HIGH-ORDER FINITE ELEMENT APPROXIMATIONS OF THE MAXWELL EQUATIONS

Domokos Sármány

Graduation committee

Chairman	
prof.dr. P.J. Gellings	University of Twente
Supervisor	
prof.dr.ir. J.J.W. van der Vegt	University of Twente
Assistant supervisor	
dr. M.A. Botchev	University of Twente
Members	
prof.dr. I. Faragó	Eötvös Loránd University
prof.dr. J.G. Verwer	Center for Mathematics and Computer Science
prof.dr. W.H.A. Schilders	Eindhoven University of Technology
prof.dr. S.A. van Gils	University of Twente
dr. M. Hammer	University of Twente

The research in this thesis was supported by the Dutch government through the national program BSIK: knowledge and research capacity, in the ICT project BRICKS (`http://www.bsik-bricks.nl`), theme MSV1

Cover design: on the front is an electromagnetic cavity, courtesy of Materials Processing Laboratory, MIT; and on the back are the fish called Des and Dom, courtesy of Denis Miretskiy

ISBN 978-90-365-2968-6

Printed by Wöhrmann Print Service, the Netherlands

HIGH-ORDER FINITE ELEMENT APPROXIMATIONS OF THE MAXWELL EQUATIONS

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof.dr. H. Brinksma,
on account of the decision of the graduation committee,
to be publicly defended
on Friday 12$^{\text{th}}$ February 2010 at 13:15

by

Domokos Sármány

born on 27$^{\text{th}}$ November 1980
in Budapest, Hungary

This dissertation has been approved by the supervisor,
prof.dr.ir. J.J.W. van der Vegt,

and the assistant supervisor,
dr. M.A. Botchev

To the memory of my father

ACKNOWLEDGMENTS

It would be difficult to overstate the role that my supervisor Jaap van der Vegt has played in my completing this work. I thank him for providing me with the opportunity to pursue a PhD at the University of Twente, and for showing me that doing mathematics in an engineering way is not only useful but it can also be fun. I am equally thankful to my daily supervisor Mike Botchev for his patience, time and the many fruitful discussions we had. His unstinting support helped me get through the most difficult of times that inevitably occur in such a project.

I have had both the pleasure and privilege to be able to closely collaborate with two other excellent mathematicians. I would like to thank Jan Verwer for his valuable input to the time-integration part of this thesis, and Ferenc Izsák for carrying out most of the theoretical work for the parameter estimation. Their contributions have greatly enhanced the quality of this thesis. My special thanks go to István Faragó, who (what seems now a long time ago) set me on the path to scientific computing and numerical analysis. Over the years, our conversations have gradually moved from the scientific to the personal, just as he has slowly turned from an MSc supervisor into a friend.

I would like to express my acknowledgment to the rest of the graduation committee, P.J. Gellings, W.H.A. Schilders, S.A. van Gils and M. Hammer for their time and effort in reviewing this thesis.

I am grateful to the whole NACM group for a friendly and simulating work environment. Most of the research in this thesis forms part of an ambitious project named hpGEM, run by the NACM group. The hpGEM meetings, held regularly, were especially useful and motivating for which I would like to thank the whole team: Jaap, Onno, Ruud, Vijaya, Henk, Sander, Lars, Alex, Tito and Shavarsh. During the past four years, I have been lucky enough to share office with a relatively large number of colleagues. They were good company, always ready to help with work or to have a chat with in case work was just, well, a bit too much. I thank them all: Sena, Chris, Vijaya, Davit, Alyona, Remco, Marcel, Lie, Bob and Sander. My special gratitude goes to Vijaya and Sander, who were always there to help, be

it DG, hpGEM, the Dutch language, tennis or a lift to Rotterdam. Many thanks to Marielle, Linda and Carin for being the heart and soul of the group, to Ivan for organising regular sport events and to Bob for keeping me environmentally conscious.

Probably the most effective diversion from work was playing for the University's basketball team. I am especially thankful to our trainer Bas for being always cheerful and encouraging, even though our team were undoubtedly the lousiest in the league.

My time in Enschede would have been an altogether different experience without the friends I made there outside work: Jon, Babsi, Denis, Katya, Des, Clare, Paul, Federica, Peter, Connie, Aimee, Uros, Mila, Nat, Katja, Pavel, Eva, Markus, Vas, Hicham and Yana. Their support and friendship are more than appreciated. Special thanks to Denis for all the fish he bought, gutted and cooked; to Jon for always offering me an extra drink when I was adamant to leave; to Paul for providing drama in a hopelessly undramatic place; and to Clare for making the Dutch classes even more fun to attend.

I am happy to thank Nico for visiting me in Enschede almost every time he had some work to do in the Netherlands. I would like to mention my great friends in Budapest, especially Bogi, Veró, Gergő, Áron, Bandi, Vera and Szó. Vitally, I also mention my entire family, including my Mum, Bori, Philip, Viki, Peti, Titi and Gábor. I am grateful to them all for making sure that I continued to feel at home in my home city.

However, the person who has contributed by far the most to this thesis is my wife, who has had to bear the brunt of the ups and downs of my PhD candidacy. And who is, of course, essential. Zsofka, my apologies and love.

CONTENTS

CHAPTER 1

BACKGROUND AND MOTIVATION

1.1 Introduction

In 2006 readers of Physics World[1] voted the Maxwell equations as the greatest mathematical equation(s) of all times. It topped a list, shown in Figure 1.3, that included such better-known competitors as Newton's second law or Einstein's famous $E = mc^2$. But is this high standing justified? Probably yes. The Maxwell equations describe electromagnetic waves and their propagative properties. The laws of electromagnetism – in the way that James Clerk Maxwell formulated them in the second half of the 19th century – gave rise to radiotechnology, from which information technology later evolved. It would be very difficult today to imagine the world without all the appliances, machines and technologies that are built on the principles of electromagnetic theory and have such a major impact on our lives. These include satellite and stealth technologies, magnetic resonance imaging (MRI) scanners, various household appliances or even lightning localisation[2] – to name but a few. (See Figure 1.2.)

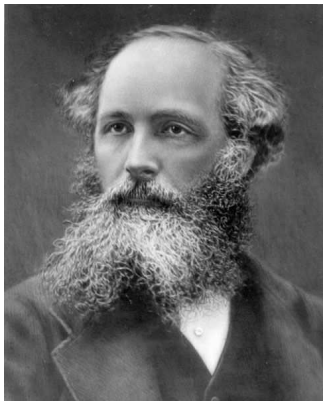


Figure 1.1: *James Clerk Maxwell (1831–1879)*

In truth, almost all of the laws of electricity and magnetism predate Maxwell's work. His contribution proved groundbreaking for two, strongly related, reasons.

[1] `http://physicsworld.com/cws/home`

[2] Insurance companies are especially interested in being able to determine where precisely a lightning has struck

Figure 1.2: *A small selection of technologies that emerged in part thanks to the electromagnetic theory of Maxwell. Starting from the top left corner and proceeding clockwise, they show a satellite dish, a stealth bomber, a lightning strike (the technology involved is the localisation of a lightning strike), a microwave oven and an MRI scanner.*

First, by modifying Ampère's circuital law, he established a link between the electric and magnetic fields that showed that the two are inherently intertwined. More specifically, it shed light on the existence and nature of electromagnetic waves. Second, he unified existing laws of electricity and magnetism in a single mathematical framework, where the physical quantities are described as three-dimensional vector fields and their change in space and time can be represented by mathematical operators. This allowed later researchers and engineers to use a whole set of mathematical tools in order to gain further insight into the electromagnetic phenomena of nature and to make pioneering technological progress.

However, solving the Maxwell equations often proves to be a formidable task, even when advanced mathematics is used. An exact solution is more often than not unattainable and one has to resort to approximating the solution by means of a numerical discretisation method instead. In broad terms, this thesis is about the numerical discretisation of the Maxwell equations. There are various existing such techniques with vast literature. Which one is best to use primarily depends on the physical problem one needs to solve but also on other, less scientific, factors such as tradition of a given scientific area. Out of the many numerical methods, this

thesis focuses on the discontinuous Galerkin finite element method (DG-FEM), a type of finite element method (FEM) that allows the discrete (i.e. approximating) representation of the fields to be discontinuous. It is a relatively new computational technology that for certain applications offers a number of advantages over existing numerical methods.

The remaining part of this chapter reviews some of the basic concepts that underlie the (DG-)FEM discretisation of the Maxwell equations. At this stage, the emphasis is not on providing rigorous definitions and derivations – that is left to later chapters. Rather, the goal here is to give an intuitive overview of the Maxwell equations and of the numerical discretisations that are most typically applied to them, with the focus being on (dis)continuous FEM. The original research presented in subsequent chapters is also outlined.

1.2 The Maxwell equations

We now recall the laws of electromagnetism, first as they were known before Maxwell's work and then in the form they took as a result of his work. See also the excellent textbook [32] for an introduction to electromagnetic theory.

1.2.1 Laws of electromagnetism before Maxwell

There are a set of four equations that were later to be dubbed the Maxwell equations. At the time, however, they were only the principles of electricity and magnetism.

Gauss's law *The electric flux through any closed surface is proportional to the enclosed electric charge.* One of the simplest applications of this law is to demonstrate that a Faraday cage[3] blocks out external electric fields. Mathematically, Gauss's law in differential form is written as

$$\nabla \cdot (\varepsilon \boldsymbol{E}) = \varrho, \tag{1.1}$$

where the operator $\nabla\cdot$ denotes divergence, $\boldsymbol{E}$ is the electric field and ϱ is the electric charge density. The material coefficient ε is the electric permittivity and assumed to be piecewise constant throughout this thesis.

Law of magnetism *Magnetic monopoles do not exist.* As it will be apparent shortly, this is the law that spoils the symmetry in the Maxwellian theory of electromagnetism. It would seem very natural indeed if magnetic charges, just like electric charges, existed. But despite unstinting efforts to identify magnetic monopoles, they remain elusive. Mathematically, the differential form of this law is written as

$$\nabla \cdot (\mu \boldsymbol{H}) = 0, \tag{1.2}$$

[3] A Faraday cage is an enclosure formed by conducting material or by a mesh of such material

where $\boldsymbol{H}$ is the magnetic field and the material coefficient μ is the magnetic permeability. Just as ε, it is assumed to be piecewise constant throughout the thesis.

Faraday's law *A changing magnetic field will create an electric field.* This principle is involved in the working of transformers, inductors, and many forms of electrical generators. In differential form, it is mathematically formulated as

$$\nabla \times \boldsymbol{E} = -\mu \frac{\partial \boldsymbol{H}}{\partial t}, \tag{1.3}$$

where $\nabla\times$ is the curl (or rotation) operator and $\frac{\partial}{\partial t}$ denotes the partial time derivative.

Ampère's law *Electric current generates a magnetic field.* The most obvious application of this principle is the way electromagnets operate. The mathematical form reads

$$\nabla \times \boldsymbol{H} = \boldsymbol{J}, \tag{1.4}$$

where $\boldsymbol{J}$ is the electric current density. In general, the current density takes the form $\boldsymbol{J} = \sigma \boldsymbol{E} + \boldsymbol{J}_s$, where σ is the electric conductivity and $\boldsymbol{J}_s$ is the density of the external source current.

1.2.2 Laws of electromagnetism since Maxwell

Maxwell realised that the last of the above physical laws is not complete. Instead, Ampère's law in the correct form should read: *Electric current generates a magnetic field and changing electric field also generates a magnetic field.* Mathematically, it takes the form

$$\nabla \times \boldsymbol{H} = \boldsymbol{J} + \varepsilon \frac{\partial \boldsymbol{E}}{\partial t}. \tag{1.5}$$

The significance of Maxwell's correction lies in the fact that (1.3) and (1.5) now represent the relation between the electric field $\boldsymbol{E}$ and the magnetic field $\boldsymbol{H}$ in a dynamically symmetric way. This symmetry directly led to the discovery of electromagnetic waves and to the realisation that light itself is an electromagnetic wave.

Note that (1.1)–(1.3) and (1.5) do not constitute the most frequently cited form of the Maxwell equations. That generally involves four equations with four unknowns (see also Figure 1.3. Indeed, at first, (1.1)–(1.3) and (1.5) may seem to be overdefined. This is not the case, however, since as long as the initial condition satisfies (1.1) and (1.2), the other two, time-dependent equations, (1.3) and (1.5), guarantee that all four equations are satisfied at any later time.

Out of the many various physical processes that are described by (1.1)–(1.3) and (1.5), the focus in this thesis is on electromagnetic waves. This involves the simultaneous solution of (1.3) and (1.5),

$$\varepsilon \frac{\partial \boldsymbol{E}}{\partial t} = \nabla \times \boldsymbol{H} - \boldsymbol{J}, \quad \mu \frac{\partial \boldsymbol{H}}{\partial t} = -\nabla \times \boldsymbol{E}, \tag{1.6}$$

under the condition that the two unknown fields satisfy (1.1) and (1.2) at initial time. Assuming a sinusoidal wave solution, (1.6) can also be written as

$$\mathrm{i}\omega\varepsilon\boldsymbol{E} = \nabla\times\boldsymbol{H} - \boldsymbol{J}, \quad \mathrm{i}\omega\mu\boldsymbol{H} = -\nabla\times\boldsymbol{E}, \tag{1.7}$$

where ω is the temporal frequency, $\mathrm{i}^2 = -1$ and all fields depend solely on space. Using straightforward substitutions, both (1.6) and (1.7) can be expressed as a second-order differential equation, rather than a first-order system,

$$\varepsilon\frac{\partial^2\boldsymbol{E}}{\partial t^2} + \nabla\times\left(\mu^{-1}\nabla\times\boldsymbol{E}\right) = -\frac{\partial\boldsymbol{J}}{\partial t}, \tag{1.8}$$

$$\nabla\times\left(\mu^{-1}\nabla\times\boldsymbol{E}\right) - \varepsilon\omega^2\boldsymbol{E} = -\mathrm{i}\omega\boldsymbol{J}. \tag{1.9}$$

The four different forms (1.6)–(1.9) are mathematically equivalent at the continuous level. However, the properties of the discretised versions may be rather different for the different formulations. In this thesis, we will discuss discretisations of three of the above four forms: (1.6) in Chapter 2, (1.9) in Chapter 4 and (1.8) in Chapter 5.

1.3 Numerical approximation

We now give a brief overview of the most widely-used numerical approximation techniques for the Maxwell equations. For this purpose, we consider the Maxwell equations in the flux form

$$Q(\boldsymbol{x})\frac{\partial\boldsymbol{q}}{\partial t} + \nabla\cdot\boldsymbol{F}(\boldsymbol{q}) = 0, \tag{1.10}$$

where $Q(\boldsymbol{x})$ represents the material properties, $\boldsymbol{q}$ is the vector of the field values and $\boldsymbol{F}(\boldsymbol{q}) = [F_1(\boldsymbol{q}), F_2(\boldsymbol{q}), F_3(\boldsymbol{q})]^T$ denotes the flux. Namely,

$$Q(\boldsymbol{x}) = \operatorname{diag}\left(\varepsilon_\mathrm{r}, \varepsilon_\mathrm{r}, \varepsilon_\mathrm{r}, \mu_\mathrm{r}, \mu_\mathrm{r}, \mu_\mathrm{r}\right), \quad \boldsymbol{q} = \begin{bmatrix}\boldsymbol{E}\\ \boldsymbol{H}\end{bmatrix}, \quad F_i(\boldsymbol{q}) = \begin{bmatrix}-\boldsymbol{e}_i\times\boldsymbol{H}\\ \boldsymbol{e}_i\times\boldsymbol{E}\end{bmatrix},$$

where $\boldsymbol{e}_i$ is the corresponding Cartesian unit vector.

The general underlying idea is to replace the continuous fields in the Maxwell equations with a discrete counterpart. The discretised version of the Maxwell equations is then possible to solve even if the (exact) solution of the original equations is hard, or even impossible, to obtain. However, the discrete solution will only be an approximate representation of the exact solution.

1.3.1 Finite difference method

Let us begin with the simplest and historically oldest method, known as the finite difference method. In this approach, a grid is laid down in space and spatial derivatives are approximated by finite differences. In the context of the Maxwell

equations, the staggered grid method of Yee [86] has proved especially popular thanks to the improved approximation properties provided by the staggering.

One of the most appealing aspects of this method is its simplicity. The discretisation of general problems and operators is often intuitive and it leads to very efficient schemes for many problems. Furthermore, the explicit semi-discrete form gives flexibility in the choice of time-integration methods if needed, and these methods are supported by an extensive body of theory [36].

It is also, however, the simplicity of the inherently one-dimensional concept of finite differences that is the main disadvantage of the method. It enforces a simple dimension-by-dimension structure in higher dimensions, and introduces additional complications around boundaries and discontinuous internal layers. This makes the finite difference method ill-suited to deal with complex geometries, both in terms of general computational domains and internal discontinuities. Changes to local order and grid size to reflect local features of the solution is also problematic.

1.3.2 Finite volume method

The above discussion highlights that to ensure geometric flexibility, one needs to abandon the simple one-dimensional approximation in favour of something more general. The most natural approach is to introduce an element-based discretisation. The computational domain is divided into a number of cells or elements[4], organised in an unstructured manner to fill the physical domain. In its simplest form, the finite volume method approximates the physical field in each element by a constant.

The scheme is purely local and thus imposes no conditions on the grid structure. In particular, all cells can have different mesh sizes. The flux terms $\boldsymbol{F}(\boldsymbol{q})$ reduce to pure surface terms by the use of the divergence theorem. This step introduces the need to evaluate the fluxes at the element boundaries. In extremely special cases, such as linear problems on equidistant grids, this method reduces to the finite difference method. But the formulation is much less restrictive in terms of the grid structure. The representation of solution values at the interfaces is a local procedure and generalises straightforwardly to unstructured grids in higher dimensions, thus ensuring the desired geometric flexibility. Furthermore, the construction of the interface fluxes can be done in various ways that are closely related to the particular equations [53, 79].

If, however, one needs to increase the order of accuracy of the method, a substantial problem may emerge in the classical finite volume approach[5]. In the simple one-dimensional case, this can be done straightforwardly by extending the size of the stencil. But in higher dimensions, extending the stencil reintroduces the need for a particular grid structure and thus compromises the geometric flexibility of the

[4] These are typically triangles or quadrilaterals in two dimensions and tetrahedra, hexahedra or prisms in three dimensions

[5] By 'classical' we refer to the approach where the fields are approximated by a constant in each cell (or element). It is possible to include higher-order moments in each cell (or element) and thus construct a compact-stencil high-order finite volume scheme. However, in that case the final form of the discretisation is little different from discontinuous Galerkin schemes.

finite volume method in higher dimensions. The main problem with the high-order representation is that it usually needs to span multiple elements as the numerical approximation of q is given by cell averages only.

1.3.3 Finite element method

Rather than representing the numerical discretisation by cell averages, one could be tempted to take a different approach and introduce more degrees of freedom in the element. This is what happens in the finite element method where, similar to the finite volume method, the computational domain is divided into a number of possibly unstructured elements. Inside each element, we assume that the local solution is expressed as the linear combination of locally defined basis functions. With this local element-based model, each element shares the nodes with other elements. If we combine the elementwise approximations, we have a global continuous representation of the unknown $\boldsymbol{q}$.

To recover a scheme to solve (1.10), we need to define the space of test functions, and require that the residual be orthogonal to all test functions in this space. The details of the scheme are determined by how this space of test functions is defined. A classical choice, leading to a Galerkin scheme, is to require that the spaces spanned by the basis functions and test functions be the same.

This approach reflects, in essence, the classical FEM and clearly allows different element sizes [76, 87, 88, 89]. Furthermore, we recall that the main motivation behind considering methods beyond the finite volume approach was the interest in high-order approximations. Such extensions are relatively simple in the finite-element setting and can be achieved by adding additional degrees of freedom to the element while maintaining shared nodes along the faces of the elements [52].

However, there are drawbacks of the classical continuous finite element formulation, for hyperbolic problems in general and for the Maxwell equations in particular. First, the globally defined basis functions and the requirement that the residual be orthogonal to the same set of globally defined test functions implies that the semi-discrete scheme becomes implicit. For time-dependent problems, this is a clear disadvantage compared with finite difference or finite volume methods.

Second, there is an issue that relates to the structure of the basis and is especially important for the Maxwell equation. That is, the original physical fields $\boldsymbol{E}$ and $\boldsymbol{H}$ are not necessarily continuous. The Maxwell equations require the tangential component of either field to be continuous, but there is no requirement for the normal component. Representing these fields with continuous functions at the discrete level may result in convergence to spurious, non-physical solutions. In the finite element context, this problem can be solved by the use of a special type of basis functions for which only tangential continuity is prescribed.

1.3.4 Discontinuous Galerkin method

An intelligent combination of the finite element and finite volume methods appears to offer many of the desired properties. This combination is exactly what leads to

the discontinuous Galerkin FEM [63, 20, 22, 4, 42].

In this approach, we maintain the definition of elements as in the finite element and finite volume schemes. But in order to ensure the locality of the scheme, we allow the variables located at each node to have multiple values – as many as the number of elements the node belongs to. In each of these elements, we again assume that the local solution can be expressed as a linear combination of locally defined basis functions. The space spanned by these local basis functions is also local and we now require that the solution be orthogonal to this local space. At this stage, however, the discretisation is useless, since it does not allow one to recover a meaningful global solution. Furthermore, the nodes at the boundary of the element take values from multiple elements so uniqueness of the solution is not guaranteed.

To overcome these problems, we connect the elements through what are called numerical fluxes. After integration by parts and the use of the divergence theorem, these can be defined as unique values to be used at the interface and obtained by combining information from both elements. The choice of the numerical flux is a central element of the scheme and is also where one can introduce knowledge of the dynamics of the problem in the numerical discretisation.

While the structure of the DG-FEM is very similar to that of the continuous FEM, there are several significant differences. Most important, the mass matrix is local rather than global and can be inverted at very little cost, yielding a semi-discrete scheme that is explicit. Furthermore, by carefully designing the numerical flux to reflect the underlying dynamics, on has more flexibility than in the classical FEM to ensure stability for wave-dominated problems. Compared with the finite volume approach, the DG-FEM overcomes the key limitation on achieving high-order accuracy on general grids by enabling this through a local element-based basis. This is all achieved while maintaining benefits such as local conservation and flexibility in the choice of the numerical flux.

All this, however, comes at a price – most notably through an increase in the total degrees of freedom as a direct result of the decoupling of the elements. For linear elements in one dimension, this yields a doubling in the total number degrees of freedom, compared with the continuous FEM. In higher dimensions, the difference is even greater for linear basis functions. In certain applications, such when the inversion of a discrete operator is needed anyway, this clearly becomes an issue of significant importance. Furthermore, for problems where the flexibility in the flux choice and the locality of the scheme is less important (e.g. in elliptic problems), the DG-FEM is not as efficient as a better-suited method such as the continuous FEM.

Nevertheless, the relative increase in degrees of freedom of high-order accurate DG-FEM compared with classical $H(\mathrm{curl})$-conforming FEM is rather small. This, together with the additional flexibility of DG methods, makes them an excellent numerical technique for accurate discretisations of the Maxwell equations.

1.4 Outline of the thesis

The remaining part of the thesis consists of four chapters. Each of these chapters is an independent article that has been published in, or submitted to, peer-reviewed international journals in the field of scientific computation.

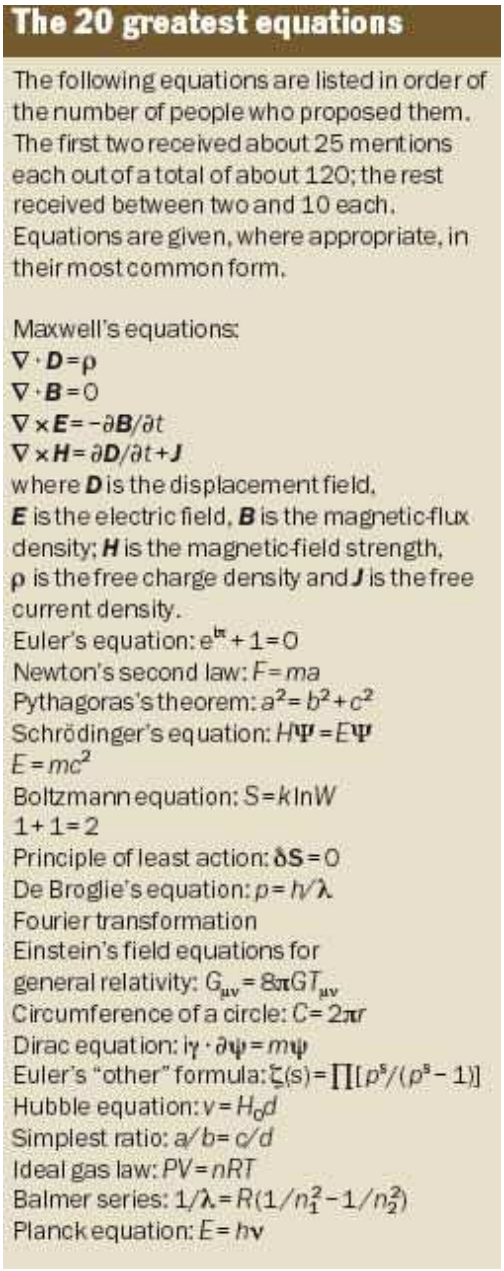
In Chapter 2, different time-stepping methods for a nodal high-order discontinuous Galerkin discretisation of the Maxwell equations are discussed. A comparison between the most popular choices of Runge-Kutta (RK) methods is made from the point of view of accuracy and computational work. By choosing the strong-stability-preserving Runge-Kutta (SSP-RK) time-integration method of order consistent with the polynomial order of the spatial discretisation, better accuracy can be attained compared with fixed-order schemes. Moreover, this comes without a significant increase in computational work. A numerical Fourier analysis is performed for this Runge-Kutta discontinuous Galerkin (RKDG) discretisation to gain insight into the dispersion and dissipation properties of the fully discrete scheme. The analysis is carried out on both the one-dimensional and the two-dimensional fully discrete schemes and, in the latter case, on uniform as well as on non-uniform meshes. It also provides practical information on the convergence of the dissipation and dispersion error up to polynomial order 10 for the one-dimensional fully discrete scheme.

Chapter 3 describes the implementation details of a hierarchic $H(\text{curl})$-conforming tetrahedral basis in the finite element discretisation of the Maxwell equations. This basis is especially useful in the development of p- and hp-adaptive methods. It can in a natural way be applied to the Maxwell equations in both $H(\text{curl})$-conforming and discontinuous Galerkin finite element discretisations. Although the construction of the basis is relatively well-known, it is essential that we draw attention to a possible practical difficulty in the implementation of such a basis for the $H(\text{curl})$-conforming finite element method. This concerns one particular type of basis function, the face-based bubble function, which is part of the basis for polynomial orders $p \geq 3$. We show, through a simple example, that face-based bubble functions do not, in general, transform from the reference tetrahedron to a physical tetrahedron in a conforming way, even after the issue of orientation has been carefully addressed. As a consequence, special care needs to be exercised to guarantee the global tangential continuity required by the $H(\text{curl})$-conforming method. Once that is ensured, the finite element discretisation can be obtained in two main steps. First, compute the discretisation for each element (and face if DG is used), e.g. build the element (and face) matrices. Second, assemble the local matrices into global ones which form the algebraic system or the system of ordinary differential equations that are to be solved to get the discrete solution.

In Chapter 4, we provide optimal parameter estimates and a priori error bounds for symmetric discontinuous Galerkin (DG) discretisations for the second-order indefinite time-harmonic Maxwell equations. More specifically, we consider two variations of symmetric DG methods: the interior penalty DG (IP-DG) method and one that makes use of the local lifting operator in the flux formulation. As a novelty, our parameter estimates bounds are *i)* valid in the pre-asymptotic regime;

ii) solely depend on the geometry and the polynomial order; and *iii)* are free of unspecified constants. Such parameter estimates are particularly important in three-dimensional (3D) simulations because in practice many 3D computations occur in the pre-asymptotic regime. Therefore, it is vital that our numerical experiments that accompany the theoretical results are also in 3D. They are carried out on tetrahedral meshes with high-order ($p = 1, 2, 3, 4$) hierarchic $H(\text{curl})$-conforming polynomial basis functions.

Chapter 5 compares the discontinuous Galerkin finite element method (DG-FEM) with the $H(\text{curl})$-conforming FEM in the discretisation of the second-order time-domain Maxwell equations with a possibly nonzero conductivity term. While DG-FEM suffers from an increased number of degrees of freedom compared with $H(\text{curl})$-conforming FEM, it has the advantage of a purely block-diagonal mass matrix. This means that, as long as an explicit time-integration scheme is used, it is no longer necessary to solve a linear system at each time step – a clear advantage over $H(\text{curl})$-conforming FEM. It is known that DG-FEM generally favours high-order methods whereas $H(\text{curl})$-conforming FEM is more suitable for low-order ones. The novelty we provide in this chapter is a direct comparison of the performance of the two methods when hierarchic $H(\text{curl})$-conforming basis functions are used up to polynomial order $p = 3$. The motivation behind this choice of basis functions is its growing importance in the development of p- and hp-adaptive FEMs. The fact that we allow for nonzero conductivity requires special attention with regards to the time-integration methods applied to the semi-discrete systems. A high-order polynomial basis warrants the use of high-order time-integration schemes, but existing high-order schemes may suffer from a too severe time-step stability restriction as result of the conductivity term. We investigate several alternatives from the point of view of accuracy, stability and computational work. Finally, we carry out a numerical Fourier analysis to study the dispersion and dissipation properties of the semi-discrete DG-FEM scheme and several of the time-integration methods. It is instructive in our approach that the dispersion and dissipation properties of the spatial discretisation and those of the time-integration methods are investigated separately, providing additional insight into the two discretisation steps.



The 20 greatest equations

The following equations are listed in order of the number of people who proposed them. The first two received about 25 mentions each out of a total of about 120; the rest received between two and 10 each. Equations are given, where appropriate, in their most common form.

Maxwell's equations:
$\nabla \cdot \boldsymbol{D} = \rho$
$\nabla \cdot \boldsymbol{B} = 0$
$\nabla \times \boldsymbol{E} = -\partial \boldsymbol{B}/\partial t$
$\nabla \times \boldsymbol{H} = \partial \boldsymbol{D}/\partial t + \boldsymbol{J}$
where $\boldsymbol{D}$ is the displacement field, $\boldsymbol{E}$ is the electric field, $\boldsymbol{B}$ is the magnetic-flux density; $\boldsymbol{H}$ is the magnetic-field strength, ρ is the free charge density and $\boldsymbol{J}$ is the free current density.
Euler's equation: $e^{i\pi} + 1 = 0$
Newton's second law: $F = ma$
Pythagoras's theorem: $a^2 = b^2 + c^2$
Schrödinger's equation: $H\Psi = E\Psi$
$E = mc^2$
Boltzmann equation: $S = k \ln W$
$1 + 1 = 2$
Principle of least action: $\delta \mathbf{S} = 0$
De Broglie's equation: $p = h/\lambda$
Fourier transformation
Einstein's field equations for general relativity: $G_{\mu\nu} = 8\pi G T_{\mu\nu}$
Circumference of a circle: $C = 2\pi r$
Dirac equation: $i\gamma \cdot \partial\psi = m\psi$
Euler's "other" formula: $\zeta(s) = \prod [p^s/(p^s - 1)]$
Hubble equation: $v = H_0 d$
Simplest ratio: $a/b = c/d$
Ideal gas law: $PV = nRT$
Balmer series: $1/\lambda = R(1/n_1^2 - 1/n_2^2)$
Planck equation: $E = h\nu$

Figure 1.3: *The twenty greatest equations based on the votes of the readers of Physics World in 2006. This image has been downloaded from and is also available at* `http://physicsworld.com/cws/article/print/20407`.

CHAPTER 2

DISPERSION AND DISSIPATION ERROR IN HIGH-ORDER RUNGE-KUTTA DISCONTINUOUS GALERKIN DISCRETISATIONS OF THE MAXWELL EQUATIONS

2.1 Introduction

As pointed out in an extensive review on the state of the art of computational electromagnetics [40], in many cases finite-difference time-domain (FDTD) schemes [77, 86] are undoubtedly the most popular methods among physicists and engineers to solve the time-domain Maxwell equations numerically. This popularity is mainly due to their simplicity and efficiency in discretising simple-domain problems. However, their inability to effectively handle complex geometries prompted some scientists to search for alternatives long ago. Finite-element (FE) methods are an obvious alternative, but early efforts were marred by the fact that standard continuous Galerkin finite-element schemes give rise to non-physical solutions. Most apparent of these are the spurious modes in the numerical solution of the frequency-domain Maxwell equations (see [51] and references therein). The revolutionary solution to this problem was to realise that by using a particular set of vector basis functions (vector elements such as Nédélec or Whitney elements [45, 59]), it is possible to mimic many of the special properties of the Maxwell equations at the discrete level. See [8] and [9]. Ever since, vector elements have been a viable alternative to FDTD and standard FE methods in computational electrodynamics, especially for frequency-domain problems with complex geometries. The practical considerations of both standard and vector finite elements in computational electromagnetics are covered in [51]. For the more theoretical aspects of Nédélec elements we refer to [56].

The need to model electromagnetic wave propagation in large and complex domains and over a relatively long time span has increased the demand for high-order methods. However, neither high-order FDTD methods nor high-order vector FE methods are devoid of practical drawbacks. High-order FDTD methods fail to effectively handle complex geometries whereas high-order vector FE methods (based on high-order Nédélec elements [59] for example) lead to global mass matrices with relatively large bandwidths (after optimal reordering). The time-integration schemes to solve such systems are in turn computationally rather expensive. These difficulties have motivated the development of discontinuous Galerkin (DG) finite-element methods [20, 22], together with spectral element methods [52]. In both the frequency-domain formulation [43, 47, 60, 61, 81] and the time-domain formulation [17, 21, 42, 57] significant progress has been made. One of the most promising methods for complicated geometries is the high-order nodal DG method of Hesthaven and Warburton [42], which proved both accurate and efficient for the spatial discretisation. In time integration, however, the low-storage Runge-Kutta (RK) method the authors applied poses a comparatively stringent time-step constraint, which may turn out to be the bottleneck for long-time integration. Furthermore, fixed-order time-integration schemes may spoil the high-order convergence of the global scheme. In the meantime, for discontinuous formulations of convection-dominated problems [22] it has been shown in [31] and in [17] that the time-step restriction may be loosened if we use Strong-Stability-Preserving Runge-Kutta (SSP-RK) methods of one order higher than the polynomial order of the spatial discretisation.

In this chapter, we study the behaviour of the high-order nodal scheme when several of the best-suited time-integration methods are used. In particular, we have a closer look at the dispersion and dissipation properties of the Runge-Kutta discontinuous Galerkin (RKDG) method comprising the nodal high-order DG method and the SSP-RK method. The main motivation for using this particular time-integration scheme is its relatively weak time-step restriction. This property implies that we can retain high-order accuracy without losing much on the computational work measured as the number of operations.

The literature on the dispersion and dissipation properties of the DG method has in recent years become abundant. A thorough analysis of the dispersion and dissipation behaviour of the DG method for the transport equation (scalar linear conservation law) was given in [1], which also provided a proof for earlier conjectures, especially from [49]. The semi-discrete system for the wave equation has also been extensively studied [3, 50, 71]. In particular, the authors in [3] provided two different dispersion analyses for the semi-discrete wave equation on tensor product elements. One for the interior penalty DG method (IP-DG) of the second-order wave equation and another for the general DG method for a first order system.

The novelty of this chapter with regards to the dispersion and dissipation behaviour of DG methods lies in including the time integration in the analysis. We consider the discretisation of the first-order system related to the Maxwell equations, so our scheme falls in the category of what the authors call the 'general DG method' in [3]. Throughout this chapter we apply a fully upwinding numerical flux,

since it has proved superior–due to stabilisation and lack of spurious modes–to the centered or mixed fluxes for time-dependent problems [43]. In wave-propagation problems it is often more advantageous to know the convergence rate of the dispersion and dissipation errors than that of the error in the L_2-norm. These convergence rates have been established in [1] for the semi-discrete transport equation. For the general DG scheme, to which the nodal DG method discussed here belongs, it has been shown in [3] that using first-order polynomials in the spatial discretisation results in a dispersion error of order $\mathcal{O}(h^4)$ and a dissipation error of order $\mathcal{O}(h^3)$ for the semi-discrete system. In this chapter we show, through numerical examples, how the dispersion and dissipation errors converge in the fully discrete high-order RKDG scheme for the linear autonomous form of the Maxwell equations.

The remaining part of this chapter is outlined as follows. In Section 2.2 we recall the system of time-domain Maxwell equations and reduce it to the linear autonomous form. The spatial discretisation is briefly reviewed in Section 2.3 and the RK schemes for the temporal discretisation in Section 2.4. One-dimensional and two-dimensional Fourier analysis is carried out in Section 2.5, and the associated numerical results, along with some other numerical tests, are presented in Section 2.6. Here we examine the behaviour of the dispersion and dissipation errors in terms of the mesh size per wave length and the size of the time step. Finally, we sum up our conclusions in Section 2.7.

2.2 Maxwell equations

We begin with deriving the dimensionless time-domain form of the Maxwell equations in the three-dimensional domain $\Omega \subset \mathbb{R}^3$. Boldface symbols here refer to vector fields, i.e. fields in $\mathbb{R}^3 \to \mathbb{R}^3$. With these notations the Maxwell equations read

$$\frac{\partial \boldsymbol{D}}{\partial t} = \nabla \times \boldsymbol{H} - \boldsymbol{J}, \qquad \frac{\partial \boldsymbol{B}}{\partial t} = -\nabla \times \boldsymbol{E}, \tag{2.1}$$

$$\nabla \cdot \boldsymbol{D} = \varrho, \qquad \nabla \cdot \boldsymbol{B} = 0, \tag{2.2}$$

with charge distribution $\varrho(\boldsymbol{x}, t)$, position vector $\boldsymbol{x} = (x, y, z) \in \Omega$, the nabla operator $\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right)$ and time t. The vector valued quantities are the electric field $\boldsymbol{E}(\boldsymbol{x}, t)$, the electric flux density $\boldsymbol{D}(\boldsymbol{x}, t)$, the magnetic field $\boldsymbol{H}(\boldsymbol{x}, t)$, the magnetic flux density $\boldsymbol{B}(\boldsymbol{x}, t)$ and the electric current density $\boldsymbol{J}(\boldsymbol{x}, t)$. For many applications it is reasonable to assume that the materials are isotropic, linear and time-invariant. Thus the system of equations is closed with the linear constitutive relations

$$\boldsymbol{D} = \varepsilon_{\mathrm{r}} \boldsymbol{E}, \qquad \boldsymbol{B} = \mu_{\mathrm{r}} \boldsymbol{H}, \tag{2.3}$$

where the scalar quantities $\varepsilon_{\mathrm{r}}(\boldsymbol{x})$ and $\mu_{\mathrm{r}}(\boldsymbol{x})$ are the permittivity and permeability, respectively. Furthermore, Ohm's law

$$\boldsymbol{J} = \sigma \boldsymbol{E}$$

also holds with electric conductivity $\sigma(\boldsymbol{x}, t)$.

To obtain the non-dimensional form of the Maxwell equations (2.1)–(2.2), we first introduce tilded variables to represent the dimensional fields. The special notations $\tilde{\varepsilon}_0$ and $\tilde{\mu}_0$ stand for the dimensional permittivity and permeability of vacuum. By using the normalised space and time variables

$$\boldsymbol{x} = \frac{\tilde{\boldsymbol{x}}}{\tilde{L}}, \quad t = \frac{\tilde{t}}{\tilde{L}/\tilde{c}_0},$$

with reference length $\tilde{L}$ and dimensional speed of light in vacuum $\tilde{c}_0 = 1/\sqrt{\tilde{\varepsilon}_0\tilde{\mu}_0}$, the physical fields are made non-dimensional through the relations

$$\boldsymbol{E} = \frac{\tilde{\boldsymbol{E}}}{\tilde{Z}_0\tilde{H}_0}, \quad \boldsymbol{H} = \frac{\tilde{\boldsymbol{H}}}{\tilde{H}_0}, \quad \boldsymbol{J} = \frac{\tilde{\boldsymbol{J}}}{\tilde{H}_0/\tilde{L}}.$$

Here $\tilde{Z}_0 = \sqrt{\tilde{\mu}_0/\tilde{\varepsilon}_0}$ and $\tilde{H}_0$ are the free-space intrinsic impedance and reference magnetic field strength, respectively.

With the constitutive relations (2.3), equations (2.2) are just the consistency conditions for (2.1). To see that point, we only need to take the divergence of (2.1), apply (2.2) and (2.3) and realise that the resultant equation represents nothing else but charge conservation, which should always hold. Consequently, as long as the initial conditions satisfy (2.2) and the fields evolve according to (2.1), the solution at any time will also satisfy (2.2). It is therefore enough to consider only

$$\varepsilon_{\mathrm{r}}\frac{\partial \boldsymbol{E}}{\partial t} = \nabla \times \boldsymbol{H} - \boldsymbol{J}, \qquad \mu_{\mathrm{r}}\frac{\partial \boldsymbol{H}}{\partial t} = -\nabla \times \boldsymbol{E}, \tag{2.4}$$

in which the constitutive relations (2.3) are also included. As for the boundary conditions, one important special case is that of perfect electric conductors (PEC). These read

$$\hat{\boldsymbol{n}} \times \boldsymbol{E} = 0, \qquad \hat{\boldsymbol{n}} \times \boldsymbol{H} = 0, \tag{2.5}$$

with outward pointing normal vector $\hat{\boldsymbol{n}}$. Between material interfaces, in the absence of surface currents and surface charge, the following conditions are valid

$$\begin{aligned} \hat{\boldsymbol{n}} \times [\![\boldsymbol{E}]\!] &= 0, \qquad & \hat{\boldsymbol{n}} \cdot [\![\varepsilon_r \boldsymbol{E}]\!] &= 0, \\ \hat{\boldsymbol{n}} \times [\![\boldsymbol{H}]\!] &= 0, \qquad & \hat{\boldsymbol{n}} \cdot [\![\mu_r \boldsymbol{H}]\!] &= 0, \end{aligned} \tag{2.6}$$

where

$$[\![\boldsymbol{u}]\!] = \boldsymbol{u}^+ - \boldsymbol{u}^-$$

denotes the jump in the field value $\boldsymbol{u}$. The expressions (2.6) represent the physical property that the tangential components of both fields are continuous across different materials, whereas the normal components may be discontinuous.

2.3 Discontinuous Galerkin discretisation in space

We approximate the solutions to the Maxwell equations in space using the high-order nodal discontinuous Galerkin method introduced in [42] and further studied in [43] and [81]. In the following we briefly review the main features of this discretisation.

We consider the Maxwell equations in the general domain $\Omega \subset \mathbb{R}^3$ filled with non-conductive materials ($\sigma = 0$) and rewrite (2.4) in the flux form

$$Q(\boldsymbol{x})\frac{\partial \boldsymbol{q}}{\partial t} + \nabla \cdot \boldsymbol{F}(\boldsymbol{q}) = 0, \tag{2.7}$$

where $Q(\boldsymbol{x})$ represents the material properties, $\boldsymbol{q}$ is the vector of the field values and $\boldsymbol{F}(\boldsymbol{q}) = [F_1(\boldsymbol{q}), F_2(\boldsymbol{q}), F_3(\boldsymbol{q})]^T$ denotes the flux. Namely,

$$Q(\boldsymbol{x}) = \operatorname{diag}\left(\varepsilon_\mathrm{r}, \varepsilon_\mathrm{r}, \varepsilon_\mathrm{r}, \mu_\mathrm{r}, \mu_\mathrm{r}, \mu_\mathrm{r}\right), \quad \boldsymbol{q} = \begin{bmatrix} \boldsymbol{E} \\ \boldsymbol{H} \end{bmatrix}, \quad F_i(\boldsymbol{q}) = \begin{bmatrix} -\boldsymbol{e}_i \times \boldsymbol{H} \\ \boldsymbol{e}_i \times \boldsymbol{E} \end{bmatrix},$$

where $\boldsymbol{e}_i$ is the corresponding Cartesian unit vector. We seek the numerical solution in the computational domain Ω_K tessellated into K non-overlapping elements, i.e.

$$\Omega \approx \Omega_K = \bigcup_{k=1}^{K} \Omega^k.$$

Here Ω_K represents a tetrahedral tessellation in three dimensions and a triangular tessellation in two dimensions.

Before formulating the discontinuous Galerkin discretisation, we introduce the standard (or reference) element $\Omega_\mathrm{st} = \mathcal{T}^d$ for different spatial dimensions d. These are defined as

$$\mathcal{T}^1 = \{\xi\colon\ -1 \le \xi \le 1\}$$

in one dimension,

$$\mathcal{T}^2 = \{\boldsymbol{\xi} = (\xi, \eta)\colon\ -1 \le \xi, \eta,\ \xi + \eta \le 0\}$$

in two dimensions and

$$\mathcal{T}^3 = \{\boldsymbol{\xi} = (\xi, \eta, \zeta)\colon\ -1 \le \xi, \eta, \zeta,\ \xi + \eta + \zeta \le -1\}$$

in three dimensions. Each element Ω^k is constructed by the invertible mapping $\mathcal{X}^k(\boldsymbol{\xi})\colon \Omega_\mathrm{st} \to \Omega^k$, which is unique for any given element. For details see the extensive book [52]. We now define the finite element space as

$$\mathcal{V}_h = \left\{ \boldsymbol{q}_N^k \in \left(L^2(\Omega)\right)^{2\mathrm{d}} : \boldsymbol{q}_N^k(\mathcal{X}^k(\xi)) \in \mathcal{P}_p^d(\Omega_\mathrm{st}),\ k = 1, \ldots, K \right\}, \tag{2.8}$$

where $L^2(\Omega)$ is the space of square integrable functions on Ω and $\mathcal{P}_p^d(\Omega_\mathrm{st})$ denotes the space of d-dimensional polynomials of maximum order p on the reference element Ω_st. Since this polynomial space is associated with

$$N = \frac{(n+d)!}{n!\,d!}$$

nodal points $\boldsymbol{\xi}_i \in \Omega_{\text{st}}$, we can now introduce the multidimensional Lagrange polynomials $L_i(\boldsymbol{\xi})$ passing through these nodes:

$$L_i(\boldsymbol{\xi}_j) = \delta_{ij}, \quad \text{with} \quad \delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Taking the Lagrange polynomials as trial functions and using the mapping $\mathcal{X}^k(\xi)$, we approximate the solution at the N nodal points within each element as

$$\boldsymbol{q}^k(\boldsymbol{x}, t) \approx \boldsymbol{q}_N^k(\boldsymbol{x}, t) = \sum_{i=1}^{N} \boldsymbol{q}_i^k(t) \left(L_i(\boldsymbol{x})\right)^{2\text{d}} \in \mathcal{P}_p^d(\Omega^k),$$

where $\boldsymbol{q}_N^k(\boldsymbol{x}, t)$ is the finite element approximation, and $\boldsymbol{q}_i^k(t)$ represents the solution at nodal point $\boldsymbol{x}_j \in \Omega^k$.

The distribution of the nodes is a key issue for the properties of the interpolation, especially for very high-order approximations. It is best measured by the Lebesgue constant associated with the Lagrange polynomials going through a particular set of nodes. The Lebesgue constant shows just how close a given polynomial approximation is to the best polynomial approximation. The most popular choices for nodes in spectral/hp element methods are the Fekete points [78] and the electrostatic points [39, 41]. It should be noted that although the Fekete points have the best interpolation properties (lowest Lebesgue constant) in a triangle for orders $p \geq 9$, no distribution for a tetrahedron has so far been provided. An (almost) optimal distribution of the electrostatic nodes, however, is given for a triangle in [39] and for a tetrahedron in [41]. Moreover, the electrostatic points also perform slightly better for orders $p \leq 8$ in triangles. The distribution of these nodes in the standard triangle is shown in Figure 2.1 for orders $p = 2, 4, 6, 10$. We also note that the nodal distributions in a triangle and tetrahedron with an L^2-norm optimal Lebesgue constant were determined in [18] and [19]. However, these nodes, in contrast with the Fekete and electrostatic points, do not have an edge distribution which can be identified with Gauss-Lobatto-Jacobi points. We refer to [52] for further overview on nodal (and modal) spectral/hp methods.

To formulate the discontinuous Galerkin scheme, we first introduce the local inner product and its associated norm on Ω^k as

$$(\boldsymbol{u}, \boldsymbol{v})_{\Omega^k} = \int_{\Omega^k} \boldsymbol{u} \cdot \boldsymbol{v} \, \mathrm{d}\boldsymbol{x}, \qquad \|\boldsymbol{u}\|_{\Omega^k}^2 = (\boldsymbol{u}, \boldsymbol{u})_{\Omega^k}$$

and on its boundary $\partial\Omega^k$ as

$$(\boldsymbol{u}, \boldsymbol{v})_{\partial\Omega^k} = \int_{\partial\Omega^k} \boldsymbol{u} \cdot \boldsymbol{v} \, \mathrm{d}\boldsymbol{s}.$$

We multiply (2.7) with the local test function $\phi \in \mathcal{P}_p^d(\Omega^k)$, chosen to be the same interpolating Lagrange polynomials $L_i(\boldsymbol{x})$ for the trial basis functions, drop the

superscript k and integrate by parts over element Ω^k to obtain the continuous weak formulation

$$\left(Q\frac{\partial \boldsymbol{q}}{\partial t}, \phi\right)_{\Omega^k} - (\boldsymbol{F}, \nabla\phi)_{\Omega^k} = -(\hat{\boldsymbol{n}} \cdot \boldsymbol{F}, \phi)_{\partial\Omega^k}, \quad \forall \Omega^k \subset \Omega_K. \tag{2.9}$$

We then replace the continuous variable $\boldsymbol{q}$ with its discrete counterpart $\boldsymbol{q}_N$, and the exact flux $\boldsymbol{F}$ with the numerical flux $\widehat{\boldsymbol{F}}$ to account for the multi-valued traces at the element boundary. Finally, integration by parts for the second time results

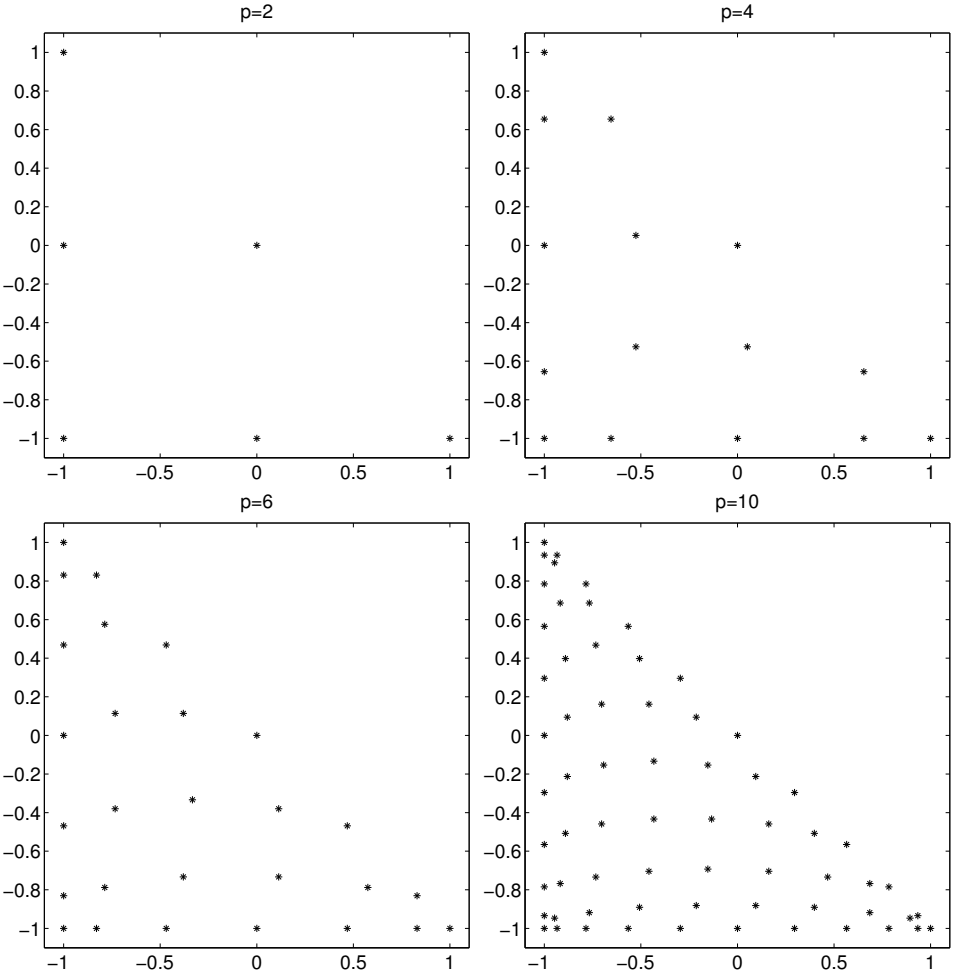


Figure 2.1: *Electrostatic points for orders* $p = 2, 4, 6, 10$

in the discrete formulation

$$\left(Q\frac{\partial \boldsymbol{q}_N}{\partial t} + \nabla \boldsymbol{F}_N, \phi\right)_{\Omega^k} = \left(\hat{\boldsymbol{n}} \cdot \left[\boldsymbol{F} - \widehat{\boldsymbol{F}}\right], \phi\right)_{\partial\Omega^k}. \tag{2.10}$$

The right-hand side of (2.10) is responsible for the communication between the elements through the numerical flux $\widehat{\boldsymbol{F}}$. The role of the numerical flux in the present spatial discretisation is discussed in [43] in the light of the Maxwell eigenvalue problem. Throughout this chapter, we use the upwind flux [55], where information travels along local wave directions.

In order to formulate the upwind flux, we first introduce the impedance Z and the conductance Y defined as

$$Z = Y^{-1} = \sqrt{\mu_\mathrm{r}/\varepsilon_\mathrm{r}}.$$

We also introduce the associated quantities

$$Z^{\pm} = \frac{1}{Y^{\pm}} = \sqrt{\frac{\mu_r^{\pm}}{\varepsilon_r^{\pm}}}, \qquad \bar{Z} = \frac{Z^- + Z^+}{2}, \qquad \bar{Y} = \frac{Y^- + Y^+}{2}.$$

The upwind flux at dielectric interfaces then reads as

$$\hat{\boldsymbol{n}} \cdot \widehat{\boldsymbol{F}} = \frac{1}{2}\begin{bmatrix} \bar{Z}^{-1}\left(-\hat{\boldsymbol{n}} \times Z^- \boldsymbol{H}_N^- - \hat{\boldsymbol{n}} \times Z^+ \boldsymbol{H}_N^+ + \hat{\boldsymbol{n}} \times \hat{\boldsymbol{n}} \times [\![\boldsymbol{E}_N]\!]\right) \\ \bar{Y}^{-1}\left(\hat{\boldsymbol{n}} \times Y^- \boldsymbol{E}_N^- + \hat{\boldsymbol{n}} \times Y^+ \boldsymbol{E}_N^+ + \hat{\boldsymbol{n}} \times \hat{\boldsymbol{n}} \times [\![\boldsymbol{H}_N]\!]\right) \end{bmatrix}, \tag{2.11}$$

where $\left(\boldsymbol{E}_N^-, \boldsymbol{H}_N^-\right)$ and $\left(\boldsymbol{E}_N^+, \boldsymbol{H}_N^+\right)$ denote the local and neighbouring solution at the boundary of Ω^k, respectively. We emphasise that the cross product is defined between vectors at each node of the element. For a detailed derivation of the upwind flux we refer to [55]. We should also recognise that

$$\hat{\boldsymbol{n}} \cdot \boldsymbol{F}_N = \begin{bmatrix} -\hat{\boldsymbol{n}} \times \boldsymbol{H}_N^- \\ \hat{\boldsymbol{n}} \times \boldsymbol{E}_N^- \end{bmatrix},$$

and combining this with (2.11), the penalising boundary term will now read

$$\hat{\boldsymbol{n}} \cdot \left(\boldsymbol{F}_N - \widehat{\boldsymbol{F}}\right) = \frac{1}{2}\begin{bmatrix} \bar{Z}^{-1}\left(Z^+ \hat{\boldsymbol{n}} \times [\![\boldsymbol{H}_N]\!] - \hat{\boldsymbol{n}} \times \hat{\boldsymbol{n}} \times [\![\boldsymbol{E}_N]\!]\right) \\ \bar{Y}^{-1}\left(-Y^+ \hat{\boldsymbol{n}} \times [\![\boldsymbol{E}_N]\!] - \hat{\boldsymbol{n}} \times \hat{\boldsymbol{n}} \times [\![\boldsymbol{H}_N]\!]\right) \end{bmatrix}.$$

To obtain the semi-discrete system we introduce the N-by-N local mass and stiffness matrices as

$$\begin{aligned} \mathbf{M}_{ij} &= \left(L_i(\boldsymbol{x}), L_j(\boldsymbol{x})\right)_{\Omega^k}, & \mathbf{S}_{ij}^x &= \left(L_i(\boldsymbol{x}), \partial_x L_j(\boldsymbol{x})\right)_{\Omega^k}, \\ \mathbf{S}_{ij}^y &= \left(L_i(\boldsymbol{x}), \partial_y L_j(\boldsymbol{x})\right)_{\Omega^k}, & \mathbf{S}_{ij}^z &= \left(L_i(\boldsymbol{x}), \partial_z L_j(\boldsymbol{x})\right)_{\Omega^k}, \end{aligned} \tag{2.12}$$

and the face-based mass matrices

$$\mathbf{F}_{il} = \left(L_i(\boldsymbol{x}), L_l(\boldsymbol{x})\right)_{\partial\Omega^k}, \tag{2.13}$$

where the second index is limited to the boundaries of Ω^k.

We can now express the semi-discrete scheme as the following system of ordinary differential equations

$$\begin{aligned}
\frac{\mathrm{d}E_N^x}{\mathrm{d}t} &= (\varepsilon_\mathrm{r}\mathbf{M})^{-1}\left(\mathbf{S}^y H_N^z - \mathbf{S}^z H_N^y\right) \\
&\quad + (\varepsilon_\mathrm{r}\mathbf{M})^{-1}\,\mathbf{F}\left(\hat{\boldsymbol{n}} \times \frac{Z^+ [\![\boldsymbol{H}_N]\!] - \hat{\boldsymbol{n}} \times [\![\boldsymbol{E}_N]\!]}{Z^+ + Z^-}\right)^x \Bigg|_{\partial\Omega^k}, \\
\frac{\mathrm{d}E_N^y}{\mathrm{d}t} &= (\varepsilon_\mathrm{r}\mathbf{M})^{-1}\left(\mathbf{S}^z H_N^x - \mathbf{S}^x H_N^z\right) \\
&\quad + (\varepsilon_\mathrm{r}\mathbf{M})^{-1}\,\mathbf{F}\left(\hat{\boldsymbol{n}} \times \frac{Z^+ [\![\boldsymbol{H}_N]\!] - \hat{\boldsymbol{n}} \times [\![\boldsymbol{E}_N]\!]}{Z^+ + Z^-}\right)^y \Bigg|_{\partial\Omega^k}, \\
\frac{\mathrm{d}E_N^z}{\mathrm{d}t} &= (\varepsilon_\mathrm{r}\mathbf{M})^{-1}\left(\mathbf{S}^x H_N^y - \mathbf{S}^y H_N^x\right) \\
&\quad + (\varepsilon_\mathrm{r}\mathbf{M})^{-1}\,\mathbf{F}\left(\hat{\boldsymbol{n}} \times \frac{Z^+ [\![\boldsymbol{H}_N]\!] - \hat{\boldsymbol{n}} \times [\![\boldsymbol{E}_N]\!]}{Z^+ + Z^-}\right)^z \Bigg|_{\partial\Omega^k},
\end{aligned} \tag{2.14}$$

$$\begin{aligned}
\frac{\mathrm{d}H_N^x}{\mathrm{d}t} &= (\varepsilon_\mathrm{r}\mathbf{M})^{-1}\left(\mathbf{S}^z E_N^y - \mathbf{S}^y E_N^z\right) \\
&\quad + (\varepsilon_\mathrm{r}\mathbf{M})^{-1}\,\mathbf{F}\left(\hat{\boldsymbol{n}} \times \frac{Y^+ [\![\boldsymbol{E}_N]\!] + \hat{\boldsymbol{n}} \times [\![\boldsymbol{H}_N]\!]}{Y^+ + Y^-}\right)^x \Bigg|_{\partial\Omega^k}, \\
\frac{\mathrm{d}H_N^y}{\mathrm{d}t} &= (\varepsilon_\mathrm{r}\mathbf{M})^{-1}\left(\mathbf{S}^x E_N^z - \mathbf{S}^z E_N^x\right) \\
&\quad + (\varepsilon_\mathrm{r}\mathbf{M})^{-1}\,\mathbf{F}\left(\hat{\boldsymbol{n}} \times \frac{Y^+ [\![\boldsymbol{E}_N]\!] + \hat{\boldsymbol{n}} \times [\![\boldsymbol{H}_N]\!]}{Y^+ + Y^-}\right)^y \Bigg|_{\partial\Omega^k}, \\
\frac{\mathrm{d}H_N^z}{\mathrm{d}t} &= (\varepsilon_\mathrm{r}\mathbf{M})^{-1}\left(\mathbf{S}^y E_N^x - \mathbf{S}^x E_N^y\right) \\
&\quad + (\varepsilon_\mathrm{r}\mathbf{M})^{-1}\,\mathbf{F}\left(\hat{\boldsymbol{n}} \times \frac{Y^+ [\![\boldsymbol{E}_N]\!] + \hat{\boldsymbol{n}} \times [\![\boldsymbol{H}_N]\!]}{Y^+ + Y^-}\right)^z \Bigg|_{\partial\Omega^k}.
\end{aligned} \tag{2.15}$$

Here the fields E_N^x, E_N^y, E_N^z, H_N^x, H_N^y, and H_N^z represent the discrete counterparts of *scalar* fields. That is the reason they are not typeset boldface, despite now being in fact vectors as a result of the discretisation. In contrast, we evaluate the numerical flux in the right-hand side of (2.14)–(2.15) at each node at the boundary of the element using the discrete counterparts of *vector* fields. Then at each node the corresponding component of the resulting vector is taken.

The advantages of the above described discretisation are discussed in detail in [42] and [81], where a number of numerical examples are also provided. Here it suffices to mention its optimal flexibility for mesh refinement, the possibility of independent adjustment of polynomial orders in each element (*hp*-adaptation), its

excellent performance on parallel computers and that only matrix-matrix multiplications are needed during the time integration. In this chapter, however, our aim is to analyse the properties of time -integration methods suitable for this spatial DG discretisation, therefore we assemble the local semi-discrete system (2.14)–(2.15) into a global matrix and consider the 'abstract' semi-discrete system

$$\frac{\mathrm{d}\boldsymbol{q}_h}{\mathrm{d}t} = \boldsymbol{\mathcal{A}}\boldsymbol{q}_h, \tag{2.16}$$

where $\boldsymbol{\mathcal{A}}$ is the global matrix and $\boldsymbol{q}_h = [\boldsymbol{E}_h, \boldsymbol{H}_h]^T$ represents the numerical approximation to the fields in the complete domain. The matrix assembly is somewhat lengthy but straightforward, and it follows the standard procedure. See [52] for example.

2.4 Runge-Kutta time-stepping methods

From the point of view of time integration, one of the main difficulties in high-order spectral/*hp* element methods is the restriction on the time step of explicit time-integration schemes. For hyperbolic systems in general, and for the advection equation in particular, it is known (see [52], for example) that the maximum eigenvalue of the semi-discrete global matrix grows as $\mathcal{O}(p^2)$ with polynomial order p, hence the time step is usually bounded by $\mathcal{O}(1/p^2)$. The time-step restriction then can generally be taken as

$$\Delta t \leq \Delta t_{\max} = \mathrm{CFL}(p)\frac{h_k}{c_k}, \tag{2.17}$$

where h_k is the minimum edge length of all elements and c_k is the maximum wave speed in the domain. Here the parameter CFL depends on the degree of the polynomials used in the spatial discretisation. If we apply any given time-integration scheme with fixed order (i.e. independent of the polynomial order p) to the semi-discrete system (2.16), we have

$$\mathrm{CFL}(p) = \mathrm{C}\,\frac{1}{p^2}, \tag{2.18}$$

where C is a constant, typically of order one. This condition may turn out to be rather restrictive as we go to higher and higher order approximations, even with a slightly increasing value for C (see Section 2.6).

The low-storage Runge-Kutta schemes introduced in [16] are among the most popular choices for time integration of the DG space-discretised Maxwell equations. Storage can be essential for large-scale computations and low-storage schemes require only two storage units per ODE variable. If we consider the ODE system

$$\frac{\mathrm{d}u}{\mathrm{d}t} = L(u), \tag{2.19}$$

Table I: *Coefficients of the fourth-order five-stage low-storage Runge-Kutta method*

i	a_i	b_i
1	0	0.14965902199923
2	−0.41789047449985	0.37921031299963
3	−1.19215169464268	0.82295502938698
4	−1.69778469247153	0.69945045594912
5	−1.51418344425716	0.15305724796815

the general m-stage low-storage Runge-Kutta scheme [84, 16] can be written in the form

$$\begin{aligned} &u^{(0)} = u^n, \quad v^{(0)} = 0, \\ &v^{(i)} = a_i v^{(i-1)} + \Delta t L(u^{(i-1)}), \quad i = 1, \ldots, m, \\ &u^{(i)} = u^{(i-1)} + b_i v^{(i)}, \qquad\quad i = 1, \ldots, m, \\ &u^{n+1} = u^{(m)}, \end{aligned} \tag{2.20}$$

where only u and an auxiliary variable v must be stored. The coefficients a_i and b_i have been determined for a number of different low-storage Runge-Kutta schemes. See [16] and [30] for more details. In this chapter we consider the fourth-order five-stage low-storage scheme also applied in [42]. The coefficients we use are listed in Table I.

One possible way to achieve a weaker time-step restriction is the application of SSP-RK schemes. In [31] it was shown that for the linear autonomous system (2.19) the class of m-stage linear SSP-RK schemes, given recursively by

$$\begin{aligned} u^{(0)} &= u^n, \\ u^{(i)} &= u^{(i-1)} + \Delta t L u^{(i-1)}, \qquad i = 1, \ldots, m-1 \\ u^{(m)} &= \sum_{k=0}^{m-2} \alpha_{m,k} u^{(k)} + \alpha_{m,m-1} \left(u^{(m-1)} + \Delta t L u^{(m-1)} \right), \\ u^{n+1} &= u^{(m)} \end{aligned} \tag{2.21}$$

where $\alpha_{1,0} = 1$ and

$$\begin{aligned} \alpha_{m,k} &= \frac{1}{k} \alpha_{m-1,k-1}, \qquad k = 1, \ldots, m-2 \\ \alpha_{m,m-1} &= \frac{1}{m!}, \qquad \alpha_{m,0} = 1 - \sum_{k=1}^{m-1} \alpha_{m,k}, \end{aligned}$$

are mth-order accurate. This was extended to linear non-autonomous systems by Chen et al. [17]. In that work the authors demonstrated that when applied together

with the classical discontinuous Galerkin method [20, 22], the SSP-RK scheme gives $(p+1)$st-order convergence with the stability bound

$$\mathrm{CFL}(p) = \mathrm{C}\,\frac{1}{2p+1} \tag{2.22}$$

with $\mathrm{C} = 1$, as long as for a given spatial discretisation of polynomial order p, the corresponding SSP-RK method has order $p+1$. The stability regions of several SSP-RK methods and the low-storage five-stage fourth-order Runge-Kutta method are displayed in Figure 2.2.

2.5 Analysis of the dispersion and dissipation error

A critical factor in the numerical simulation of wave-propagation is the artificial dissipation and/or dispersion inflicted on the waves due to numerical discretisation errors. In order to analyse these properties of the different schemes, we resort to the one-dimensional and two-dimensional forms of the Maxwell equations with periodic boundary conditions. First, these reduced models are formulated and then we perform a numerical Fourier analysis of the fully discrete schemes to investigate the dispersion and dissipation errors as a function of mesh size per wave length and time step. This analysis provides important information on the accuracy of the schemes regarding wave motion and the relation between time step, mesh size and polynomial order.

2.5.1 Wave equation in one and two dimensions

The Maxwell equations in one dimension read

$$\varepsilon_\mathrm{r}\frac{\partial E}{\partial t} = -\frac{\partial H}{\partial x}, \qquad \mu_\mathrm{r}\frac{\partial H}{\partial t} = -\frac{\partial E}{\partial x}, \tag{2.23}$$

or they can be expressed by the wave equation

$$\frac{\partial^2 E}{\partial t^2} - \frac{1}{\varepsilon_\mathrm{r}\mu_\mathrm{r}}\frac{\partial^2 E}{\partial x^2} = 0,$$

in the domain $\Omega \subset \mathbb{R}$. In conservative form (2.7), this reads

$$Q\frac{\partial \boldsymbol{q}}{\partial t} + \nabla \cdot \boldsymbol{F}(\boldsymbol{q}) = 0 \tag{2.24}$$

with

$$Q = \mathrm{diag}(\varepsilon_\mathrm{r}, \mu_\mathrm{r}), \quad \boldsymbol{q} = \begin{bmatrix} E \\ H \end{bmatrix}, \quad \boldsymbol{F}(\boldsymbol{q}) = \begin{bmatrix} H \\ E \end{bmatrix}.$$

For the two-dimensional analysis we take the transverse magnetic (TM) polarisation of the Maxwell equations

$$\mu_\mathrm{r}\frac{\partial H^x}{\partial t} = -\frac{\partial E^z}{\partial y},$$

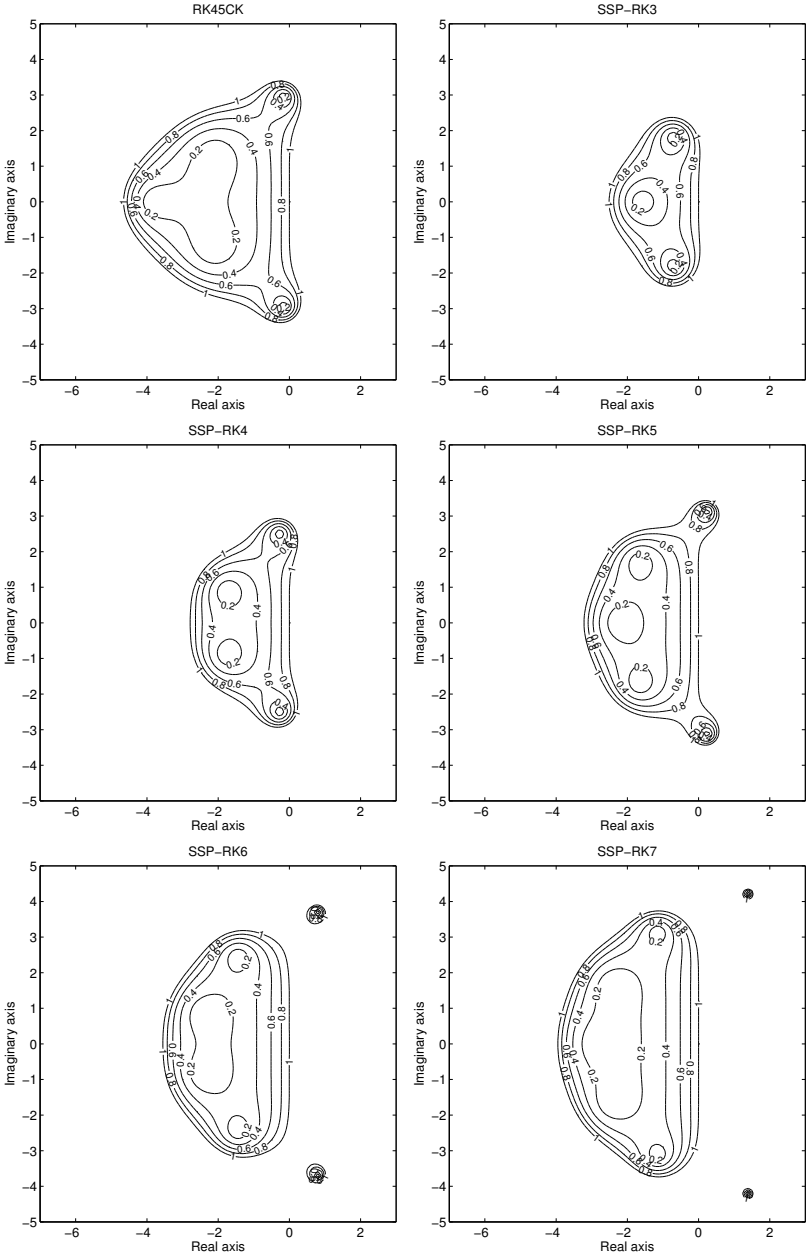


Figure 2.2: *Stability regions for the five-stage fourth-order Runge-Kutta method (top left) and for five different SSP-RK methods of order* $m = 3, 4, 5, 6, 7$

$$\mu_{\mathrm{r}} \frac{\partial H^y}{\partial t} = \frac{\partial E^z}{\partial x},$$
$$\varepsilon_{\mathrm{r}} \frac{\partial E^z}{\partial t} = \frac{\partial H^y}{\partial x} - \frac{\partial H^x}{\partial y},$$

which is again equivalent to the second-order wave equation,

$$\frac{\partial^2 E^z}{\partial t^2} - \frac{1}{\varepsilon_{\mathrm{r}} \mu_{\mathrm{r}}} \nabla^2 E^z = 0.$$

Thus we arrive at the first-order system (2.7)

$$Q \frac{\partial \boldsymbol{q}}{\partial t} + \nabla \cdot \boldsymbol{F}(\boldsymbol{q}) = 0, \tag{2.25}$$

with

$$Q = \mathrm{diag}(\mu_{\mathrm{r}}, \mu_{\mathrm{r}}, \varepsilon_{\mathrm{r}}), \quad \boldsymbol{q} = \begin{bmatrix} H^x \\ H^y \\ E^z \end{bmatrix}, \quad \boldsymbol{F}(\boldsymbol{q}) = \begin{bmatrix} 0 & -E^z \\ E^z & 0 \\ H^y & -H^x \end{bmatrix}.$$

2.5.2 Dispersion and dissipation analysis of the global scheme

For the analysis of the fully discrete schemes, we consider (2.24) and (2.25) in the domains $\Omega = [-1, 1]$ and $\Omega = [-1, 1]^2$, respectively. Furthermore, we use uniform meshes and assume that the boundaries are periodic.

One-dimensional Fourier analysis

We are primarily interested in wave propagation and in the associated dispersion and dissipation error of the RKDG scheme. We consider the one-dimensional semi-discrete scheme (2.16)

$$\frac{\mathrm{d}\boldsymbol{q}_h}{\mathrm{d}t} = \boldsymbol{\mathcal{A}} \boldsymbol{q}_h$$

in the domain $\Omega = [-1, 1]$ filled with vacuum (or air) and assume a monochromatic plane wave (which is also a Fourier mode)

$$\boldsymbol{q}_h(0) = \boldsymbol{q}_h^0 = \left[\mathrm{e}^{\mathrm{i}kx_h}, \ \mathrm{e}^{\mathrm{i}kx_h} \right]^T$$

as initial condition. Here, x_h represents the vector of the nodes used for the spatial discretisation. We denote the angular wave frequency with ω. The exact wave number k is given by the dispersion relation $k^2 = \omega^2 / c^2$, with c being the speed of light. The time-exact discrete Fourier mode at time level $t_n = n\Delta t$ will read

$$\boldsymbol{q}_h(n\Delta t) = \nu^n \boldsymbol{q}_h(0) = \mathrm{e}^{-\mathrm{i}\omega n \Delta t} \left[\mathrm{e}^{\mathrm{i}kx_h}, \ \mathrm{e}^{\mathrm{i}kx_h} \right]^T \tag{2.26}$$

with exact amplification factor $\nu^n = \mathrm{e}^{-\mathrm{i}\omega n\Delta t}$ and $\mathrm{i}^2 = -1$. To see the effect of the time-stepping method, we replace the exact amplification factor ν^n with its discrete counterpart ν_h^n and take the fully discrete Fourier mode as

$$\boldsymbol{q}_h^n = \nu_h^n \boldsymbol{q}_h^0 = \nu_h^n \left[\mathrm{e}^{\mathrm{i}kx_h},\ \mathrm{e}^{\mathrm{i}kx_h}\right]^T. \tag{2.27}$$

In addition, we write the SSP-RK scheme as a two-level explicit scheme. Thus

$$\boldsymbol{q}_h^{n+1} = \boldsymbol{\mathcal{B}}\boldsymbol{q}_h^n \tag{2.28}$$

holds with amplification matrix

$$\boldsymbol{\mathcal{B}} = \sum_{l=0}^{m} \frac{1}{l!} \left(\Delta t \boldsymbol{\mathcal{A}}\right)^l \tag{2.29}$$

and with m being the order of the SSP-RK time-stepping scheme. Substituting (2.27) into (2.28) results in the equation

$$\nu_h^{n+1} \left[\mathrm{e}^{\mathrm{i}kx_h},\ \mathrm{e}^{\mathrm{i}kx_h}\right]^T = \boldsymbol{\mathcal{B}}\nu_h^n \left[\mathrm{e}^{\mathrm{i}kx_h},\ \mathrm{e}^{\mathrm{i}kx_h}\right]^T,$$

which, after division with ν_h^n, reduce to the eigenvalue problem

$$\nu_h \boldsymbol{q}_h^0 = \boldsymbol{\mathcal{B}}\boldsymbol{q}_h^0. \tag{2.30}$$

Solving this eigenvalue equation will produce $p+1$ different values for $\nu_{h,j}$ (and as many corresponding eigenvectors $\boldsymbol{q}_{h,j}^0$). Bearing in mind that

$$\nu_{h,j} = \mathrm{e}^{-\mathrm{i}\tilde{\omega}_{h,j}\Delta t},$$

with *complex* numerical frequencies $\tilde{\omega}_{h,j}$, we can establish the dispersion and dissipation properties of the scheme. For that, we consider the real (for dispersion) and imaginary (for dissipation) parts of the complex numerical frequencies $\tilde{\omega}_{h,j} = (\mathrm{i}/\Delta t)\ln\nu_{h,j}$, that is

$$\omega_{h,j} = \mathrm{Re}\left[\tilde{\omega}_{h,j}\right] = \mathrm{Re}\left[(\mathrm{i}/\Delta t)\ln\nu_{h,j}\right], \quad \rho_{h,j} = \mathrm{Im}\left[\tilde{\omega}_{h,j}\right] = \mathrm{Im}\left[(\mathrm{i}/\Delta t)\ln\nu_{h,j}\right],$$

with *real* numerical frequency $\omega_{h,j}$ and numerical dissipation $\rho_{h,j}$, both corresponding to the eigenvalue $\nu_{h,j}$. One of the computed modes will be close to the frequency of the physical mode. This represents the approximation properties of our scheme. The other modes are spurious. To decide which eigenvalue should be considered, we define the dissipation error as $\mathrm{err}_{h,j}^{\mathrm{diss}} = |\rho_{h,j}| - 1$ and the dispersion error as the absolute value of dispersion error $\mathrm{err}_{h,j}^{\mathrm{disp}} = |\omega - \omega_{h,j}|$. The numerical Fourier mode is now taken as the closest eigenvalue to the physical mode

$$\nu_h := \left\{ \nu_{h,j} \colon \min_j \sqrt{\left(\mathrm{err}_{h,j}^{\mathrm{disp}}\right)^2 + \left(\mathrm{err}_{h,j}^{\mathrm{diss}}\right)^2} \right\}. \tag{2.31}$$

In the numerical dispersion and dissipation analysis of the fully discrete schemes, we consider a wide range of values for Δt and h_k. The eigenvalues of (2.30) are computed in `Matlab`.

It is also important to consider the convergence of the numerical dispersion and dissipation error. In [1] a complete dispersion and dissipation analysis of the semi-discrete advection equation was carried out for discontinuous Galerkin methods with high-order tensor-product elements. In that chapter it was proven that in the asymptotic region $hk = \frac{2\pi h}{\lambda} \to 0$ (with λ being the wave length) the dispersion relation for a pth-order method is accurate to order $2p+3$ for the dispersion error and order $2p+2$ for the dissipation error. See [49] and [1] for more details. In a more recent work [3] the dispersion and dissipation analysis of the semi-discrete wave equation was provided for some low-order schemes (linear elements for the general DG and up to third order elements for the IP-DG) and the authors also conjectured on how the results would extend to arbitrary order elements. For the fully discrete system we consider here, the rate of convergence is also influenced by the time-stepping method. However, for most polynomial orders p, the convergence of the dispersion and dissipation error still by far supersedes that of the error measured in the l^2-norm. The numerical results are discussed in Section 2.6.

Two-dimensional Fourier analysis

As in one dimension, we assemble our right-hand side into a global matrix and consider the abstract Cauchy problem (2.16)

$$\frac{\mathrm{d}\boldsymbol{q}_h}{\mathrm{d}t} = \boldsymbol{\mathcal{A}}\boldsymbol{q}_h$$

where now $\boldsymbol{q} = [H_h^x, H_h^y, E_h^z]^T$ and the matrix $\boldsymbol{\mathcal{A}}$ represents the semi-discrete system resulting from the discretisation of (2.25). The only difference in the mathematical formulation to the one-dimensional case is that the dispersion relation now reads $k_x^2 + k_y^2 = \omega^2/c^2$. The time-exact discrete Fourier mode satisfying (2.25) is equal to

$$\boldsymbol{q}_h^n = \nu^n \left[(k_y/\omega)\, \mathrm{e}^{\mathrm{i}k_x x_h + \mathrm{i}k_y y_h},\ (-k_x/\omega)\, \mathrm{e}^{\mathrm{i}k_x x_h + \mathrm{i}k_y y_h},\ \mathrm{e}^{\mathrm{i}k_x x_h + \mathrm{i}k_y y_h} \right]^T \tag{2.32}$$

in the two-dimensional domain $\Omega = [-1,1]^2$. From here we follow exactly the same line as in the one-dimensional case. As initial condition we take a monochromatic plane wave with different wave numbers k_y and k_x between which the relation $k_y = 2k_x$ always holds. This represents a monochromatic plane wave travelling at an angle of about $26.565°$ against the x axis. As in one dimension, a range of values of Δt and h_k are considered. We note that for computing the matrix exponential in (2.29) the simple Horner's rule is applied (see [28] for example).

2.6 Numerical results

2.6.1 One-dimensional cavity

In order to investigate if the SSP-RK scheme retains the high-order convergence of the spatial discretisation [42], we consider the one-dimensional cavity problem in the domain $x \in [-1, 1]$, with two different non-magnetic ($\mu_{\mathrm{r},1} = \mu_{\mathrm{r},2} = 1$) materials. The material interface is situated at $x = 0$ and the two different materials have a relative permittivity of $\varepsilon_{\mathrm{r},1} = 1$ and $\varepsilon_{\mathrm{r},2} = 2.25$, respectively. The error is measured against the exact solution, which is included in the Appendix. We set the frequency of the wave at $\omega = 2\pi$, the same throughout the whole domain, which entails the corresponding wave numbers $k_1 = 2\pi$ and $k_2 = 3\pi$, respectively. In Table II we show the l_2-error $\|\boldsymbol{q}_N - \boldsymbol{q}_{\mathrm{exact}}\|$ at final time $T = 1$ for different orders of the local polynomials. For the time integration, we use the $(p+1)$st-order SSP-RK method (2.4) with corresponding maximum time step (2.22). We can see that in the asymptotic region $hk \ll 1$, $(p+1)$st-order convergence is achieved for polynomial orders p in the range of $1 \leq p \leq 9$.

In the next example we compare the performance of the two different time-integration schemes defined in Section 2.4. We also include the ubiquitous standard fourth-order RK scheme as a reference. The constant in the time-step restriction (2.22) of the SSP-RK schemes is set to $\mathrm{C} = 1$ for all values of polynomial order p. For the low-storage Runge-Kutta scheme and the standard Runge-Kutta scheme, we use the value $\mathrm{C} = 1$ in (2.18) for $p \leq 5$, and the value $\mathrm{C} = 2$ when $5 < p \leq 10$. We consider the same cavity, but now filled with vacuum (or air) and integrate for a relatively long time, until $T = 1000$ time periods. The results are shown in Table III for different orders and different number of elements. We measure the computational work as the number of operations, which is simply computed as

$$\mathtt{ops} = N_{\mathrm{T}}\, m,$$

where N_T is the number of time steps needed until final time T, and m represents the number of stages in a given RK scheme. For each RK scheme we use the corresponding maximum time step defined in Section 2.4. The test was carried out for orders $p = 3$, $p = 6$ and $p = 10$. Significant differences in accuracy, up to $\mathcal{O}(4)$, occur between the schemes for orders $p = 6$ and $p = 10$. For each order, we include two subtables, one for a relatively fine spatial grid and one for a comparatively coarse one. The most favourable characteristic of the SSP-RK schemes is that the better accuracy occurs without significant–or indeed, any–increase in the number of operations.

Perhaps even more illuminating is to see how much computational work is needed to obtain a given accuracy. In Table IV we show the number of operations necessary to achieve the error $l^2_{\mathrm{err}} = 10^{-5}$ at final time $T = 1000$. The comparisons of the different time-integration schemes are made for fixed polynomial order and spatial mesh, so that only the time step was lowered in order to decrease the errors.

Table II: *Convergence of the global error for the one-dimensional metallic cavity filled with two different materials. In each case the $(p+1)$st-order SSP-RK scheme is applied.*

	$p=1$		$p=2$		$p=3$	
	l^2-error	Order	l^2-error	Order	l^2-error	Order
$N_{\text{el}}=2$	1.7557E-00		2.5843e+00		1.3544E-00	
$N_{\text{el}}=4$	1.5732E-00	1.5840	7.9723E-01	1.6967	2.0100E-01	2.7524
$N_{\text{el}}=8$	8.4106E-01	9.0339	1.0048E-01	2.9880	1.7313E-02	3.5372
$N_{\text{el}}=16$	1.9518E-01	2.1074	1.4226E-02	2.8204	1.1295E-03	3.9382
$N_{\text{el}}=32$	3.8904E-02	2.3268	1.8855E-03	2.9155	6.8300E-05	4.0476
$N_{\text{el}}=64$	9.0807E-03	2.0990	2.3897E-04	2.9800	4.3243E-06	3.9813
$N_{\text{el}}=128$	2.2310E-03	2.0251	2.9985E-05	2.9945	2.7049E-07	3.9988
$N_{\text{el}}=256$	5.5755E-04	2.0005	3.7547E-06	2.9975	1.6909E-08	3.9997
$N_{\text{el}}=512$	1.3944E-04	1.9995	4.6972E-07	2.9988	1.0568E-09	4.0000
	$p=4$		$p=5$		$p=6$	
	l^2-error	Order	l^2-error	Order	l^2-error	Order
$N_{\text{el}}=2$	7.7746E-01		2.7975E-01		1.6624E-01	
$N_{\text{el}}=4$	5.7034E-02	3.7689	1.0551E-02	4.7286	1.8202E-03	6.5130
$N_{\text{el}}=8$	1.9711E-03	4.8548	2.0173E-04	5.7089	1.6440E-05	6.7907
$N_{\text{el}}=16$	6.5391E-05	4.9138	3.1504E-06	6.0007	1.3514E-07	6.9267
$N_{\text{el}}=32$	2.0736E-06	4.9789	4.9661E-08	5.9873	1.0574E-09	6.9978
$N_{\text{el}}=64$	6.4733E-08	5.0015	7.7594E-10	6.0000	8.3515E-12	6.9843
$N_{\text{el}}=128$	1.9895E-09	5.0240	1.2161E-11	5.9956		
$N_{\text{el}}=256$	6.2218E-11	4.9989	2.7547E-13	5.4642		
	$p=7$		$p=8$		$p=9$	
	l^2-error	Order	l^2-error	Order	l^2-error	Order
$N_{\text{el}}=2$	1.4124E-02		1.5533E-02		7.3602E-04	
$N_{\text{el}}=4$	2.7117E-04	5.7028	3.7314E-05	8.7014	4.4752E-06	7.3616
$N_{\text{el}}=8$	1.2568E-06	7.7533	8.2859E-08	8.8149	4.9118E-09	9.8315
$N_{\text{el}}=16$	5.0129E-09	7.9699	1.6508E-10	8.9714	4.9240E-12	9.9622
$N_{\text{el}}=32$	1.9730E-11	7.9891	3.2760E-13	8.9770		

2.6.2 Numerical dispersion and dissipation error

To conduct the dispersion and dissipation analysis described in Section 2.5, we carry out two types of experiments. First, we consider the convergence of the dispersion and dissipation errors in one dimension. The polynomials we apply for the spatial discretisation range from $p=1$ to $p=10$, and for the time discretisation we use the corresponding $(p+1)$st-order SSP-RK scheme (2.4) with maximum time step (2.22). Because the actual errors may be rather small for large values of p, we increase the wave number (thus decrease the wave length) for higher-order polynomials. Consequently, the following wave numbers are used in our one-dimensional Fourier

Table III: *Errors and computational work when different time-stepping schemes are applied to the cavity problem with polynomial order p, number of elements* N_{el} *and degrees of freedom* DoF *after integrating over* $T = 1000$ *time periods*

$p = 3$, $N_{\text{el}} = 40$ (DoF $= 160$)	$\Delta t = \Delta t_{\max}$	ops	l^∞-error	l^2-error
SSP-RK4	7.1428E-03	560004	4.2370E-04	2.1220E-04
Carpenter&Kennedy	5.5555E-03	900005	6.2719E-05	3.1442E-05
Standard RK4	5.5555E-03	720004	1.5559E-04	7.7676E-05
$p = 3$, $N_{\text{el}} = 20$ (DoF $= 80$)	$\Delta t = \Delta t_{\max}$	ops	l^∞-error	l^2-error
SSP-RK4	1.4286E-02	280004	6.6828E-03	3.3938E-03
Carpenter&Kennedy	1.1111E-02	450005	9.5898E-04	5.2903E-04
Standard RK4	1.1111E-02	360004	2.4328E-03	1.2492E-03
$p = 6$, $N_{\text{el}} = 12$ (DoF $= 84$)	$\Delta t = \Delta t_{\max}$	ops	l^∞-error	l^2-error
SSP-RK7	1.2820E-02	546007	1.5364E-07	6.4372E-08
Carpenter&Kennedy	9.2592E-03	540005	4.7966E-04	2.3993E-04
Standard RK4	9.2592E-03	432004	1.1982E-03	5.9976E-04
$p = 6$, $N_{\text{el}} = 6$ (DoF $= 42$)	$\Delta t = \Delta t_{\max}$	ops	l^∞-error	l^2-error
SSP-RK7	2.5640E-02	273007	1.7646E-05	8.0618E-06
Carpenter&Kennedy	1.8518E-02	270005	7.6505E-03	3.8385E-03
Standard RK4	1.8518E-02	216004	1.9066E-02	9.5896E-03
$p = 10$, $N_{\text{el}} = 4$ (DoF $= 44$)	$\Delta t = \Delta t_{\max}$	ops	l^∞-error	l^2-error
SSP-RK11	2.3810E-02	462000	1.9624E-08	9.7129E-09
Carpenter&Kennedy	1.0000E-02	500000	6.5246E-04	2.6565E-04
Standard RK4	1.0000E-02	400000	1.6297E-03	6.6450E-04
$p = 10$, $N_{\text{el}} = 2$ (DoF $= 22$)	$\Delta t = \Delta t_{\max}$	ops	l^∞-error	l^2-error
SSP-RK11	4.7619E-02	231000	3.3137E-06	3.2341E-06
Carpenter&Kennedy	2.0000E-02	250000	1.0199E-02	4.5032E-03
Standard RK4	2.0000E-02	200000	2.5387E-02	1.1272E-02

analysis:

$$k = \begin{cases} \pi & \text{if } p = 1, 2, \\ 2\pi & \text{if } p = 3, 4, \\ 4\pi & \text{if } p = 5, 6, 7, \\ 8\pi & \text{if } p = 8, 9, 10. \end{cases}$$

The errors, defined in Section 2.5, and the rate of the convergence are shown in Table V for the dispersion and in Table VI for the dissipation. Although we cannot establish precise convergence rates due the influence of the time discretisation on the dispersion and dissipation properties, an order of convergence of approximately

Table IV: *Computational work needed to achieve at least* $l^2_{\text{err}} = 10^{-5}$ *accuracy when different time-stepping schemes are applied to the cavity problem with polynomial order* p*, number of elements* N_{el} *and degrees of freedom* DoF *after integrating over* $T = 1000$ *time periods*

$p = 3$, $N_{\text{el}} = 54$ (DoF $= 216$)	Δt	ops	l^2-error
SSP-RK4	3.2922E-03	1215004	9.7349E-06
Carpenter&Kennedy	4.1152E-03	1215005	9.4976E-06
Standard RK4	3.2922E-03	1215004	9.7349E-06
$p = 6$, $N_{\text{el}} = 6$ (DoF $= 42$)	Δt	ops	l^2-error
SSP-RK7	2.5640E-02	273007	8.0618E-06
Carpenter&Kennedy	3.7037E-03	1350005	9.5547E-06
Standard RK4	2.7778E-03	1440004	8.9196E-06
$p = 10$, $N_{\text{el}} = 2$ (DoF $= 22$)	Δt	ops	l^2-error
SSP-RK11	4.7619E-02	231000	3.2341E-06
Carpenter&Kennedy	4.0000E-03	1250000	7.7825E-06
Standard RK4	3.0000E-03	1333336	6.4107E-06

$2p$ is achieved for both the dissipation and dispersion error. Comparing these results to the findings of [1] and [3] already implies that the SSP-RK time integration has a less dominant effect on the dispersion and dissipation errors.

In the second experiment, we consider a wide range of time steps Δt and mesh sizes h, and examine the dispersion and dissipation errors of the schemes as a function of wave length per mesh size (λ/h), degrees of freedom per wave length (DoF/λ) and relative time step ($\Delta t/\Delta t_{\max}$). The value $\Delta t/\Delta t_{\max} = 1$ indicates the size of the maximum time step, defined as in (2.22) in Section 2.4. The contour plots of the dispersion error (or dispersion error) for the one-dimensional analysis with wave number $k = 2\pi$ is shown in Figure 2.3 for orders $p = 1, 2$. The same plots for the dissipation error are displayed in Figure 2.4. For this range of values the dispersion error is generally at least one order higher than the dissipation error, and it can be improved by both taking smaller time steps and/or refining the spatial grid.

We carry out the same experiment in two dimensions on a uniform grid with wave numbers $k_x = \pi$ and $k_y = 2\pi$ for orders $p = 1, 2, 3, 4, 5$ and $k_x = 2\pi$ and $k_y = 4\pi$ for order $p = 6$. The number of elements used to construct the uniform meshes range from 200 elements to 8 elements for $p = 1, 2, 3, 4$; from 128 elements to 8 elements for $p = 5$; and from 72 elements to 8 elements for $p = 6$. The contour plots of the dispersion and dissipation errors are shown in Figure 2.5 and in Figure 2.6, respectively. Note that the studied range of wave lengths per mesh size differs significantly from that of the one-dimensional case and occasionally from one another. It can be deduced that the spatial discretisation dominates the dispersion error, which is all the more relevant because the dispersion error is at least one

Table V: *Convergence of the dispersion error for the one-dimensional wave equation in a periodic domain with wave numbers $k = \pi$ for $p = 1, 2$; $k = 2\pi$ for $p = 3, 4$; $k = 4\pi$ for $p = 5, 6, 7$; and $k = 8\pi$ for $p = 8, 9, 10$. In each case the $(p+1)$st-order SSP-RK scheme is applied.*

	$p=1$		$p=2$			
	Error	Order	Error	Order		
$N_{\text{el}} = 2$	1.1219E-00		5.4535E-02			
$N_{\text{el}} = 4$	1.8866E-01	2.5721	1.9858E-03	4.7794		
$N_{\text{el}} = 8$	3.9271E-02	2.2642	8.0576E-05	4.6232		
$N_{\text{el}} = 16$	9.1970E-03	2.0942	4.2539E-06	4.2435		
$N_{\text{el}} = 32$	2.2573E-03	2.0266	2.5327E-07	4.0700		
$N_{\text{el}} = 64$	5.6163E-04	2.0069	1.5631E-08	4.0182		
	$p=3$		$p=4$			
	Error	Order	Error	Order		
$N_{\text{el}} = 2$	2.8401E-01		4.3239E-02			
$N_{\text{el}} = 4$	1.8427E-03	7.2680	8.7429E-05	8.9500		
$N_{\text{el}} = 8$	1.1103E-04	4.0528	9.1380E-08	9.9020		
$N_{\text{el}} = 16$	8.1803E-06	3.7626	3.1783E-09	4.8456		
$N_{\text{el}} = 32$	5.1772E-07	3.9819	5.1479E-11	5.9481		
$N_{\text{el}} = 64$	3.2403E-08	3.9980	8.0824E-13	5.9930		
	$p=5$		$p=6$		$p=7$	
	Error	Order	Error	Order	Error	Order
$N_{\text{el}} = 2$	2.2250E-00		6.2763E-01		1.4307E-01	
$N_{\text{el}} = 4$	7.5731E-03	8.1987	4.3307E-04	10.501	1.8263E-05	12.936
$N_{\text{el}} = 8$	4.8796E-06	10.600	4.9182E-08	13.104	3.2505E-10	15.778
$N_{\text{el}} = 16$	2.1995E-08	7.7934	1.5744E-11	11.609	4.6185E-13	9.4590
$N_{\text{el}} = 32$	3.2993E-10	6.0589				
$N_{\text{el}} = 64$	5.1621E-12	5.9981				
	$p=8$		$p=9$		$p=10$	
	Error	Order	Error	Order	Error	Order
$N_{\text{el}} = 2$	3.8264E-00		8.3880E-01		1.5821E-00	
$N_{\text{el}} = 4$	4.9854E-02	6.2621	6.5271E-03	7.0057	6.6177E-04	11.223
$N_{\text{el}} = 8$	1.1951E-06	15.348	3.1177E-08	17.676	6.6458E-10	19.925
$N_{\text{el}} = 16$	6.6187E-12	17.462				

order higher than the dissipation error for all polynomial orders considered. Another important feature–shown in Figure 2.9–is the decreasing number of degrees of freedom needed to obtain a given accuracy as we go to higher-order elements. For instance, to attain a dispersion error of 10^{-3}, we need about 14-15 degrees of freedom per wave length for discretisation with second-order polynomials and

Table VI: *Convergence of the dissipation error for the one-dimensional wave equation in a periodic domain with wave numbers $k = \pi$ for $p = 1, 2$; $k = 2\pi$ for $p = 3, 4$; $k = 4\pi$ for $p = 5, 6, 7$; and $k = 8\pi$ for $p = 8, 9, 10$. In each case the $(p+1)$st-order SSP-RK scheme is applied.*

	$p=1$		$p=2$			
	Error	Order	Error	Order		
$N_{\text{el}} = 2$	2.5464E-01		2.1860E-02			
$N_{\text{el}} = 4$	1.6710E-02	3.9297	7.5294E-04	4.8596		
$N_{\text{el}} = 8$	1.1239E-03	3.8941	3.1441E-05	4.5818		
$N_{\text{el}} = 16$	7.2499E-05	3.9544	1.6831E-06	4.2235		
$N_{\text{el}} = 32$	4.5729E-06	3.9868	1.0063E-07	4.0640		
$N_{\text{el}} = 64$	2.8649E-07	3.9966	6.2172E-09	4.0166		
	$p=3$		$p=4$			
	Error	Order	Error	Order		
$N_{\text{el}} = 2$	6.1234E-02		6.2971E-03			
$N_{\text{el}} = 4$	6.8779E-04	6.4762	1.4252E-05	8.7873		
$N_{\text{el}} = 8$	4.2695E-06	7.3318	1.8226E-08	9.6110		
$N_{\text{el}} = 16$	2.8124E-08	7.2461	5.9037E-10	4.9482		
$N_{\text{el}} = 32$	2.7328E-10	6.6852	9.5566E-12	5.9490		
$N_{\text{el}} = 64$	3.6039E-12	6.2447	1.5099E-13	5.9840		
	$p=5$		$p=6$		$p=7$	
	Error	Order	Error	Order	Error	Order
$N_{\text{el}} = 2$	1.7209E-01		3.8318E-02		8.3102E-03	
$N_{\text{el}} = 4$	5.4651E-04	8.2987	3.2754E-05	10.192	1.4167E-06	12.518
$N_{\text{el}} = 8$	2.8854E-07	10.887	4.0972E-09	12.965	3.7758E-11	15.195
$N_{\text{el}} = 16$	5.7239E-11	12.300	1.3922E-12	11.523		
	$p=8$		$p=9$		$p=10$	
	Error	Order	Error	Order	Error	Order
$N_{\text{el}} = 2$	8.7179E-02		1.1417E-01		4.0277E-02	
$N_{\text{el}} = 4$	1.5076E-03	5.8536	2.0997E-04	9.0867	2.2413E-05	10.811
$N_{\text{el}} = 8$	4.6922E-08	14.972	1.2343E-09	17.376	2.6468E-11	19.692
$N_{\text{el}} = 16$	2.7467E-13	17.382				

about six for the discretisation with sixth-order polynomials.

One of the main advantages of DG methods is that they are relatively insensitive to the uniformity of the mesh from the point of view of accuracy and convergence. In order to illustrate that this is the case for the dispersion and dissipation behaviour as well, we perform the same two-dimensional analysis on non-uniform random meshes. To construct the non-uniform mesh we randomly relocate all inner (ie. not lying on the boundary) vertices of the uniform mesh within the range

$[-\frac{h_{\text{ed}}}{3}, \frac{h_{\text{ed}}}{3}]$ in both directions, where h_{ed} is length of the uniformly distributed (one-dimensional) 'boundary' elements. Since the total number of elements in the non-uniform mesh is the same as in the uniform mesh, the values DoF/λ should also be the same (because this is the case 'on average'). However, on the bottom axes the smallest value of h is taken for computing λ/h. The results are shown in Figure 2.7 for the dispersion error and in Figure 2.8 for the dissipation error. They are qualitatively the same as the corresponding results on uniform meshes, which demonstrates the robustness of the RKDG method.

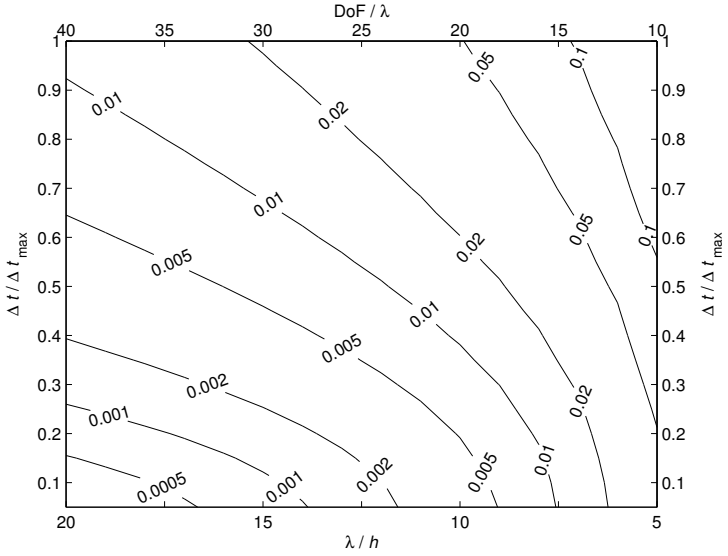
2.7 Concluding remarks

The main purpose of this chapter has been to study the global dispersion and dissipation errors of a high-order DG spectral element discretisation combined with the high-order SSP-RK time integration. We have shown that by applying the $(p+1)$st-order SSP-RK scheme to a spatial discretisation with pth-order polynomials we can retain $(p+1)$st-order convergence (without preprocessing) in the l_2-norm. Even when the order of the discretisation is increased beyond the accuracy of the fixed-order schemes, the computational work for the SSP-RK scheme is not significantly higher than that of the fixed-order ones. This favourable property can be explained by the much looser time-step restriction. It should be noted, however, that for large systems, the SSP-RK schemes could have a major disadvantage over low-storage RK schemes due to storage requirements. This is because an m-stage SSP-RK scheme requires m storage units per time step, whereas a low storage scheme requires only two storage units per time step.

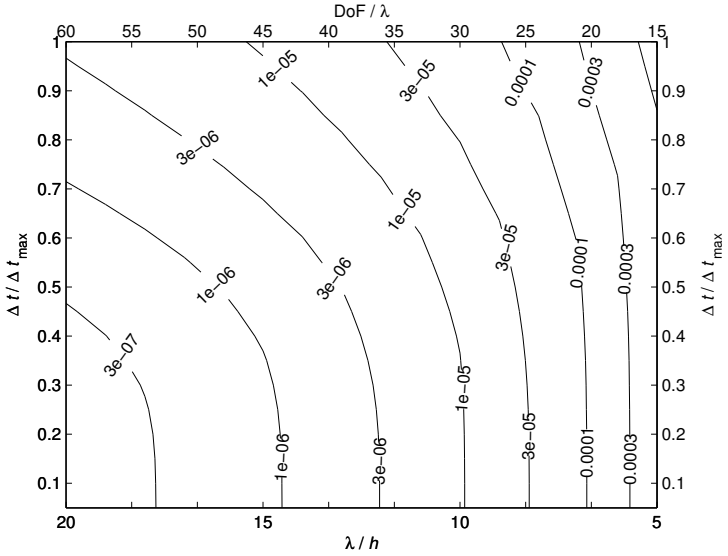
Through numerical Fourier analysis, it has been demonstrated that the dispersion error of the global scheme is generally at least one order higher than the dissipation error, irrespective of the actual order of the discretisation. It has also turned out that within the studied range of mesh sizes h and time steps Δt we cannot gain anything on the dispersion error by decreasing the time step, i.e. it is worth using the largest one permitted by the CFL condition. This seemingly unusual property can be explained by the fact that the maximum time step allowed by the CFL condition already inflicts a far smaller dispersion error than that of the space discretisation. Using this condition it has also been shown that the convergence rate of the dispersion and dissipation error is far higher, namely $\mathcal{O}(2p)$, than that of the error measured in the l^2-norm. Another important feature of the global scheme is that the number of degrees of freedom to obtain a given accuracy also decreases as the order of the approximation grows.

Appendix

The exact solution to (2.23) in the one-dimensional metallic cavity described in Section 2.6.1–with $\kappa = 1, 2$ signifying the two regions filled with different materials–

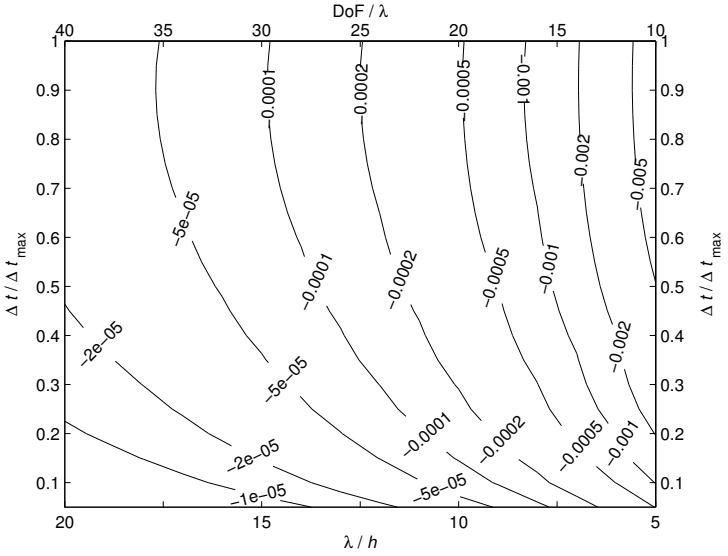


(a) $p = 1$

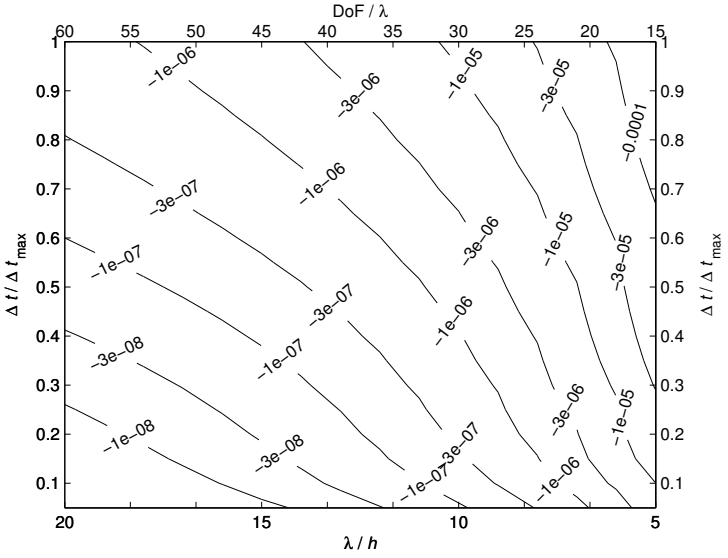


(b) $p = 2$

Figure 2.3: *Absolute value of the dispersion error as a function of wave length per mesh size (λ/h), degrees of freedom per wavelength (DoF/λ) and relative time step ($\Delta t/\Delta t_{\max}$) for polynomial orders $p = 1, 2$*

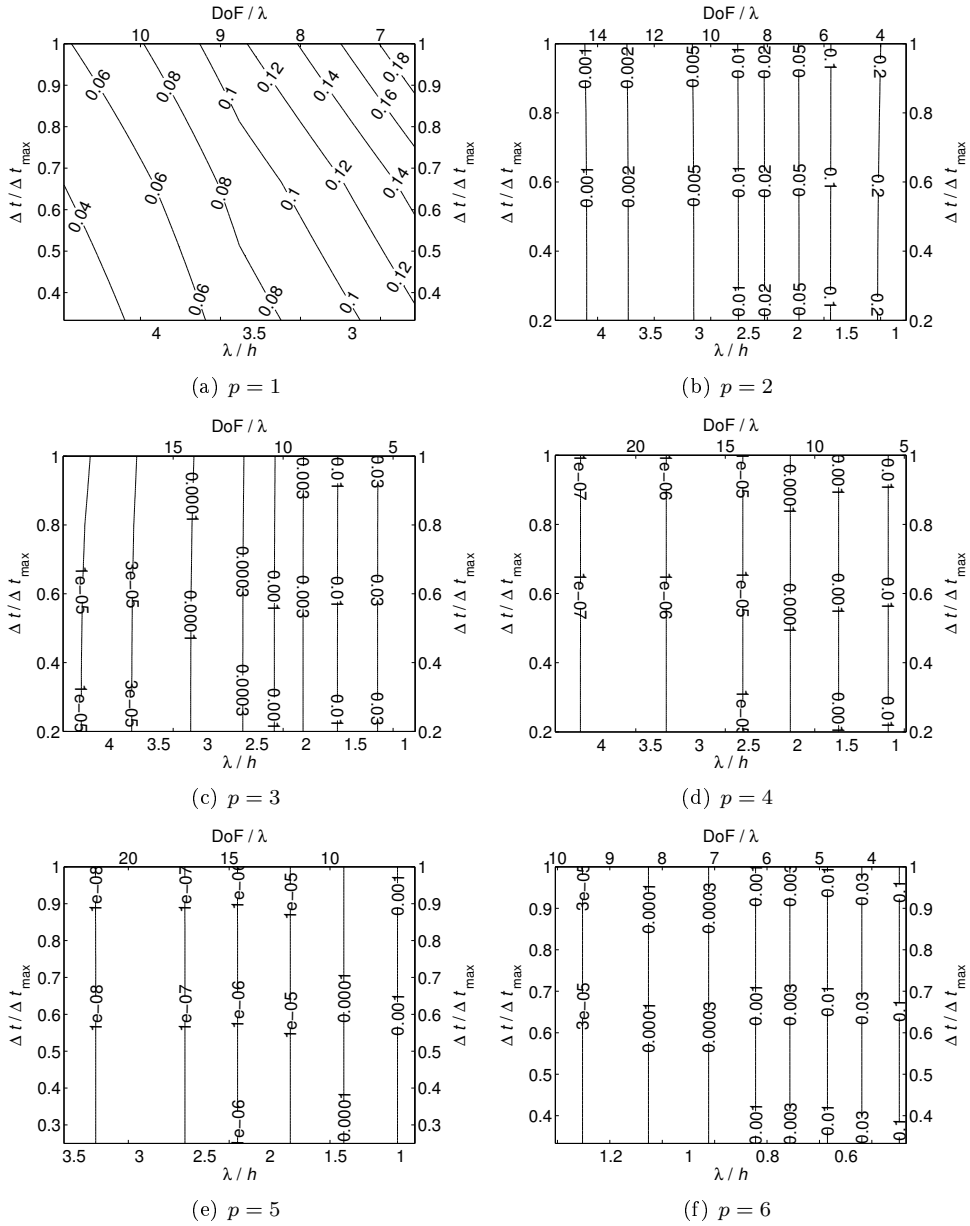


(a) $p = 1$



(b) $p = 2$

Figure 2.4: *Dissipation error as a function of wave length per mesh size (λ/h), degrees of freedom per wavelength (DoF/λ) and relative time step ($\Delta t/\Delta t_{\max}$) for polynomial orders $p = 1, 2$*

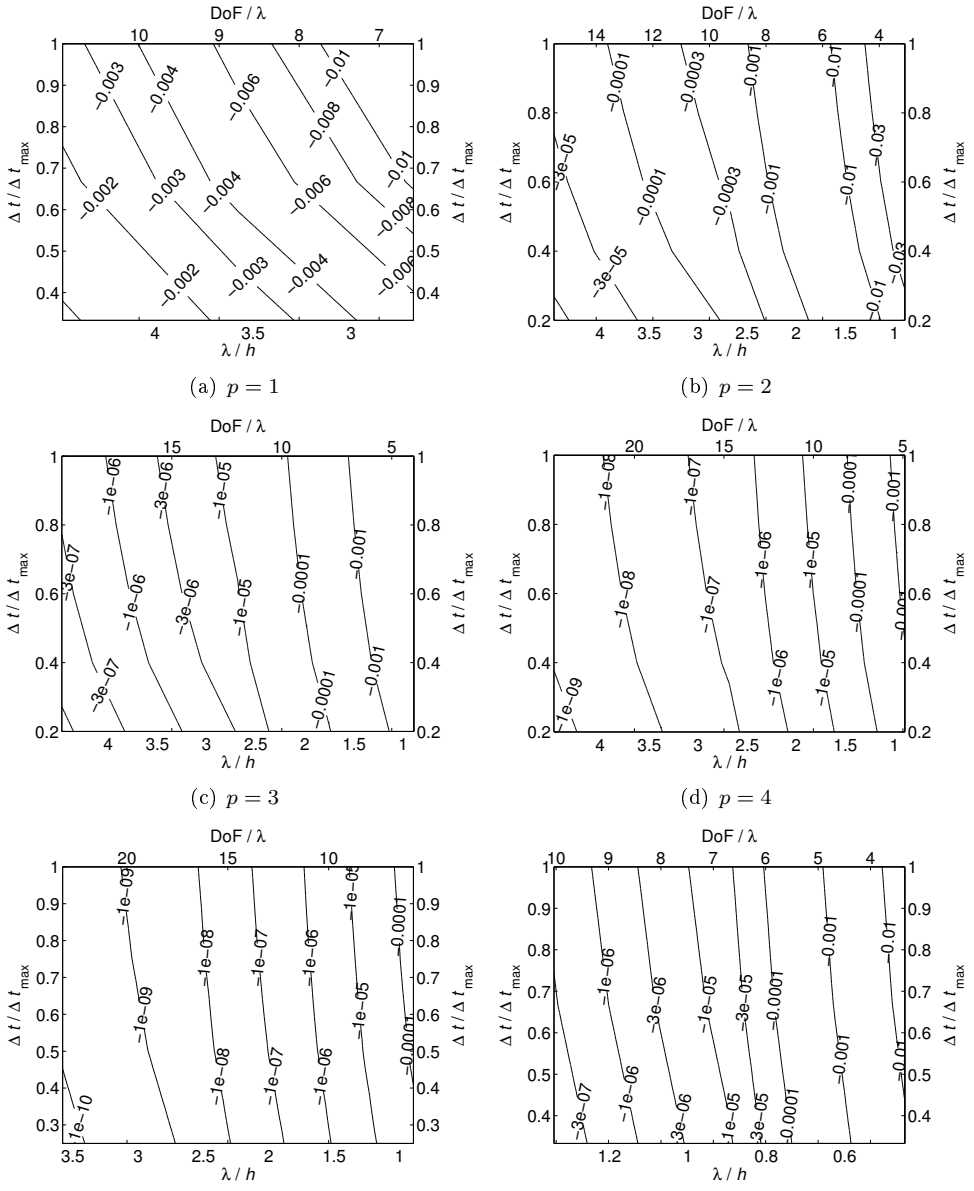


(a) $p = 1$ (b) $p = 2$

(c) $p = 3$ (d) $p = 4$

(e) $p = 5$ (f) $p = 6$

Figure 2.5: *Absolute value the of dispersion error as a function of wave length per mesh size (λ/h), degrees of freedom per wavelength (*DoF$/\lambda$*) and relative time step ($\Delta t/\Delta t_{\max}$) for polynomial orders $p = 1, 2, 3, 4, 5, 6$ on a uniform mesh*

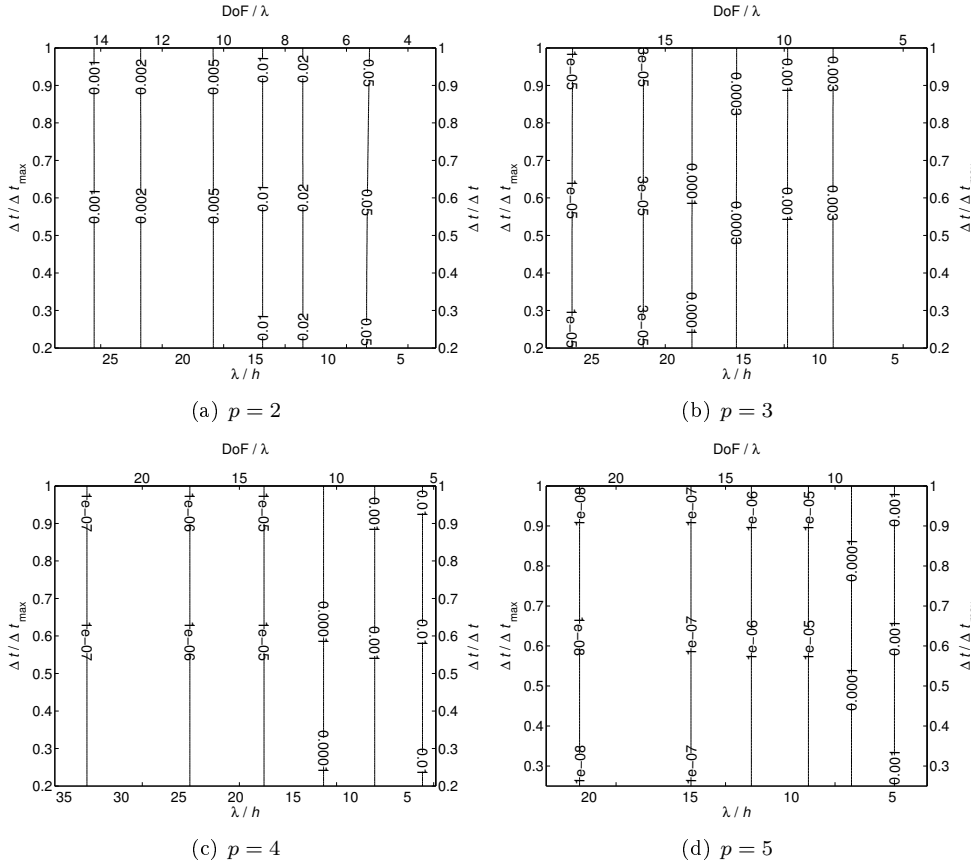


(a) $p = 1$ (b) $p = 2$

(c) $p = 3$ (d) $p = 4$

(e) $p = 5$ (f) $p = 6$

Figure 2.6: *Dissipation error as a function of wave length per mesh size (λ/h), degrees of freedom per wavelength (*DoF$/\lambda$*) and relative time step ($\Delta t/\Delta t_{\max}$) for polynomial orders $p = 1, 2, 3, 4, 5, 6$ on a uniform mesh*



(a) $p = 2$ (b) $p = 3$

(c) $p = 4$ (d) $p = 5$

Figure 2.7: *Absolute value of the dispersion error as a function of wave length per mesh size (λ/h), degrees of freedom per wavelength (DoF/λ) and relative time step ($\Delta t/\Delta t_{\max}$) for polynomial orders $p = 2, 3, 4, 5$ on a non-uniform random mesh*

is given by

$$E_\kappa = \left[-A_\kappa \mathrm{e}^{\mathrm{i}n_\kappa \omega x} + B_\kappa \mathrm{e}^{-\mathrm{i}n_\kappa \omega x}\right] \mathrm{e}^{\mathrm{i}\omega t},$$
$$H_\kappa = \left[\ \ A_\kappa \mathrm{e}^{\mathrm{i}n_\kappa \omega x} + B_\kappa \mathrm{e}^{-\mathrm{i}n_\kappa \omega x}\right] \mathrm{e}^{\mathrm{i}\omega t},$$

where

$$A_1 = \frac{n_2 \cos(n_2\omega)}{n_1 \cos(n_1\omega)}, \quad A_2 = \mathrm{e}^{-\mathrm{i}\omega(n_1+n_2)},$$

and

$$B_1 = A_1 \mathrm{e}^{-\mathrm{i}2n_1\omega}, \quad B_2 = A_2 \mathrm{e}^{\mathrm{i}2n_2\omega}.$$

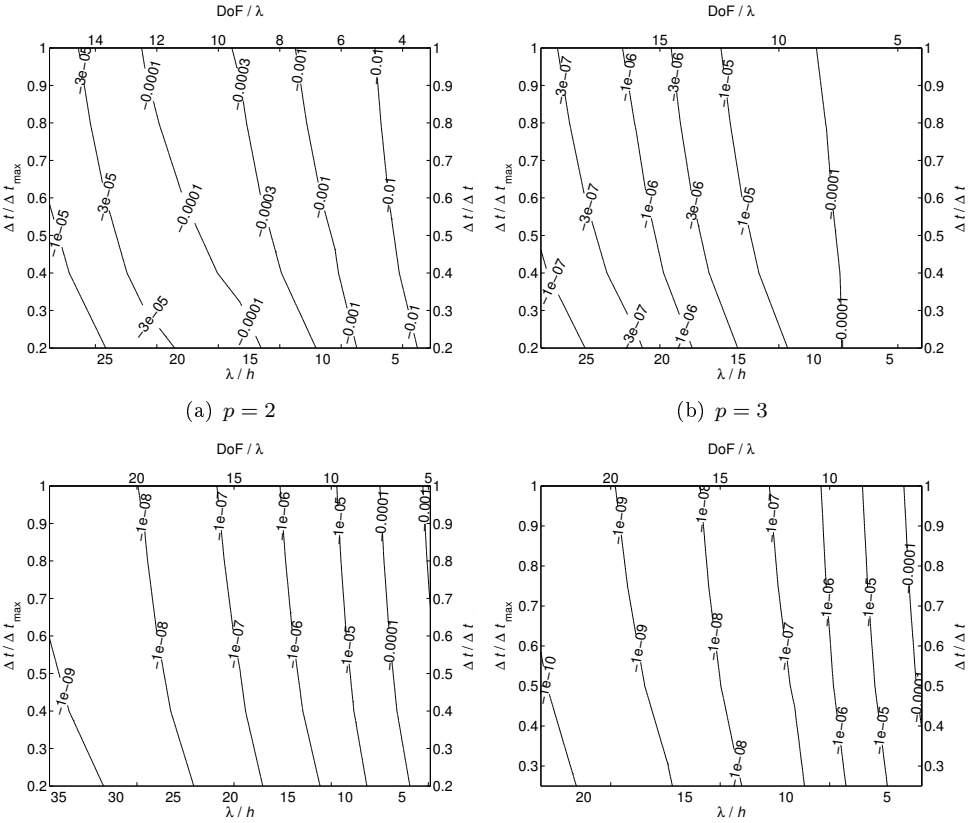


Figure 2.8: *Dissipation error as a function of wave length per mesh size (λ/h), degrees of freedom per wavelength (*DoF$/\lambda$*) and relative time step ($\Delta t/\Delta t_{\max}$) for polynomial orders $p = 2, 3, 4, 5$ on a non-uniform random mesh*

Here $n_\kappa = \sqrt{\varepsilon_\kappa}$ represents the local index of refraction, and the frequency takes the value $\omega = 2\pi/n$ if $n_1 = n_2 = n$ or is found as the solution to the equation

$$-n_2 \tan(n_1\omega) = n_1 \tan(n_2\omega).$$

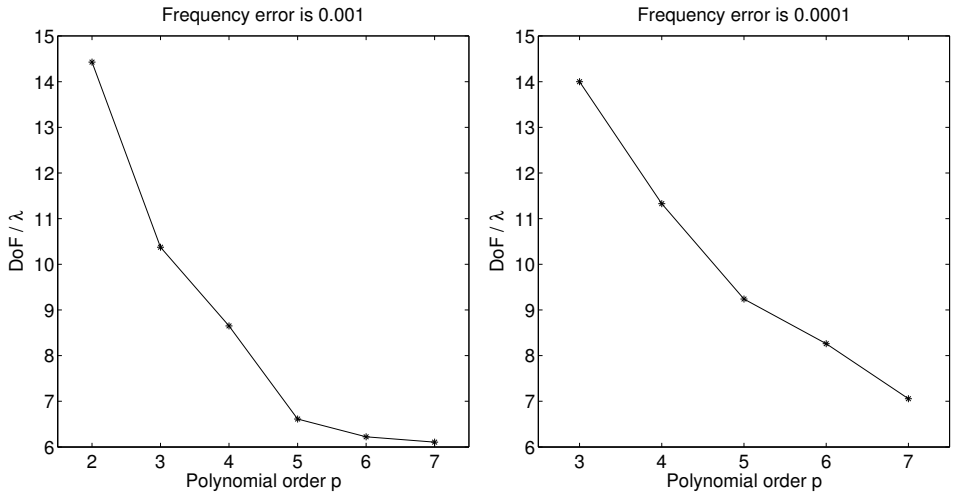


Figure 2.9: *Degrees of freedom as a function of polynomial order for given dispersion error* 10^{-3} *and* 10^{-4} *on uniform meshes*

CHAPTER 3

ON THE INCONSISTENCY IN THE IMPLEMENTATION OF H(CURL)-CONFORMING BASIS FUNCTIONS ON TETRAHEDRAL MESHES

3.1 Introduction

Hierarchic finite element bases are especially useful in the construction of p- and hp-adaptive finite element methods [23] thanks to the possibility to increase the polynomial order by only adding extra basis functions and leave the existing ones unchanged. An arbitrary high-order construction of such bases for spaces H^1, H(curl), H(div) and L^2 was given in [2, 74], where the authors also addressed the question of global conformity. The importance of H(curl)- and H(div)-conforming finite element methods (FEM) lies in the fact that they have better approximating properties than H^1-conforming elements when the physical spaces they discretise also belong to these spaces. Therefore, H(curl)- and H(div)-conforming FEMs are especially suitable for the numerical solution of the Maxwell equations, where the unknown fields generally belong to either of these spaces [45, 56].

The discussion on the families of hierarchic basis functions of arbitrary order is fairly complete in [2, 74] and we refer to those works and the references therein for a full description. In this chapter, we only focus on the H(curl)-conforming version of these bases in order to highlight a practical difficulty that is not fully addressed in the existing literature.

A key issue for hierarchic bases in the H(curl)-conforming discretisation is how to avoid a mismatch between neighbouring elements. One elegant way of doing this, proposed in [2], is through renumbering the mesh vertices in the preprocessing stage so that each element can be classified as either of two possible types based on their enumeration. Then it is possible to use two reference tetrahedra, one for each type, and no further consideration is needed about orientation. How-

ever, this approach is not always feasible because some software packages do not support the renumbering of vertices after mesh generation. An alternative solution is then to store the orientation data for each element and modify the definitions of the reference-element basis functions in such a way that it corresponds to the orientation of the given physical element.

We follow the second approach here since this thesis forms part of the development of *hp*GEM [62], a finite element package that is designed to deal with the orientation in this way. Unfortunately, this approach comes with an additional difficulty concerning the face-bubble functions. As we will illustrate through a simple example, the standard procedure to define the basis function on the reference element and map it to the mesh element may fail. Fixing this problem is quite straightforward but it is important that we highlight this phenomenon since recognising it can be much more of a struggle.

Once it is ensured that all basis functions are defined correctly throughout the whole domain, the finite element discretisation of a typical (system of) linear partial differential equation(s) can be broadly described in two steps. First, we compute the local matrices. This includes the computation of the element matrices for the $H(\mathrm{curl})$-conforming FEM and the element plus face matrices for the discontinuous Galerkin FEM (DG-FEM). Second, using the connectivity information of the generated mesh, we assemble the local matrices into global ones that form the final linear system (or system of ODEs) to be solved to obtain the discrete solution.

The chapter is organised as follows. We recall the construction of the $H(\mathrm{curl})$-conforming basis from [2] in Section 3.2. We describe an example of face-bubble functions in Section 3.3, show how the transformation fails in Section 3.4, and offer a quick fix and a brief discussion in Section 3.5. Section 3.6 is devoted to the implementation details, after which we conclude in Section 3.7.

3.2 $H(\mathrm{curl})$-conforming hierarchic basis

We begin by introducing the tessellation $\mathcal{T}_h$ that partitions the polyhedral domain $\Omega \subset \mathbb{R}^3$ into a set of tetrahedra $\{\mathbf{t}\}$. Suppose we have a unique enumeration of the vertices of the mesh. Then an arbitrary tetrahedron $\mathbf{t} = [o\ i\ j\ k] \in \mathcal{T}_h$ is defined by the vertices $\mathbf{v}_o$, $\mathbf{v}_i$, $\mathbf{v}_j$, $\mathbf{v}_k$, and the face $\mathbf{f} = [o\ i\ j]$ is likewise formed from the global numbering of the vertices. The unique directed tangent on an edge $[o\ i]$ is defined by

$$\boldsymbol{\tau}^{[o\ i]} = \mathbf{v}_i - \mathbf{v}_o,$$

while a unique pair of tangent vectors to face $\mathbf{f} = [o\ i\ j]$ is given by $\boldsymbol{\tau}^{[o\ i]}$ and $\boldsymbol{\tau}^{[o\ j]}$. A unique parameterisation of the edge can be defined as

$$\xi_{oi} = \lambda_i - \lambda_o$$

with λ_o and λ_i being the barycentric function of vertices $\mathbf{v}_o$ and $\mathbf{v}_i$, respectively.

We can now briefly recall the definition of the $H(\mathrm{curl})$-conforming polynomial basis, where L_n denotes the nth-order Legendre polynomial and we also assume that $o < i < j < k$.

- For edge $\mathbf{e} = [o\ i]$ and polynomial order p, the *edge functions* read

$$\begin{aligned} \boldsymbol{\psi}^{\mathbf{e}}_0 &= \lambda_i \nabla\lambda_o - \lambda_o \nabla\lambda_i, \\ \boldsymbol{\psi}^{\mathbf{e}}_1 &= \lambda_i \nabla\lambda_o + \lambda_o \nabla\lambda_i, \\ \boldsymbol{\psi}^{\mathbf{e}}_{l+1} &= \frac{2l+1}{l+1} L_l(\xi_{oi}) \boldsymbol{\psi}^{\mathbf{e}}_1 - \frac{l}{l+1} L_{l-1}(\xi_{oi}) \boldsymbol{\psi}^{\mathbf{e}}_0, \quad 1 \le l \le p-1. \end{aligned} \tag{3.1}$$

- For face $\mathbf{f} = [o\ i\ j]$ and edge $\mathbf{e} \subset \partial\mathbf{f}$, the *edge-based face functions* read

$$\boldsymbol{\psi}^{\mathbf{f},\mathbf{e}}_l = \lambda_o \lambda_i L_l(\xi_{oi}) \nabla\lambda_{\mathbf{f}/\mathbf{e}}, \quad 0 \le l \le p-2, \tag{3.2}$$

  where $\mathbf{f}/\mathbf{e}$ denotes the vertex opposite edge $\mathbf{e}$ in face $\mathbf{f}$.

- For face $\mathbf{f} = [o\ i\ j]$, the *face-bubble functions* read

$$\left.\begin{aligned} \boldsymbol{\psi}^{\mathbf{f},1}_{l,m} &= \lambda_o \lambda_i \lambda_j L_l(\xi_{oi}) L_m(\xi_{oj}) \boldsymbol{\tau}^{[o\ i]} \\ \boldsymbol{\psi}^{\mathbf{f},2}_{l,m} &= \lambda_o \lambda_i \lambda_j L_l(\xi_{oi}) L_m(\xi_{oj}) \boldsymbol{\tau}^{[o\ j]} \end{aligned}\right\} \quad 0 \le l, m, l+m \le p-3. \tag{3.3}$$

- For element $\mathbf{t} = [o\ i\ j\ k]$ and face $\mathbf{f} \subset \mathbf{t}$, the face-based interior functions read

$$\boldsymbol{\psi}^{\mathbf{t},\mathbf{f}}_{l,m} = \lambda_o \lambda_i \lambda_j L_l(\xi_{oi}) L_m(\xi_{oj}) \nabla\lambda_{\mathbf{t}/\mathbf{f}}, \quad 0 \le l, m, l+m \le p-3, \tag{3.4}$$

  where $\mathbf{t}/\mathbf{f}$ denotes the vertex opposite face $\mathbf{f}$.

- For element $\mathbf{t} = [o\ i\ j\ k]$, the interior bubble functions read

$$\boldsymbol{\psi}^{\mathbf{t},d}_{l,m,n} = \lambda_o \lambda_i \lambda_j \lambda_k L_l(\xi_{oi}) L_m(\xi_{oj}) L_n(\xi_{ok}) \boldsymbol{\tau}^d, \tag{3.5}$$

  with $0 \le l, m, n, l+m+n \le p-4$ and with $\boldsymbol{\tau}^d$, $d = 1, 2, 3$ denoting the directed tangent vectors $\boldsymbol{\tau}^1 = \boldsymbol{\tau}^{[o\ i]}$, $\boldsymbol{\tau}^2 = \boldsymbol{\tau}^{[o\ j]}$ and $\boldsymbol{\tau}^3 = \boldsymbol{\tau}^{[o\ k]}$, respectively.

In the above construction, the definitions are *global* in the sense that they apply to every $\mathbf{t} \in \mathcal{T}_h$. It is known that in this case every basis function is $H(\mathrm{curl})$-conforming (thanks to the condition $o < i < j < k$), i.e. their tangential components are continuous but there is no restriction on the normal components. (See [2] for detailed proofs.)

However, the standard approach is to define all basis functions (3.1)–(3.5) on the reference tetrahedron $\hat{\mathbf{t}} = [\hat{0}\ \hat{1}\ \hat{2}\ \hat{3}]$, given by its vertices in Cartesian coordinates $(\xi, \eta, \zeta) \in \mathbb{R}^3$ as

$$\hat{\mathbf{v}}_{\hat{0}} = (0,0,0)\,, \quad \hat{\mathbf{v}}_{\hat{1}} = (1,0,0)\,, \quad \hat{\mathbf{v}}_{\hat{2}} = (0,1,0)\,, \quad \hat{\mathbf{v}}_{\hat{3}} = (0,0,1)\,, \tag{3.6}$$

and then to transform them to a given mesh tetrahedron $\mathbf{t} \in \mathcal{T}_h$ using the transformation rule

$$\boldsymbol{\psi}^{\mathbf{t}} = \mathbb{J}_{\mathbf{t}}^{-T} \hat{\boldsymbol{\psi}}. \tag{3.7}$$

Here $\hat{\boldsymbol{\psi}}$ is the basis function on the reference element $\hat{\mathbf{t}}$, $\boldsymbol{\psi}^{\mathbf{t}}$ is the basis function on the physical element $\mathbf{t}$, and $\mathbb{J}$ is the Jacobian of the mapping between the reference element and the physical element. The derivation of this transformation rule is standard and can be found in many textbooks [56, 45, 51].

If (3.7) is to result in a globally conforming approximation, one has to make sure that the intrinsic orientation of the reference element matches that of the physical element for each $\mathbf{t} \in \mathcal{T}_h$. This is necessary in order to avoid a mismatch of a given basis function between neighbouring elements, and its implementation is not straightforward for hierarchic bases. One elegant way to do this is through the renumbering of the mesh vertices in the preprocessing stage so that each element can be classified as either of two possible types based on their enumeration. Then it is possible to use two reference tetrahedra, one for each type, and no further consideration is needed about orientation. See [2] again for more details.

If this approach is not feasible (e.g. some FEM software packages do not support the renumbering of vertices after mesh generation), an alternative solution is to store the orientation data (i.e. the enumeration) for each element. Then one can modify the definitions of the reference-element basis functions in such a way that the construction of the basis (3.1)–(3.5) follows the numbering of the physical element and not that of the original reference element (3.6). For example, if the global numbering of a tetrahedron is $\mathbf{t} = [15\ 96\ 8\ 24]$, then the vertex numbering in (3.6) changes to

$$\hat{\mathbf{v}}_{\hat{1}} = (0,0,0)\,, \quad \hat{\mathbf{v}}_{\hat{3}} = (1,0,0)\,, \quad \hat{\mathbf{v}}_{\hat{0}} = (0,1,0)\,, \quad \hat{\mathbf{v}}_{\hat{2}} = (0,0,1)\,.$$

In the following, we will abuse the notation slightly and refer to this situation as $\hat{\mathbf{t}} = [\hat{1}\ \hat{3}\ \hat{0}\ \hat{2}]$.

Whichever approach is used the crux of the problem is to have the same intrinsic orientation for the reference tetrahedron as for the physical tetrahedron. Once that is done, all basis functions are expected to be globally $H(\mathrm{curl})$-conforming after the transformation (3.7) is applied. In the remaining part of the chapter, we show, through a simple example, that when the second approach is used, the face bubble-functions fail to satisfy this criterion.

3.3 An example of global face-bubble functions

Let us consider a mesh that divides the unit cube $\Omega = (0,1)^3$ into five tetrahedra. We enumerate the vertices of the mesh as

$$\begin{aligned} &\mathbf{v}_0 = (0,0,0)\,, \quad \mathbf{v}_1 = (1,0,0)\,, \quad \mathbf{v}_2 = (0,1,0)\,, \quad \mathbf{v}_3 = (1,1,0)\,, \\ &\mathbf{v}_4 = (0,0,1)\,, \quad \mathbf{v}_5 = (1,0,1)\,, \quad \mathbf{v}_6 = (0,1,1)\,, \quad \mathbf{v}_7 = (1,1,1)\,, \end{aligned}$$

where the vertices are determined by their Cartesian coordinates $(x,y,z) \in \mathbb{R}^3$. We define each tetrahedron by the global numbers of its vertices

$$\begin{aligned} \mathbf{t}_0 = [0\ 1\ 2\ 4]\,, \quad \mathbf{t}_1 = [3\ 2\ 1\ 7]\,, \quad \mathbf{t}_2 = [5\ 4\ 7\ 1]\,, \\ \mathbf{t}_3 = [6\ 7\ 4\ 2]\,, \quad \mathbf{t}_4 = [4\ 1\ 2\ 7]\,. \end{aligned} \tag{3.8}$$

We have chosen the numbering in (3.8) in such a way that the positions of the vertices in the physical element correspond to the positions of the vertices in the reference element, once the mapping is performed. So for example, for $\mathbf{t}_2$, vertex $\mathbf{v}_5$ will be mapped to $\hat{\mathbf{v}}_{\hat{0}}$, vertex $\mathbf{v}_4$ will be mapped to $\mathbf{v}_{\hat{1}}$, and so on. We emphasise again that the numbering in (3.8) is correct in a sense that it preserves orientation.

Throughout this example, the subject of our investigation is the face $\mathbf{f} = [2\ 4\ 7]$ with left element $\mathbf{t}_3$ and right element $\mathbf{t}_4$. (Of course, definitions 'left' and 'right' are arbitrary in three dimensions.) For our purpose, it suffices to consider the lowest order face-bubble functions, i.e. those with $l = m = 1$ in (3.3). These first appear in the construction of the third-order polynomial space, when there are two face-bubble functions associated with face $\mathbf{f} = [2\ 4\ 7]$. They are defined as

$$\boldsymbol{\psi}_0^{\mathbf{f}} = \lambda_2\lambda_7\lambda_4\boldsymbol{\tau}^{[2\ 4]} \quad \text{and} \quad \boldsymbol{\psi}_1^{\mathbf{f}} = \lambda_2\lambda_7\lambda_4\boldsymbol{\tau}^{[2\ 7]}, \tag{3.9}$$

where λ_i is the barycentric function of (global) vertex $\mathbf{v}_i$. Furthermore, $\boldsymbol{\tau}^{[2\ 4]} = \mathbf{v}_4 - \mathbf{v}_2$ and $\boldsymbol{\tau}^{[2\ 7]} = \mathbf{v}_7 - \mathbf{v}_2$.

It is easy to see that (3.9) is $H(\mathrm{curl})$-conforming. In practice, however, one aims to define all basis functions locally on a single element (the reference element) and use the transformation (3.7) between the reference-element basis functions and the physical-element basis functions.

3.4 Transformation of the face-bubble functions

We now look at how the face-bubble functions in (3.9) transform from the reference element to the physical elements. Since there are two elements that contain the face $\mathbf{f} = [2\ 4\ 7]$, we need to look at four separate transformations (two basis function in both elements). Desirably, the result would be tangential continuity for both basis functions through the face.

3.4.1 Left element

Let us first consider $\mathbf{t}_3 = [6\ 7\ 4\ 2]$. The local numbers corresponding to face $\mathbf{f} = [2\ 7\ 4]$ are $[\hat{3}\ \hat{1}\ \hat{2}]$. So we define on the reference element the tangential vectors $\hat{\boldsymbol{\tau}}^{[\hat{3}\ \hat{2}]} = \hat{\mathbf{v}}_{\hat{2}} - \hat{\mathbf{v}}_{\hat{3}} = (0, 1, -1)$ and $\hat{\boldsymbol{\tau}}^{[\hat{3}\ \hat{1}]} = \hat{\mathbf{v}}_{\hat{1}} - \hat{\mathbf{v}}_{\hat{3}} = (1, 0, -1)$. Since we have

$$\hat{\lambda}_{\hat{0}} = 1 - \xi - \eta - \zeta, \quad \hat{\lambda}_{\hat{1}} = \xi, \quad \hat{\lambda}_{\hat{2}} = \eta, \quad \hat{\lambda}_{\hat{3}} = \zeta,$$

where ξ, η and ζ are the Cartesian coordinates on the reference element, the two basis functions are defined as

$$\hat{\boldsymbol{\psi}}_0^{\mathbf{f}} = \xi\eta\zeta\hat{\boldsymbol{\tau}}^{[\hat{3}\ \hat{2}]} = \begin{pmatrix} 0 \\ \xi\eta\zeta \\ -\xi\eta\zeta \end{pmatrix} \quad \text{and} \quad \hat{\boldsymbol{\psi}}_1^{\mathbf{f}} = \xi\eta\zeta\hat{\boldsymbol{\tau}}^{[\hat{3}\ \hat{1}]} = \begin{pmatrix} \xi\eta\zeta \\ 0 \\ -\xi\eta\zeta \end{pmatrix}.$$

Again, $H(\mathrm{curl})$-conforming fields transform as

$$\boldsymbol{\psi}^{\mathbf{t}_i} = \mathbb{J}_{\mathbf{t}_i}^{-T}\hat{\boldsymbol{\psi}},$$

where $\mathbb{J}$ is the Jacobian of the function that maps reference element coordinates (ξ, η, ζ) to the coordinates (x, y, z) of physical element $\mathbf{t}_i$. For $\mathbf{t}_3$ this mapping is $\boldsymbol{F}_3 : (\xi, \eta, \zeta) \rightarrow \mathbf{v}_6 + \xi\,(\mathbf{v}_7 - \mathbf{v}_6) + \eta\,(\mathbf{v}_4 - \mathbf{v}_6) + \zeta\,(\mathbf{v}_2 - \mathbf{v}_6)$, and therefore we have

$$\mathbb{J}_{\mathbf{t}_3}^{-T} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

On the physical element $\mathbf{t}_3$, this in turn results in

$$\boldsymbol{\psi}_0^L = \begin{pmatrix} 0 \\ -x\,(1-y)\,(1-z) \\ x\,(1-y)\,(1-z) \end{pmatrix} \quad \text{and} \quad \boldsymbol{\psi}_1^L = \begin{pmatrix} x\,(1-y)\,(1-z) \\ 0 \\ x\,(1-y)\,(1-z) \end{pmatrix},$$

where we have used the superscript L to represent the 'left' element, in this case $\mathbf{t}_3$.

3.4.2 Right element

As a second step, let us turn our attention to element $\mathbf{t}_4 = [4\ 1\ 2\ 7]$. The local numbers corresponding to face $\mathbf{f} = [2\ 4\ 7]$ are $\left[\hat{2}\ \hat{0}\ \hat{3}\right]$. Since $\hat{\boldsymbol{\tau}}^{[\hat{2}\ \hat{3}]} = \hat{\mathbf{v}}_{\hat{3}} - \hat{\mathbf{v}}_{\hat{2}} = (0, -1, 1)$ and $\hat{\boldsymbol{\tau}}^{[\hat{2}\ \hat{0}]} = \hat{\mathbf{v}}_{\hat{0}} - \hat{\mathbf{v}}_{\hat{2}} = (0, -1, 0)$, on the reference element we now have

$$\hat{\boldsymbol{\psi}}_0^{\mathbf{f}} = (1-\xi-\eta-\zeta)\,\eta\zeta\hat{\boldsymbol{\tau}}^{[\hat{2}\ \hat{0}]} = \begin{pmatrix} 0 \\ -(1-\xi-\eta-\zeta)\,\eta\zeta \\ 0 \end{pmatrix},$$

$$\hat{\boldsymbol{\psi}}_1^{\mathbf{f}} = (1-\xi-\eta-\zeta)\,\eta\zeta\hat{\boldsymbol{\tau}}^{[\hat{2}\ \hat{3}]} = \begin{pmatrix} 0 \\ -(1-\xi-\eta-\zeta)\,\eta\zeta \\ (1-\xi-\eta-\zeta)\,\eta\zeta \end{pmatrix}.$$

For the current element, $\mathbf{t}_4$, we have the mapping $\boldsymbol{F}_4 : (\xi, \eta, \zeta) \rightarrow \mathbf{v}_4 + \xi\,(\mathbf{v}_1 - \mathbf{v}_4) + \eta\,(\mathbf{v}_2 - \mathbf{v}_4) + \zeta\,(\mathbf{v}_7 - \mathbf{v}_4)$, and thus the transformation matrix

$$\mathbb{J}_{\mathbf{t}_4}^{-T} = \frac{1}{2}\begin{pmatrix} 1 & -1 & 1 \\ -1 & 1 & 1 \\ -1 & -1 & 1 \end{pmatrix}.$$

As a result, the physical basis functions read

$$\boldsymbol{\psi}_0^R = \frac{1}{16}\begin{pmatrix} (z-x-y+1)\,(y-x-z+1)\,(x+y+z-1) \\ -(z-x-y+1)\,(y-x-z+1)\,(x+y+z-1) \\ (z-x-y+1)\,(y-x-z+1)\,(x+y+z-1) \end{pmatrix},$$

$$\boldsymbol{\psi}_1^R = \frac{1}{8}\begin{pmatrix} (z-x-y+1)\,(y-x-z+1)\,(x+y+z-1) \\ 0 \\ (z-x-y+1)\,(y-x-z+1)\,(x+y+z-1) \end{pmatrix}.$$

Finally, note that face $\mathbf{f} = [2\ 4\ 7]$ lies on the plane $z = x - y + 1$ and that

$$x\,(1-y)\,(1-z)\,\big|_{z=x-y+1} = \frac{1}{8}\,(z-x-y+1)\,(y-x-z+1)\,(x+y+z-1)\,\big|_{z=x-y+1}.$$

Using this latter relation it becomes clear that while $\boldsymbol{\psi}_1^L = \boldsymbol{\psi}_1^R$ on the face $\mathbf{f}$ as expected, the identity simply does not hold for the other basis function: $\boldsymbol{\psi}_0^L \neq \boldsymbol{\psi}_0^R$.

This inconsistency causes the method to fail for polynomial orders $p \geq 0$, as it dashes the possibility to define all basis functions on the reference element and transform them to any given physical element.

3.5 Brief discussion of the example

The simplest way to circumvent the problem highlighted in the previous example is to first transform the scalar face-bubble functions,

$$\psi^{\mathbf{f}} = \lambda_o \lambda_i \lambda_j L_l(\xi_{oi}) L_l(\xi_{oj}), \tag{3.10}$$

and multiply them with the tangential vectors afterwards. This solution also identifies the cause of the problem: that the tangent vectors that explicitly appear in the definition of the face-bubble functions do not transform in a conforming way. This also explains why we do not experience similar problems in the transformation of the edge functions and edge-based face functions. They only use the barycentric functions in their definitions and not the tangential vector explicitly. (For the interior functions conformity is trivial since their tangential components are zero.)

As a remark, we mention that the situation is very similar if the curl of the basis functions is also needed, as it is often the case. The curl of the face-bubble functions can be written as

$$\nabla \times \boldsymbol{\psi}_0^{\mathbf{f}} = \nabla\psi^{\mathbf{f}} \times \boldsymbol{\tau}^{[o\ i]} \quad \text{and} \quad \nabla \times \boldsymbol{\psi}_1^{\mathbf{f}} = \nabla\psi^{\mathbf{f}} \times \boldsymbol{\tau}^{[o\ j]}, \tag{3.11}$$

where $\nabla\psi^{\mathbf{f}}$ is the gradient of the scalar face-bubble function (3.10). Note that scalar function (3.10) transform in a H^1-conforming way while its gradient transforms in $H(\mathrm{curl})$-conforming way.

3.6 Implementation details for the second-order Maxwell equation

To illustrate the main implementation steps of a finite-element discretisation when the above-described basis functions are used, we consider the time-harmonic Maxwell equation

$$\begin{aligned} \nabla \times \frac{1}{\mu_{\mathrm{r}}} \nabla \times \boldsymbol{E} - k^2 \varepsilon_{\mathrm{r}} \boldsymbol{E} &= \boldsymbol{J} \quad \text{in } \Omega, \\ \boldsymbol{n} \times \boldsymbol{E} &= \boldsymbol{g} \quad \text{on } \Gamma. \end{aligned} \tag{3.12}$$

Here Ω is an open bounded Lipschitz polyhedron on $\mathbb{R}^3$ with boundary $\Gamma = \partial\Omega$ and outward normal unit vector $\boldsymbol{n}$. The right-hand side $\boldsymbol{J}$ is the external source and k is the (real-valued) wave number with the assumption that k^2 is not a Maxwell eigenvalue. The (relative) permittivity and the (relative) permeability correspond to vacuum (or dry air), i.e. we set $\varepsilon_{\mathrm{r}} = 1$ and $\mu_{\mathrm{r}} = 1$.

We will now describe the implementation of two discontinuous Galerkin (DG) discretisations and a high-order $H(\mathrm{curl})$-conforming finite element discretisation of (3.12), all of which result in a symmetric algebraic system. One of the DG methods is the interior-penalty DG (IP-DG) – see e.g. [46] –, the other is the DG method originally introduced in [6, 12]. Both DG methods are derived and analysed in detail in the next chapter as well as in [68]. Here it suffices to provide the weak formulations only, for which we first need to introduce some notation. We view the $H(\mathrm{curl})$-conforming FEM as a special case of the DG methods, where the flux terms are absent and the tangential continuity is taken care of through the assembly procedure.

As before, let $\mathcal{T}_h$ denote the tessellation that partitions the polyhedral domain $\Omega \subset \mathbb{R}^3$ into a set of tetrahedra $\{\mathbf{t}\}$. The notations $\mathcal{F}_h$, $\mathcal{F}_h^i$ and $\mathcal{F}_h^b$ stand respectively for the set of all faces $\{\mathbf{f}\}$, the set of all internal faces, and the set of all boundary faces. Furthermore, we introduce $(\cdot,\cdot)_D$ for the standard inner product in $\left[L^2(D)\right]^3$,

$$(\boldsymbol{u},\boldsymbol{v})_D = \int_D \boldsymbol{u}\cdot\boldsymbol{v}\,\mathrm{d}V,$$

and the operator ∇_h for the elementwise application of $\nabla = (\partial/\partial x, \partial/\partial y, \partial/\partial z)^T$.

3.6.1 Discontinuous Galerkin weak formulations

With Q^p denoting the pth-order polynomial space that is spanned by the basis functions described in Section 3.2, we now define the space Σ_h^p as

$$\Sigma_h^p := \left\{ \boldsymbol{\sigma} \in [L^2(\Omega)]^3 \;\middle|\; \boldsymbol{\sigma}|_{\mathbf{t}} \in Q^p,\, \forall \mathbf{t} \in \mathcal{T}_h \right\}. \tag{3.13}$$

Consider an interface $\mathbf{f} \in \mathcal{F}_h$ between element $\mathbf{t}^L$ and element $\mathbf{t}^R$, and let $\boldsymbol{n}^L$ and $\boldsymbol{n}^R$ represent their respective outward pointing normal vectors. We define the tangential jump and the average of the quantity $\boldsymbol{u}$ across interface $\mathbf{f}$ as

$$[\![\boldsymbol{u}]\!]_T = \boldsymbol{n}^L \times \boldsymbol{u}^L + \boldsymbol{n}^R \times \boldsymbol{u}^R \quad \text{and} \quad \{\!\{\boldsymbol{u}\}\!\} = \left(\boldsymbol{u}^L + \boldsymbol{u}^R\right)/2,$$

respectively. Here $\boldsymbol{u}^L$ and $\boldsymbol{u}^R$ are the values of the trace of $\boldsymbol{u}$ at $\partial\mathbf{t}^L$ and $\partial\mathbf{t}^R$, respectively. At the boundary Γ, we set $\{\!\{\boldsymbol{u}\}\!\} = \boldsymbol{u}$ and $[\![\boldsymbol{u}]\!]_T = \boldsymbol{n}\times\boldsymbol{u}$. For a given face $\mathbf{f} \in \mathcal{F}_h$, we will also need the local lifting operator $\mathcal{R}_{\mathbf{f}} : \left[L^2(\mathbf{f})\right]^3 \to \Sigma_h^p$, defined as

$$(\mathcal{R}_{\mathbf{f}}(\boldsymbol{u}),\boldsymbol{v})_\Omega = \int_{\mathbf{f}} \boldsymbol{u}\cdot\{\!\{\boldsymbol{v}\}\!\}\,\mathrm{d}A, \quad \forall \boldsymbol{v} \in \Sigma_h^p. \tag{3.14}$$

Note that $\mathcal{R}_{\mathbf{f}}(\boldsymbol{u})$ vanishes outside the elements connected to the face $\mathbf{f}$.

The weak formulation of the IP-DG method for the time-harmonic Maxwell equation reads

$$\begin{aligned}\mathcal{B}_h^{ip}(\mathbf{E}_h,\boldsymbol{\phi}) :=& \\ &(\nabla_h\times\boldsymbol{E}_h,\nabla_h\times\boldsymbol{\phi})_\Omega - k^2\,(\boldsymbol{E}_h,\boldsymbol{\phi})_\Omega - \int_{\mathcal{F}_h}[\![\boldsymbol{E}_h]\!]_T\cdot\{\!\{\nabla_h\times\boldsymbol{\phi}\}\!\}\,\mathrm{d}A \\ &-\int_{\mathcal{F}_h}\{\!\{\nabla_h\times\boldsymbol{E}_h\}\!\}\cdot[\![\boldsymbol{\phi}]\!]_T\,\mathrm{d}A+\int_{\mathcal{F}_h}\mathsf{a}_h\,[\![\boldsymbol{E}]\!]_T\cdot[\![\boldsymbol{\phi}]\!]_T\,\mathrm{d}A \\ &=(\boldsymbol{J},\boldsymbol{\phi})_\Omega-\int_{\mathcal{F}_h^b}\boldsymbol{g}\cdot(\nabla_h\times\boldsymbol{\phi})\,\mathrm{d}A+\int_{\mathcal{F}_h^b}\mathsf{a}_h\boldsymbol{g}\cdot(\boldsymbol{n}\times\boldsymbol{\phi})\,\mathrm{d}A,\end{aligned}\tag{3.15}$$

where the penalty parameter a_F depends on both the mesh size and the polynomial order. Note that in the left-hand side we no longer distinguish explicitly between internal and boundary faces. This is permissible thanks to the definitions of the average and the tangential jump at the boundary.

The weak formulation of the DG method with what we call here the Brezzi formulation, originally introduced in [6, 12], is formulated in the following way. First introduce the bilinear form $\mathcal{B}_h^{br}:\Sigma_h^p\times\Sigma_h^p\to\mathbb{R}$ and the linear form $\mathcal{J}_h^{br}:\Sigma_h^p\to\mathbb{R}$ as

$$\begin{aligned}\mathcal{B}_h^{br}(\boldsymbol{E}_h,\boldsymbol{\phi}) =&\,(\nabla_h\times\boldsymbol{E}_h,\nabla_h\times\boldsymbol{\phi})_\Omega-k^2\,(\boldsymbol{E}_h,\boldsymbol{\phi})_\Omega \\ &-\int_{\mathcal{F}_h}[\![\boldsymbol{E}_h]\!]_T\cdot\{\!\{\nabla_h\times\boldsymbol{\phi}\}\!\}\,\mathrm{d}A-\int_{\mathcal{F}_h}\{\!\{\nabla_h\times\boldsymbol{E}_h\}\!\}\cdot[\![\boldsymbol{\phi}]\!]_T\,\mathrm{d}A \\ &+\sum_{F\in\mathcal{F}_h}(\eta_F+n_f)\,(\mathcal{R}_F([\![\boldsymbol{E}]\!]_T),\mathcal{R}_F([\![\boldsymbol{\phi}]\!]_T))_\Omega\,,\end{aligned}\tag{3.16}$$

and

$$\begin{aligned}\mathcal{J}_h^{br}(\boldsymbol{\phi}) =&\,(\boldsymbol{J},\boldsymbol{\phi})_\Omega-\int_{\mathcal{F}_h^b}\boldsymbol{g}\cdot(\nabla_h\times\boldsymbol{\phi})\,\mathrm{d}A \\ &+\sum_{F\in\mathcal{F}_h^b}(\eta_F+n_f)\,(\mathcal{R}_F(\boldsymbol{g}),\mathcal{R}_F(\boldsymbol{n}\times\boldsymbol{\phi}))_\Omega\,,\end{aligned}\tag{3.17}$$

respectively. In contrast to IP-DG, the penalty parameter η_F is now independent of the both the mesh size and the polynomial order, while n_f is the number of sides of the element ($n_f=4$ for a tetrahedron). Then the discrete formulation for the time-harmonic Maxwell equations can be written as follows. Find $\boldsymbol{E}_h\in\Sigma_h^p$ such that for all $\boldsymbol{\phi}\in\Sigma_h^p$ the relation

$$\mathcal{B}_h^{br}(\boldsymbol{E}_h,\boldsymbol{\phi})=\mathcal{J}_h^{br}(\boldsymbol{\phi})\tag{3.18}$$

is satisfied. See Chapter 4 for more the analysis and parameter estimates of these two methods.

3.6.2 Weak formulation of the $H(\mathrm{curl})$-conforming discretisation

Instead of (3.13), we now define the discrete space of globally $H(\mathrm{curl})$-conforming functions as

$$\Upsilon_h^p := \left\{ \boldsymbol{v} \in [H_0(\mathrm{curl}, \Omega)]^3 \;\middle|\; \boldsymbol{v}|_K \in Q^p, \forall K \in \mathcal{T}_h \right\}, \tag{3.19}$$

and let the set of basis functions $\{\boldsymbol{\psi}_i\}$ span the space Υ_h^p. See [56] for a detailed discussion on both continuous and discrete $H(\mathrm{curl})$-conforming spaces. We approximate the electric field $\mathbf{E}$ as

$$\mathbf{E} \approx \mathbf{E}_h = \sum_i u_i(t) \boldsymbol{\psi}_i(x), \tag{3.20}$$

from which the discrete weak formulation reads as follows. Find $\mathbf{E}_h \in \Upsilon_h^p$ such that $\forall \boldsymbol{\phi} \in \Upsilon_h^p$ the relation

$$(\nabla \times \mathbf{E}_h, \nabla \times \boldsymbol{\phi})_\Omega - k^2 (\mathbf{E}_h, \boldsymbol{\phi})_\Omega = (\mathbf{J}, \boldsymbol{\phi})_\Omega \tag{3.21}$$

is satisfied. Note that here the ∇ operator is defined globally, and not in an elementwise fashion as for the DG methods, because the tangential continuity is now enforced strongly.

The discontinuous nature of the weak formulations (3.15) and (3.18) makes it especially natural to first compute the element and face matrices and then assemble them into a global matrix. We will follow the same approach also for the $H(\mathrm{curl})$-conforming method, where only element matrices are needed. However, we then have to make sure that the basis functions associated with a given edge or face in the global mesh are assembled into the same entry of the global matrix.

3.6.3 Elemental matrices

Using the basis functions defined in Section 3.2, we can now express the unknown field with the polynomial expansion locally in each element (cf. (3.20)) as

$$\boldsymbol{E}_h^{\mathbf{t}}(x) = \sum_{j=1}^{N_p} E_j^{\mathbf{t}} \boldsymbol{\psi}_j^{\mathbf{t}}(x), \quad \forall x \in \mathbf{t}. \tag{3.22}$$

In order to compute the elemental matrices, let again $\hat{\mathbf{t}}$ be the reference tetrahedron and let $\mathbf{t} \in \mathcal{T}_h$ be any given physical tetrahedron. Then the entries of the elemental stiffness matrix $\mathbf{S}_\mathbf{t}$ and the elemental mass matrix $\mathbf{M}_\mathbf{t}$ can be expressed as

$$\mathbf{S}_{ij}^{\mathbf{t}} = \int_{\mathbf{t}} (\nabla_h \times \boldsymbol{\psi}_i) \cdot (\nabla_h \times \boldsymbol{\psi}_j) \, \mathrm{d}V \quad \text{and} \quad \mathbf{M}_{ij}^{\mathbf{t}} = k^2 \int_{\mathbf{t}} \boldsymbol{\psi}_i \cdot \boldsymbol{\psi}_j \, \mathrm{d}V,$$

respectively. Here, the indices run from $i, j = 1, \ldots, N_\mathbf{t}$, with $N_\mathbf{t}$ being the number of degrees of freedom in element $\mathbf{t}$. Exploiting the transformation rules

$$\boldsymbol{\psi}^{\mathbf{t}_i} = \mathbb{J}_{\mathbf{t}_i}^{-T} \hat{\boldsymbol{\psi}} \quad \text{and} \quad \nabla \times \boldsymbol{\psi}^{\mathbf{t}_i} = \frac{1}{\det(\mathbb{J}_{\mathbf{t}_i})} \nabla \times \hat{\boldsymbol{\psi}},$$

the above integrals can be computed on the reference element $\hat{\mathbf{t}}$ using quadrature rules (the question of quadrature rules is briefly addressed in the next paragraph).

As for the face contributions $\mathbf{f} \in \mathcal{F}_h$ in (3.15) and (3.16), we need to consider values in the two elements $\mathbf{t}^L$ and $\mathbf{t}^R$ which are connected through face $\mathbf{f}$. So we (abuse the notation slightly and) define the matrices $\mathbf{D}^{LR}$, $\mathbf{G}^{LR}$ and $\mathbf{H}^{LR}$ as

$$\begin{aligned}
\mathbf{D}_{ij}^{LR} &= \int_{\mathbf{f}} \boldsymbol{\psi}_i^L \cdot \left(\boldsymbol{n}^R \times \boldsymbol{\psi}_j^R \right) \mathrm{d}A, \\
\mathbf{G}_{ij}^{LR} &= \int_{\mathbf{f}} \left(\nabla_h \times \boldsymbol{\psi}_i^L \right) \cdot \left(\boldsymbol{n} \times \boldsymbol{\psi}_j^R \right) \mathrm{d}A, \\
\mathbf{H}_{ij}^{LR} &= \mathtt{a}_h \int_{\mathbf{f}} \left(\boldsymbol{n} \times \boldsymbol{\psi}_i^L \right) \cdot \left(\boldsymbol{n} \times \boldsymbol{\psi}_j^R \right) \mathrm{d}A.
\end{aligned}$$

The indices i and j now run between 1 and N_L and between 1 and N_R, respectively, with N_L and N_R being the number of degrees of freedom in element $\mathbf{t}_L$ and $\mathbf{t}_R$. Note that the face matrices are 'sparse' as many of the basis functions' tangential components vanish at a given interface. This is especially true for higher order elements.

We now focus on computing the lifting operators in the last term of (3.16). Using the definition of the local lifting operator (3.14) for a given face $\mathbf{f} \in \mathcal{F}$, we first recover

$$\begin{aligned}
(\mathcal{R}_{\mathbf{f}}([\![\boldsymbol{E}]\!]_T), \mathcal{R}_{\mathbf{f}}([\![\boldsymbol{\phi}]\!]_T))_\Omega &= \int_{\mathbf{f}} [\![\boldsymbol{\phi}]\!]_T \cdot \{\!\{ \mathcal{R}_{\mathbf{f}}([\![\boldsymbol{E}_h]\!]_T) \}\!\} \, \mathrm{d}A = \\
&\frac{1}{2} \int_{\mathbf{f}} \left(\boldsymbol{n}^L \times \boldsymbol{\phi}^L + \boldsymbol{n}^R \times \boldsymbol{\phi}^R \right) \cdot \left(\mathcal{R}_{\mathbf{f}}^L([\![\boldsymbol{E}_h]\!]_T) + \mathcal{R}_{\mathbf{f}}^R([\![\boldsymbol{E}_h]\!]_T) \right) \mathrm{d}A. \quad (3.23)
\end{aligned}$$

Since $\mathcal{R}_{\mathbf{f}}$ is only nonzero in the two elements $\mathbf{t}_L$ and $\mathbf{t}_R$ which are connected to the face $\mathbf{f}$, we have

$$\begin{aligned}
&\int_{\mathbf{t}_L} \boldsymbol{\phi}^L \cdot \mathcal{R}_{\mathbf{f}}^L([\![\boldsymbol{E}_h]\!]_T) \, \mathrm{d}V + \int_{\mathbf{t}_R} \boldsymbol{\phi}^R \cdot \mathcal{R}_{\mathbf{f}}^R([\![\boldsymbol{E}_h]\!]_T) \, \mathrm{d}V = \\
&\qquad \frac{1}{2} \int_{\mathbf{f}} \left(\boldsymbol{\phi}^L + \boldsymbol{\phi}^R \right) \cdot \left(\boldsymbol{n}^L \times \boldsymbol{E}_h^L + \boldsymbol{n}^R \times \boldsymbol{E}_h^R \right) \mathrm{d}A, \quad \forall \boldsymbol{\phi}^L, \boldsymbol{\phi}^R \in \boldsymbol{\Sigma}_h^p. \quad (3.24)
\end{aligned}$$

We approximate the lifting operator $\mathcal{R}_{\mathbf{f}}$–using the same basis as for the discretisation of $\boldsymbol{E}_h$–as

$$\mathcal{R}_{\mathbf{f}}^{\mathbf{t}}([\![\boldsymbol{E}_h]\!]_T)(x) = \sum_{j=1}^{N_p} R_j^{\mathbf{t},\mathbf{f}} \boldsymbol{\psi}_j^{\mathbf{t}}(x), \quad \forall x \in \mathbf{t}.$$

If we substitute these into (3.24) and use the fact that this equation must be satisfied for arbitrary test functions $\boldsymbol{\phi}^L$ and $\boldsymbol{\phi}^R$, then we obtain the following matrix relations

$$\begin{aligned}
\mathbf{M}^L R^L &= \frac{1}{2} \mathbf{D}^{LL} E^L + \frac{1}{2} \mathbf{D}^{LR} E^R, \\
\mathbf{M}^R R^R &= \frac{1}{2} \mathbf{D}^{RL} E^L + \frac{1}{2} \mathbf{D}^{RR} E^R.
\end{aligned} \quad (3.25)$$

We will use these relations to compute the last term of (3.16) during the assembly procedure. But first we briefly review the Gauss quadrature rules which we use to compute the above integrals.

3.6.4 Gauss quadratures

We evaluate the integrals by Gauss quadratures. One way to define Gauss quadratures on a tetrahedra is to compute them for the cube and 'collapse' the quadratures points (and the associated weights) into the tetrahedron. However, this procedure turns out to be very expensive for higher-order discretisation. Instead, we are making use of the *economical* Gauss quadratures, which have been derived for polynomials for orders $p \leq 9$. The construction of these points and weights is based on topological symmetries within the tetrahedron, and is considerable more complicated for orders $p > 9$. Since we implement basis functions up to polynomial degree five, the highest order quadrature rule we need (to compute the entries of the mass matrix, for example) is $p = 10$. Table I (from [74]) shows the number of quadrature points needed to integrate polynomials up to order $p \leq 13$. (The table is taken from [74] and we are not aware of any improvements on the quadrature rules since.) We can immediately see that numerical integration over a tetrahedron becomes increasingly costly, which practically prohibits the use of very high-order polynomials for three-dimensional problems. This problem can be partially circumvented by using nodal-based polynomial bases. See [39, 52, 44] for example.

Table I: *Known or predicted minimum numbers and achieved numbers of quadrature points for the Gauss integration rule over triangles and tetrahedra*

	Triangles		Tetrahedra	
Poly. order	Min.	Achieved	Min.	Achieved
1	1	1	1	1
2	3	3	4	4
3	4	4	5	5
4	6	6	11	11
5	7	7	14	14
6	12	12	24	24
7	13	13	28	31
8	16	16	40	43
9	19	19	52	53
10	24	25	68	
11	27	27		126
12	33	33		
13	36	37		210

3.6.5 The assembly

After computing the element and face matrices, we now have to assemble these matrices into a global matrix $\boldsymbol{\mathcal{A}}$. For the DG methods, one needs to assemble both element and face matrices. For the $H(\mathrm{curl})$-conforming method, only element matrices are assembled but in such a way that the tangential continuity is enforced strongly.

Assembly of the DG methods

We first loop over all the elements and for each element $\mathbf{t}$ we add the following contributions to the global matrix,

$$[\boldsymbol{\mathcal{A}}]^{\mathbf{tt}} \leftarrow \mathbf{S}^{\mathbf{t}} - \mathbf{M}^{\mathbf{t}}.$$

As a second step, we loop over all the faces. A given face $\mathbf{f}$ either connects to two elements, $\mathbf{t}_R$ and $\mathbf{t}_L$, or is a boundary face and connects to only $\mathbf{t}_L$. In any case, for the IP-DG method we add the following matrices to the corresponding part of the global matrix $\boldsymbol{\mathcal{A}}$,

$$\begin{aligned}
[\boldsymbol{\mathcal{A}}]^{LL} &\leftarrow -\frac{1}{2}\left(\mathbf{F}^{LL} + \mathbf{G}^{LL}\right) + \mathbf{H}^{LL},\\
[\boldsymbol{\mathcal{A}}]^{LR} &\leftarrow -\frac{1}{2}\left(\mathbf{F}^{LR} + \mathbf{G}^{LR}\right) + \mathbf{H}^{LR},\\
[\boldsymbol{\mathcal{A}}]^{RL} &\leftarrow -\frac{1}{2}\left(\mathbf{F}^{RL} + \mathbf{G}^{RL}\right) + \mathbf{H}^{RL},\\
[\boldsymbol{\mathcal{A}}]^{RR} &\leftarrow -\frac{1}{2}\left(\mathbf{F}^{RR} + \mathbf{G}^{RR}\right) + \mathbf{H}^{RR},
\end{aligned}$$

where $\mathbf{F}^{LR} = \left(\mathbf{G}^{RL}\right)^T$. If $\mathbf{f}$ is a boundary face, we just simply ignore all the contributions except for $[\boldsymbol{\mathcal{A}}]^{LL}$ and we replace $\frac{1}{2}$ there with 1.

For the DG method given by (3.16), the computation of the numerical flux is a bit more involved. However, if we combine (3.25) with (3.23) we can recover the contributions for the lifting operators. We add these in place of $\mathbf{H}$ matrices above:

$$\begin{aligned}
[\boldsymbol{\mathcal{A}}]^{LL} \leftarrow &-\frac{1}{2}\left(\mathbf{F}^{LL} + \mathbf{G}^{LL}\right)\\
&+\frac{\eta_F + n_f}{4}\left(\mathbf{C}^{LL}\left(\mathbf{M}^{L}\right)^{-1}\mathbf{D}^{LL} + \mathbf{C}^{LR}\left(\mathbf{M}^{R}\right)^{-1}\mathbf{D}^{RL}\right),\\
[\boldsymbol{\mathcal{A}}]^{LR} \leftarrow &-\frac{1}{2}\left(\mathbf{F}^{LR} + \mathbf{G}^{LR}\right)\\
&+\frac{\eta_F + n_f}{4}\left(\mathbf{C}^{LL}\left(\mathbf{M}^{L}\right)^{-1}\mathbf{D}^{LR} + \mathbf{C}^{LR}\left(\mathbf{M}^{R}\right)^{-1}\mathbf{D}^{RR}\right),
\end{aligned}$$

$$
\begin{aligned}
[\boldsymbol{\mathcal{A}}]^{RL} \leftarrow & -\frac{1}{2}\left(\mathbf{F}^{RL}+\mathbf{G}^{RL}\right) \\
& +\frac{\eta_F+n_f}{4}\left(\mathbf{C}^{RL}\left(\mathbf{M}^{L}\right)^{-1}\mathbf{D}^{LL}+\mathbf{C}^{RR}\left(\mathbf{M}^{R}\right)^{-1}\mathbf{D}^{RL}\right), \\
[\boldsymbol{\mathcal{A}}]^{RR} \leftarrow & -\frac{1}{2}\left(\mathbf{F}^{RR}+\mathbf{G}^{RR}\right) \\
& +\frac{\eta_F+n_f}{4}\left(\mathbf{C}^{RL}\left(\mathbf{M}^{L}\right)^{-1}\mathbf{D}^{LR}+\mathbf{C}^{RR}\left(\mathbf{M}^{R}\right)^{-1}\mathbf{D}^{RR}\right),
\end{aligned}
$$

where $\mathbf{C}^{LR}=\left(\mathbf{D}^{RL}\right)^{T}$.

We perform a similar assembly on the global right-hand side b_h to arrive at the linear system

$$\boldsymbol{\mathcal{A}}E_h = b_h, \tag{3.26}$$

where the matrix $\boldsymbol{\mathcal{A}}$ is symmetric indefinite.

Assembly of the $H(\mathrm{curl})$-conforming methods

We only need the element matrices in this case. However, we now also have to make sure that they are assembled in such a way that the tangential continuity of the discrete space is guaranteed on the whole domain Ω. Since the basis functions (3.1)– (3.5) are designed so precisely in order to satisfy that condition, we only have to carry out a few bookkeeping steps.

- Determine the total number of global degrees of freedom. Depending on the polynomial order p, we know the number of basis functions associated to each edge, face and element interior. Counting the number of edges, faces and elements is straightforward, and this information is often provided by the mesh generator.

- Assign each local basis function a global number that represents the global basis function. Two or more local basis functions should carry the same global number if they are of the same type and order, and are associated with the same edge or face.

- The local matrices and the local right-hand-side vectors are assembled so that the contributions are entered in the matrix according to their global number.

It is clear that one could do the same for the DG methods. But then every local basis function represents only a single global one so the renumbering is trivial.

The resulting linear system for a given mesh is now smaller than in the DG case. However, the mass matrix in the DG methods is block-diagonal while it is not in the $H(\mathrm{curl})$-conforming case. A block-diagonal mass matrix may not be much of an advantage when the time-harmonic Maxwell equation is discretised but it can play an important role in the time integration of the time-dependent Maxwell equations. We will return to this in more detail in Chapter 5.

3.7 Concluding remarks

We have provided a framework to implement high-order finite element methods on tetrahedral meshes when hierarchic $H(\text{curl})$-conforming basis functions are used. The framework covers the implementation of both symmetric DG methods and globally $H(\text{curl})$-conforming discretisations. The single most important contribution of this chapter is that, through a simple example, we have highlighted a possible discrepancy in the definition of one type of basis functions in the construction of the hierarchic $H(\text{curl})$-conforming basis. If this is not carefully addressed, the tangential-continuity condition in the globally $H(\text{curl})$-conforming discretisation may be breached for polynomial orders $p \geq 3$. A simple but effective solution to the problem has been proposed and the main steps of the implementation have also been discussed in details.

CHAPTER 4

OPTIMAL PENALTY PARAMETERS FOR SYMMETRIC DISCONTINUOUS GALERKIN DISCRETISATIONS OF THE TIME-HARMONIC MAXWELL EQUATIONS

4.1 Introduction

The difficulties of solving the Maxwell equations usually lie in the complexity of the geometry, the presence of material discontinuities and the fact that the curl operator has a large kernel. Moreover, the unknown fields in the Maxwell equations have special geometric characteristics. These are most pronounced in the three-dimensional version of the equations, and manifest themselves in the de Rham diagram; see e.g. [9, 45, 56]. However, many of the popular numerical discretisation techniques do not satisfy the de Rham diagram at the discrete level, and often contaminate the numerical solution by producing spurious modes. One notable exception is the $H(\mathrm{curl})$-conforming finite-element method, which makes use of special vector-valued polynomials to mimic the geometric properties of the electromagnetic fields at the discrete level. Based on the concept introduced by Whitney in the context of algebraic topology [83], they were proposed for the Maxwell system by Nédélec and Bossavit [8, 58, 59]. A hierarchic construction of high-order basis functions that satisfy the same properties are given in [2] for tetrahedral meshes and in [74] for more general three-dimensional meshes. The fact that these functions preserve the geometric properties of the Maxwell equations has motivated many to study the Maxwell system and its numerical discretisation in the framework of differential geometry [10, 45].

However, such elements suffer from a couple of practical hurdles. In particular, although they are capable of handling complex geometrical features and material

discontinuities, implementation is increasingly difficult when high-order basis functions are used. Furthermore, extending the approach to non-conforming meshes–where the local polynomial order can vary between elements and hanging nodes can be present–poses considerable difficulties.

One attractive alternative is the discontinuous Galerkin (DG) finite element method. It can handle non-conforming meshes relatively easily and the implementation of high-order basis functions is also comparatively straightforward. Research in the field of DG methods has been very active in the past ten years or so; see the recent books [26] and [44] and references therein. In the context of the Maxwell equations, a nodal approach was developed in [42], and further studied in [43]. This approach had originally been based on Lax-Friedrichs type numerical fluxes, and was later applied to the local discontinuous Galerkin method [81]. In the meantime, various DG discretisations of the low-frequency Maxwell equations [47, 48] as well as the high-frequency Maxwell equations [46, 14, 13] have also been extensively studied. The question of spurious modes in DG discretisations has been addressed in [14, 81, 13] for conforming meshes and, more recently, in [15] for two-dimensional non-conforming meshes.

In this work, we investigate the time-harmonic Maxwell equations in a lossless medium with inhomogeneous boundary conditions, i.e. find the (scaled) electric field $\boldsymbol{E} = \boldsymbol{E}(\boldsymbol{x})$ that satisfies

$$\begin{aligned} \nabla \times \frac{1}{\mu_{\mathrm{r}}} \nabla \times \boldsymbol{E} - k^2 \varepsilon_{\mathrm{r}} \boldsymbol{E} &= \boldsymbol{J} \quad \text{in } \Omega, \\ \boldsymbol{n} \times \boldsymbol{E} &= \boldsymbol{g} \quad \text{on } \Gamma, \end{aligned} \tag{4.1}$$

where Ω is an open bounded Lipschitz polyhedron on $\mathbb{R}^3$ with boundary $\Gamma = \partial\Omega$ and outward normal unit vector $\boldsymbol{n}$. The right-hand side $\boldsymbol{J}$ is the external source and k is the (real-valued) wave number with the assumption that k^2 is not a Maxwell eigenvalue. Throughout this chapter the (relative) permittivity and the (relative) permeability correspond to vacuum (or dry air). That is, we set $\varepsilon_{\mathrm{r}} = 1$ and $\mu_{\mathrm{r}} = 1$.

Out of the many different incarnations of DG discretisations for (4.1) we focus on symmetric ones, simply because they provide the possibility to use linear solvers – such as `MINRES` – that are efficient but only applicable to symmetric matrices. The symmetric interior penalty DG (IP-DG) method is probably the most popular such method thanks to the simple penalisation term in the flux formulation. However, the penalisation term grows quite sharply as the polynomial order is increased or the mesh is refined. As an alternative, one may opt for a numerical flux formulation that makes use of a local lifting operator, such as the ones introduced in [6] and [12]. These formulations, together with a large number of other flux choices, were analysed in [4] for the Laplace operator, and we refer to that work and references therein for further details.

The asymptotic convergence behaviour of the IP-DG discretisation for (4.1) was first established in [46]. In [14], the asymptotic spectral properties of the associated

eigenvalue problem

$$\begin{aligned} \nabla \times \frac{1}{\mu_{\mathrm{r}}} \nabla \times \boldsymbol{E} - k^2 \varepsilon_{\mathrm{r}} \boldsymbol{E} = \boldsymbol{0} \quad & \text{in } \Omega, \\ \boldsymbol{n} \times \boldsymbol{E} = \boldsymbol{0} \quad & \text{on } \Gamma, \end{aligned} \tag{4.2}$$

were analysed for the IP, incomplete IP, non-symmetric IP, and local DG (LDG) methods. An a priori estimate for each of these methods results as a direct corollary of the spectral analysis.

We take a slightly different approach in this chapter. If the problem is three-dimensional it is often more instructive to look at the discretisation in the pre-asymptotic regime, since in many practical applications the desired error falls into that region. Such an approach was taken in [81], where it was shown that for a given mesh the discrete eigenvalues of the symmetric LDG method tend to the $H(\mathrm{curl})$-conforming discrete eigenvalues as the penalty parameter tends to infinity. The same result is naturally valid for other symmetric DG discretisations, such as the ones considered here.

However, taking a too large penalty term comes at a computational cost. It results in a larger number of iterations when an iterative solver is used for the discrete linear system corresponding to (4.1) or (4.2). Furthermore, if that system is used as a semi-discrete system in time-domain computations, a large penalty term results in a particularly stringent time-step restriction for explicit time-integration methods. It is therefore essential that an optimal estimate for the penalty parameter be given that guarantees stability but does not significantly compromise computational efficiency.

An explicit expression of the IP parameter for the Poisson equations on simplicial meshes was derived in [70] and more recently in [25]. We extend these results to the Maxwell equation (4.1) for IP-DG and also provide an explicit expression of the DG method originally introduced in [12] as a slightly modified version of [6]. Our results are based on the trace inverse inequality [82] and on an extension of an accurate estimate for the lifting operators [69].

For our DG discretisation we use a hierarchic construction of $H(\mathrm{curl})$-conforming basis functions [2, 74]. They satisfy the global de Rham diagram in the continuous finite element setting. However, because of the discontinuous nature of the methods discussed here, we cannot expect our discretisation to be globally $H(\mathrm{curl})$-conforming and to satisfy the de Rham diagram. Nevertheless, we believe that the use of $H(\mathrm{curl})$-conforming basis function is beneficial, since it entails that the average across any face is also $H(\mathrm{curl})$-conforming. For higher-order polynomials, it also results in a sparser stiffness matrix (i.e. discrete curl-curl operator) than standard scalar H^1-conforming basis functions.

We implement the basis functions up to order five. In principle, it is possible to increase the order further, but implementation in three dimensions is hindered by a number of practical difficulties. First, high-order (i.e. $p > 9$) quadrature rules for tetrahedra are still sub-optimal and computationally expensive, making the assembly a lengthy procedure. Second, iterative solvers for indefinite linear

systems are known to converge slowly, a property exacerbated by the use of very high-order $H(\mathrm{curl})$-conforming basis functions.

The chapter is organised as follows. We define the tessellation and function spaces in Section 4.2 and derive the DG discretisation for (4.1) in Section 4.3. We derive explicit lower bounds for the penalty parameters in the DG methods and a priori upper bounds for the DG methods themselves in Section 4.4. Three-dimensional numerical computations are carried out in Section 4.5 to show the validity of the estimates. Finally, in Section 4.6, we conclude and provide an outlook.

4.2 Tessellation and function spaces

We consider a tessellation $\mathcal{T}_h$ that partitions the polyhedral domain $\Omega \subset \mathbb{R}^3$ into a set of tetrahedra $\{K\}$. Throughout the chapter we assume that the mesh is shape-regular and that each tetrahedron is straight-sided. The notations $\mathcal{F}_h$, $\mathcal{F}_h^i$ and $\mathcal{F}_h^b$ stand respectively for the set of all faces $\{F\}$, the set of all internal faces, and the set of all boundary faces. For a bounded domain $D \subset \mathbb{R}^d$, $d = 2, 3$, we denote by $H^s(D)$ the standard Sobolev space of functions with regularity exponent $s \geq 0$ and norm $\|\cdot\|_{s,D}$. When $D = \Omega$, we write $\|\cdot\|_s$. On the computational domain Ω, we introduce the space

$$H(\mathrm{curl};\Omega) := \left\{ \boldsymbol{u} \in \left[L^2(\Omega)\right]^3 : \nabla \times \boldsymbol{u} \in \left[L^2(\Omega)\right]^3 \right\},$$

with the norm $\|\boldsymbol{u}\|^2_{\mathrm{curl}} = \|\boldsymbol{u}\|^2_0 + \|\nabla \times \boldsymbol{u}\|^2_0$. Let $H_0(\mathrm{curl};\Omega)$ denote the subspace of $H(\mathrm{curl};\Omega)$ of functions with zero tangential trace. We will also use the notation $(\cdot,\cdot)_D$ for the standard inner product in $\left[L^2(D)\right]^3$,

$$(\boldsymbol{u},\boldsymbol{v})_D = \int_D \boldsymbol{u}\cdot\boldsymbol{v}\,\mathrm{d}V,$$

and the operator ∇_h for the elementwise application of $\nabla = (\partial/\partial x, \partial/\partial y, \partial/\partial z)^T$.

We now introduce the finite element space associated with the tessellation $\mathcal{T}_h$. Let $\mathcal{P}_p(K)$ be the space of polynomials of degree at most $p \geq 1$ on $K \in \mathcal{T}_h$. Over each element K the $H(\mathrm{curl})$-conforming polynomial space is defined as

$$Q^p = \left\{ \boldsymbol{u} \in [\mathcal{P}_p(K)]^3 ;\ \boldsymbol{u}_T|_{s_i} \in [\mathcal{P}_p(s_i)]^2 ;\ \boldsymbol{u}\cdot\boldsymbol{\tau}_j|_{e_j} \in \mathcal{P}_p(e_j) \right\}, \tag{4.3}$$

where s_i, $i = 1,2,3,4$ are the faces of the element; e_j, $j = 1,2,3,4,5,6$ are the edges of the element; $\mathbf{u}_T$ is the tangential component of $\mathbf{u}$; and $\boldsymbol{\tau}_j$ is the directed tangential vector on edge e_j. We define the space Σ_h^p as

$$\Sigma_h^p := \left\{ \boldsymbol{\sigma} \in [L^2(\Omega)]^3 \;\middle|\; \boldsymbol{\sigma}|_K \in Q^p,\ \forall K \in \mathcal{T}_h \right\}.$$

Consider an interface $F \in \mathcal{F}_h$ between element K^L and element K^R, and let $\boldsymbol{n}^L$ and $\boldsymbol{n}^R$ represent their respective outward pointing normal vectors. We define

the tangential jump and the average of the quantity $\boldsymbol{u}$ across interface F as

$$[\![\boldsymbol{u}]\!]_T = \boldsymbol{n}^L \times \boldsymbol{u}^L + \boldsymbol{n}^R \times \boldsymbol{u}^R \quad \text{and} \quad \{\!\{\boldsymbol{u}\}\!\} = \left(\boldsymbol{u}^L + \boldsymbol{u}^R\right)/2,$$

respectively. Here $\boldsymbol{u}^L$ and $\boldsymbol{u}^R$ are the values of the trace of $\boldsymbol{u}$ at ∂K^L and ∂K^R, respectively. At the boundary Γ, we set $\{\!\{\boldsymbol{u}\}\!\} = \boldsymbol{u}$ and $[\![\boldsymbol{u}]\!]_T = \boldsymbol{n} \times \boldsymbol{u}$. In case we only need the average of the tangential components, we use the notation $\{\!\{\mathbf{u}\}\!\}_T$.

For the analysis in Section 4.4, we also define the DG norm

$$\|\mathbf{v}\|_{\mathrm{DG}} = (\|\mathbf{v}\|_0^2 + \|\nabla_h \times \mathbf{v}\|_0^2 + \|\mathtt{h}^{-\frac{1}{2}} [\![\mathbf{v}]\!]_T \|_{0,\mathcal{F}_h}^2)^{\frac{1}{2}},$$

where $\|\cdot\|_{0,\mathcal{F}_h}$ denotes the $L^2(\mathcal{F})$ norm, and $\mathtt{h}(\boldsymbol{x}) = h_F$, which is the diameter of face F containing $\boldsymbol{x}$, i.e. $\|\mathtt{h}^{-\frac{1}{2}} [\![\mathbf{v}]\!]_T \|_{0,\mathcal{F}_h}^2 = \sum_{F \in \mathcal{F}_h} h_F \|[\![\mathbf{v}]\!]_T\|_{0,F}^2$. Similarly, h_K denotes the diameter of element K. Note that the shape-regular property of the mesh implies that there is a positive constant C_d independent of the mesh size such that for all faces F and the associated elements K^R and K^L we have

$$h_F \leq C_d \min\{h_{K^L}, h_{K^R}\}. \tag{4.4}$$

To derive the DG formulations (in the next section) we first need to introduce global lifting operators for $\boldsymbol{u} \in \Sigma_h^p$. The global lifting operator $\mathcal{L} : \left[L^2(\mathcal{F}_h^i)\right]^3 \to \Sigma_h^p$ is defined as

$$(\mathcal{L}(\boldsymbol{u}), \boldsymbol{v})_\Omega = \int_{\mathcal{F}_h^i} \boldsymbol{u} \cdot [\![\boldsymbol{v}]\!]_T \, \mathrm{d}A, \quad \forall \boldsymbol{v} \in \Sigma_h^p, \tag{4.5}$$

and the global lifting operator $\mathcal{R}\colon \left[L^2(\mathcal{F}_h)\right]^3 \to \Sigma_h^p$ as

$$(\mathcal{R}(\boldsymbol{u}), \boldsymbol{v})_\Omega = \int_{\mathcal{F}_h} \boldsymbol{u} \cdot \{\!\{\boldsymbol{v}\}\!\} \, \mathrm{d}A, \quad \forall \boldsymbol{v} \in \Sigma_h^p. \tag{4.6}$$

For a given face $F \in \mathcal{F}_h$, we will also need the local lifting operator $\mathcal{R}_F\colon \left[L^2(F)\right]^3 \to \Sigma_h^p$, defined as

$$(\mathcal{R}_F(\boldsymbol{u}), \boldsymbol{v})_\Omega = \int_F \boldsymbol{u} \cdot \{\!\{\boldsymbol{v}\}\!\} \, \mathrm{d}A, \quad \forall \boldsymbol{v} \in \Sigma_h^p. \tag{4.7}$$

Note that $\mathcal{R}_F(\boldsymbol{u})$ vanishes outside the elements connected to the face F so that for a given element $K \in \mathcal{T}_h$ we have the relation

$$\mathcal{R}(\boldsymbol{u}) = \sum_{F \in \mathcal{F}_h} \mathcal{R}_F(\boldsymbol{u}), \quad \forall \boldsymbol{u} \in \left[L^2(\mathcal{F}_h)\right]^3. \tag{4.8}$$

We also use the notation $H^r(\Omega)$ for the Sobolev space (with a possibly non-integer exponent) and the notation

$$H^r(\mathcal{T}_h) := \left\{u \in L^2(\Omega)\colon \nabla \times \boldsymbol{u}|_K \in H^r(K), \forall K \in \mathcal{T}_h\right\},$$

4.3 Discontinuous Galerkin discretisation

We now derive the DG formulation for (4.1). We first provide a general bilinear form where the choice of the numerical flux is not yet specified. Then we consider two different definitions of the numerical flux, each of which results in a symmetric algebraic system.

4.3.1 Derivation of the bilinear form

The derivation follows the same lines as the one in [80] for the Laplace operator. However, this time it is carried out for the curl-curl operator. We also refer to [4] for a unified analysis on DG methods for elliptic problems.

We first introduce the auxiliary variable $\boldsymbol{q} \in \left[L^2(\Omega)\right]^3$ so that, instead of (4.1), we can consider the first-order system

$$\begin{aligned} \nabla \times \boldsymbol{q} - k^2 \boldsymbol{E} &= \boldsymbol{J} \quad \text{in } \Omega, \\ \boldsymbol{q} &= \nabla \times \boldsymbol{E} \quad \text{in } \Omega, \\ \boldsymbol{n} \times \boldsymbol{E} &= \boldsymbol{g} \quad \text{on } \Gamma. \end{aligned} \tag{4.9}$$

From here we follow the standard DG approach (given, for example, in [4] or [80] for elliptic operators): *a)* multiply both equations in (4.9) with arbitrary test functions $\boldsymbol{\phi}, \boldsymbol{\pi} \in \Sigma_h^p$ and integrate by parts; *b)* in the element boundary integrals substitute the numerical fluxes $\boldsymbol{q}_h^*$ and $\boldsymbol{E}_h^*$ for their original counterparts; *c)* and finally integrate again the second equation in (4.9) by parts. Then we seek the pair $(\boldsymbol{E}_h, \boldsymbol{q}_h) \in \boldsymbol{\Sigma}_h^p \times \boldsymbol{\Sigma}_h^p$ such that for all test functions $(\boldsymbol{\phi}, \boldsymbol{\pi}) \in \boldsymbol{\Sigma}_h^p \times \boldsymbol{\Sigma}_h^p$:

$$(\boldsymbol{q}_h, \nabla_h \times \boldsymbol{\phi})_\Omega - k^2 (\boldsymbol{E}_h, \boldsymbol{\phi})_\Omega + \sum_{K \in \mathcal{T}_h} (\boldsymbol{n} \times \boldsymbol{q}_h^*, \boldsymbol{\phi})_{\partial K} = (\boldsymbol{J}, \boldsymbol{\phi})_\Omega \,, \tag{4.10}$$

$$(\boldsymbol{q}_h, \boldsymbol{\pi})_\Omega = (\nabla_h \times \boldsymbol{E}_h, \boldsymbol{\pi})_\Omega + \sum_{K \in \mathcal{T}_h} (\boldsymbol{n} \times (\boldsymbol{E}_h^* - \boldsymbol{E}_h), \boldsymbol{\pi})_{\partial K} \,. \tag{4.11}$$

Before we proceed, we make use of the following result: for any given $\boldsymbol{u}, \boldsymbol{v} \in \Sigma_h^p$, the identity

$$\begin{aligned} \sum_{K \in \mathcal{T}_h} (\boldsymbol{n} \times \boldsymbol{u}, \boldsymbol{v})_{\partial K} = \\ - \int_{\mathcal{F}_h^i} \{\!\{\boldsymbol{u}\}\!\} \cdot [\![\boldsymbol{v}]\!]_T \, \mathrm{d}A + \int_{\mathcal{F}_h^i} \{\!\{\boldsymbol{v}\}\!\} \cdot [\![\boldsymbol{u}]\!]_T \, \mathrm{d}A + \int_{\mathcal{F}_h^b} (\boldsymbol{n} \times \boldsymbol{u}) \cdot \boldsymbol{v} \, \mathrm{d}A \end{aligned} \tag{4.12}$$

holds. Combine this with (4.10) and (4.11) to obtain

$$\begin{aligned} (\boldsymbol{q}_h, \nabla_h \times \boldsymbol{\phi})_\Omega - k^2 (\boldsymbol{E}_h, \boldsymbol{\phi})_\Omega - \int_{\mathcal{F}_h^i} \{\!\{\boldsymbol{q}_h^*\}\!\} \cdot [\![\boldsymbol{\phi}]\!]_T \, \mathrm{d}A \\ + \int_{\mathcal{F}_h^i} \{\!\{\boldsymbol{\phi}\}\!\} \cdot [\![\boldsymbol{q}_h^*]\!]_T \, \mathrm{d}A + \int_{\mathcal{F}_h^b} (\boldsymbol{n} \times \boldsymbol{q}_h^*) \cdot \boldsymbol{\phi} \, \mathrm{d}A = (\boldsymbol{J}, \boldsymbol{\phi})_\Omega \end{aligned} \tag{4.13}$$

and

$$(\boldsymbol{q}_h, \boldsymbol{\pi})_\Omega = (\nabla_h \times \boldsymbol{E}_h, \boldsymbol{\pi})_\Omega - \int_{\mathcal{F}_h^i} \{\!\{\boldsymbol{E}_h^* - \boldsymbol{E}_h\}\!\} \cdot [\![\boldsymbol{\pi}]\!]_T \,\mathrm{d}A + \int_{\mathcal{F}_h^i} \{\!\{\boldsymbol{\pi}\}\!\} \cdot [\![\boldsymbol{E}_h^* - \boldsymbol{E}_h]\!]_T \,\mathrm{d}A + \int_{\mathcal{F}_h^b} (\boldsymbol{n} \times (\boldsymbol{E}_h^* - \boldsymbol{E}_h)) \cdot \boldsymbol{\pi} \,\mathrm{d}A. \quad (4.14)$$

We can use the lifting operators to express–and thus eliminate–the auxiliary variable $\boldsymbol{q}_h$ as a function of $\boldsymbol{E}_h$. From (4.14) and from the definition of the lifting operators (4.5) and (4.6), it follows that

$$\boldsymbol{q}_h = \nabla_h \times \boldsymbol{E}_h - \mathcal{L}(\{\!\{\boldsymbol{E}_h^* - \boldsymbol{E}_h\}\!\}) + \mathcal{R}([\![\boldsymbol{E}_h^* - \boldsymbol{E}_h]\!]_T). \quad (4.15)$$

Here we have also used the boundary definition of $[\![\cdot]\!]_T$. Substituting (4.15) into (4.13) and applying (4.11) results in the weak form

$$\begin{aligned} &\mathcal{B}(\boldsymbol{E}_h, \boldsymbol{\phi}) := (\nabla_h \times \boldsymbol{E}_h, \nabla_h \times \boldsymbol{\phi})_\Omega - k^2 (\boldsymbol{E}_h, \boldsymbol{\phi})_\Omega \\ &- \int_{\mathcal{F}_h^i} \{\!\{\boldsymbol{E}_h^* - \boldsymbol{E}_h\}\!\} \cdot [\![\nabla_h \times \boldsymbol{\phi}]\!]_T \,\mathrm{d}A + \int_{\mathcal{F}_h^i} [\![\boldsymbol{E}_h^* - \boldsymbol{E}_h]\!]_T \cdot \{\!\{\nabla_h \times \boldsymbol{\phi}\}\!\} \,\mathrm{d}A \\ &- \int_{\mathcal{F}_h^i} \{\!\{\boldsymbol{q}_h^*\}\!\} \cdot [\![\boldsymbol{\phi}]\!]_T \,\mathrm{d}A + \int_{\mathcal{F}_h^i} [\![\boldsymbol{q}_h^*]\!]_T \cdot \{\!\{\boldsymbol{\phi}\}\!\} \,\mathrm{d}A \\ &+ \int_{\mathcal{F}_h^b} (\boldsymbol{n} \times (\boldsymbol{E}_h^* - \boldsymbol{E}_h)) \cdot (\nabla_h \times \boldsymbol{\phi}) \,\mathrm{d}A - \int_{\mathcal{F}_h^b} \boldsymbol{q}_h^* \cdot (\boldsymbol{n} \times \boldsymbol{\phi}) \,\mathrm{d}A = (\boldsymbol{J}, \boldsymbol{\phi})_\Omega . \end{aligned} \quad (4.16)$$

This is the general primal formulation where one still has freedom to make choices about the numerical fluxes $\boldsymbol{E}_h^*$ and $\boldsymbol{q}_h^*$ that are most suitable for the problem. An overview of different fluxes for the Poisson equation is given in [4].

4.3.2 Numerical fluxes

At this point, we specify the numerical fluxes $\boldsymbol{E}_h^*$ and $\boldsymbol{q}_h^*$ in (4.16). We investigate two different formulations, one of which results in the IP-DG formulation that was thoroughly analysed in [46]. The other is similar to the stabilised central flux, except that in the stabilisation term we use the local lifting operator (4.7). Note that in both cases the numerical fluxes are consistent, i.e. $\forall \mathbf{E}, \boldsymbol{q} \in H(\mathrm{curl}, \Omega)$ the relations $\{\!\{\mathbf{E}\}\!\}_T = \mathbf{n} \times \mathbf{E}$, $\{\!\{\boldsymbol{q}\}\!\} = \mathbf{n} \times \boldsymbol{q}_h$, $[\![\mathbf{E}]\!]_T = \mathbf{0}$ and $[\![\boldsymbol{q}]\!]_T = \mathbf{0}$ hold.

Interior-penalty flux

First, we define the numerical fluxes so that they correspond to the IP flux,

$$\begin{aligned} &\boldsymbol{E}_h^* = \{\!\{\boldsymbol{E}_h\}\!\}, \quad &&\boldsymbol{q}_h^* = \{\!\{\nabla_h \times \boldsymbol{E}_h\}\!\} - \mathtt{a}_F [\![\boldsymbol{E}_h]\!]_T, \quad &&\text{if} \quad F \in \mathcal{F}_h^i, \\ &\boldsymbol{n} \times \boldsymbol{E}_h^* = \boldsymbol{g}, \quad &&\boldsymbol{q}_h^* = \nabla_h \times \boldsymbol{E}_h - \mathtt{a}_F (\boldsymbol{n} \times \boldsymbol{E}_h) + \mathtt{a}_F \boldsymbol{g}, \quad &&\text{if} \quad F \in \mathcal{F}_h^b, \end{aligned} \quad (4.17)$$

with $\mathtt{a}_F$ being the penalty parameter. We can now transform the following face integrals as

$$\begin{aligned}
&\int_{\mathcal{F}_h^i} [\![\boldsymbol{E}_h^* - \boldsymbol{E}_h]\!]_T \cdot \{\!\{\nabla_h \times \boldsymbol{\phi}\}\!\} \,\mathrm{d}A = -\int_{\mathcal{F}_h^i} [\![\boldsymbol{E}_h]\!]_T \cdot \{\!\{\nabla_h \times \boldsymbol{\phi}\}\!\} \,\mathrm{d}A,\\
&\int_{\mathcal{F}_h^b} (\boldsymbol{n} \times (\boldsymbol{E}_h^* - \boldsymbol{E}_h)) \cdot (\nabla_h \times \boldsymbol{\phi}) \,\mathrm{d}A = \int_{\mathcal{F}_h^b} (\boldsymbol{g} - \boldsymbol{n} \times \boldsymbol{E}_h) \cdot (\nabla_h \times \boldsymbol{\phi}) \,\mathrm{d}A,\\
&\int_{\mathcal{F}_h^i} \{\!\{\boldsymbol{q}_h^*\}\!\} \cdot [\![\boldsymbol{\phi}]\!]_T \,\mathrm{d}A = \int_{\mathcal{F}_h^i} \{\!\{\nabla_h \times \boldsymbol{E}_h\}\!\} \cdot [\![\boldsymbol{\phi}]\!]_T \,\mathrm{d}A - \int_{\mathcal{F}_h^i} \mathtt{a}_F [\![\boldsymbol{E}_h]\!]_T \cdot [\![\boldsymbol{\phi}]\!]_T \,\mathrm{d}A,\\
&\int_{\mathcal{F}_h^b} (\boldsymbol{n} \times \boldsymbol{q}_h^*) \cdot \boldsymbol{\phi} \,\mathrm{d}A = -\int_{\mathcal{F}_h^b} (\nabla_h \times \boldsymbol{E}_h) \cdot (\boldsymbol{n} \times \boldsymbol{\phi}) \,\mathrm{d}A\\
&\qquad\qquad + \int_{\mathcal{F}_h^b} \mathtt{a}_F (\boldsymbol{n} \times \boldsymbol{E}_h) \cdot (\boldsymbol{n} \times \boldsymbol{\phi}) \,\mathrm{d}A - \int_{\mathcal{F}_h^b} \mathtt{a}_F \boldsymbol{g} \cdot (\boldsymbol{n} \times \boldsymbol{\phi}) \,\mathrm{d}A,
\end{aligned}$$

while the other face integrals are zero. If we plug these back to (4.16), define the bilinear form $\mathcal{B}_h^{ip} : \Sigma_h^p \times \Sigma_h^p \to \mathbb{R}$ as

$$\begin{aligned}
\mathcal{B}_h^{ip}(\mathbf{E}_h, \boldsymbol{\phi}) :=\\
&(\nabla_h \times \boldsymbol{E}_h, \nabla_h \times \boldsymbol{\phi})_\Omega - k^2 (\boldsymbol{E}_h, \boldsymbol{\phi})_\Omega - \int_{\mathcal{F}_h} [\![\boldsymbol{E}_h]\!]_T \cdot \{\!\{\nabla_h \times \boldsymbol{\phi}\}\!\} \,\mathrm{d}A\\
&\qquad - \int_{\mathcal{F}_h} \{\!\{\nabla_h \times \boldsymbol{E}_h\}\!\} \cdot [\![\boldsymbol{\phi}]\!]_T \,\mathrm{d}A + \int_{\mathcal{F}_h} \mathtt{a}_F [\![\boldsymbol{E}]\!]_T \cdot [\![\boldsymbol{\phi}]\!]_T \,\mathrm{d}A \qquad (4.18)
\end{aligned}$$

and the linear form $\mathcal{J}_h^{ip} : \Sigma_h^p \to \mathbb{R}$ as

$$\mathcal{J}_h^{ip}(\boldsymbol{\phi}) := (\boldsymbol{J}, \boldsymbol{\phi})_\Omega - \int_{\mathcal{F}_h^b} \boldsymbol{g} \cdot (\nabla_h \times \boldsymbol{\phi}) \,\mathrm{d}A + \int_{\mathcal{F}_h^b} \mathtt{a}_F \boldsymbol{g} \cdot (\boldsymbol{n} \times \boldsymbol{\phi}) \,\mathrm{d}A, \qquad (4.19)$$

we have the IP-DG method for the time-harmonic Maxwell equations, formulated as follows. Find $\boldsymbol{E}_h \in \Sigma_h^p$ such that for all $\boldsymbol{\phi} \in \Sigma_h^p$ the relation

$$\mathcal{B}_h^{ip}(\boldsymbol{E}_h, \boldsymbol{\phi}) = \mathcal{J}_h^{ip}(\boldsymbol{\phi}) \qquad (4.20)$$

is satisfied. Note that in (4.18) we no longer distinguish explicitly between internal and boundary faces. This is permissible thanks to the definitions of the average and the tangential jump at the boundary.

Numerical flux of Brezzi formulation

As a next step, we define the numerical fluxes in the manner of Brezzi et al. [12]:

$$\begin{aligned}
&\boldsymbol{E}_h^* = \{\!\{\boldsymbol{E}_h\}\!\}, \qquad &&\boldsymbol{q}_h^* = \{\!\{\boldsymbol{q}_h\}\!\} - \alpha_{\mathcal{R}}([\![\boldsymbol{E}_h]\!]_T), \qquad &&\text{if} \quad F \in \mathcal{F}_h^i,\\
&\boldsymbol{n} \times \boldsymbol{E}_h^* = \boldsymbol{g}, \qquad &&\boldsymbol{q}_h^* = \boldsymbol{q}_h - \alpha_{\mathcal{R}}(\boldsymbol{n} \times \boldsymbol{E}_h) + \alpha_{\mathcal{R}}(\boldsymbol{g}), \qquad &&\text{if} \quad F \in \mathcal{F}_h^b. \qquad (4.21)
\end{aligned}$$

where $\alpha_{\mathcal{R}}(\boldsymbol{u}) = \eta_F \left\{\!\!\left\{\mathcal{R}_F(\boldsymbol{u}_h)\right\}\!\!\right\}$ for $F \in \mathcal{F}_h$ and $\eta_F \in \mathbb{R}^+$. Following the same line of argument as before and using (4.15), the bilinear form (4.16) now transforms as

$$\begin{aligned}\mathcal{B}(\boldsymbol{E}_h, \boldsymbol{\phi}) := (\nabla_h \times \boldsymbol{E}_h, \nabla_h \times \boldsymbol{\phi})_\Omega - k^2 (\boldsymbol{E}_h, \boldsymbol{\phi})_\Omega \\ - \int_{\mathcal{F}_h} [\![\mathbf{E}_h]\!]_T \cdot \left\{\!\!\left\{\nabla_h \times \boldsymbol{\phi}\right\}\!\!\right\} \mathrm{d}A - \int_{\mathcal{F}_h} \left\{\!\!\left\{\nabla_h \times \mathbf{E}_h\right\}\!\!\right\} \cdot [\![\boldsymbol{\phi}]\!]_T \, \mathrm{d}A \\ - \int_{\mathcal{F}_h} \left\{\!\!\left\{\mathcal{R}([\![\mathbf{E}_h^* - \mathbf{E}_h]\!]_T)\right\}\!\!\right\} \cdot [\![\boldsymbol{\phi}]\!]_T \, \mathrm{d}A + \sum_{F \in \mathcal{F}_h} \int_F \eta_F \left\{\!\!\left\{\mathcal{R}_F([\![\mathbf{E}_h]\!]_T)\right\}\!\!\right\} \cdot [\![\boldsymbol{\phi}]\!]_T \, \mathrm{d}A \\ + \int_{\mathcal{F}_h^b} \boldsymbol{g} \cdot (\nabla_h \times \boldsymbol{\phi}) \, \mathrm{d}A - \sum_{F \in \mathcal{F}_h^b} \int_F \eta_F \mathcal{R}_F(\boldsymbol{g}) \cdot (\mathbf{n} \times \boldsymbol{\phi}) \, \mathrm{d}A. \end{aligned} \tag{4.22}$$

We can now use the relation

$$\begin{aligned}\int_{\mathcal{F}_h} \left\{\!\!\left\{\mathcal{R}([\![\mathbf{E}_h^* - \mathbf{E}_h]\!]_T)\right\}\!\!\right\} \cdot [\![\boldsymbol{\phi}]\!]_T \, \mathrm{d}A &= (\mathcal{R}([\![\mathbf{E}_h^* - \mathbf{E}_h]\!]_T), \mathcal{R}([\![\boldsymbol{\phi}]\!]_T))_\Omega \\ &\approx n_f \sum_{F \in \mathcal{F}_h} (\mathcal{R}_F([\![\mathbf{E}_h^* - \mathbf{E}_h]\!]_T), \mathcal{R}_F([\![\boldsymbol{\phi}]\!]_T))_\Omega \\ &= -n_f \sum_{F \in \mathcal{F}_h^i} (\mathcal{R}_F([\![\mathbf{E}_h]\!]_T), \mathcal{R}_F([\![\boldsymbol{\phi}]\!]_T))_\Omega \\ &\quad + n_f \sum_{F \in \mathcal{F}_h^b} (\mathcal{R}_F(\boldsymbol{g} - [\![\mathbf{E}_h]\!]_T), \mathcal{R}_F([\![\boldsymbol{\phi}]\!]_T))_\Omega \\ &= -n_f \sum_{F \in \mathcal{F}_h} (\mathcal{R}_F([\![\mathbf{E}_h]\!]_T), \mathcal{R}_F([\![\boldsymbol{\phi}]\!]_T))_\Omega + n_f \sum_{F \in \mathcal{F}_h^b} (\mathcal{R}_F(\boldsymbol{g}), \mathcal{R}_F([\![\boldsymbol{\phi}]\!]_T))_\Omega \,, \end{aligned}$$

where n_f is the number of faces of an element.

Let us introduce the bilinear form $\mathcal{B}_h^{br} : \Sigma_h^p \times \Sigma_h^p \to \mathbb{R}$ and the linear form $\mathcal{J}_h^{br} : \Sigma_h^p \to \mathbb{R}$ as

$$\begin{aligned}\mathcal{B}_h^{br}(\boldsymbol{E}_h, \boldsymbol{\phi}) = (\nabla_h \times \boldsymbol{E}_h, \nabla_h \times \boldsymbol{\phi})_\Omega - k^2 (\boldsymbol{E}_h, \boldsymbol{\phi})_\Omega \\ - \int_{\mathcal{F}_h} [\![\boldsymbol{E}_h]\!]_T \cdot \left\{\!\!\left\{\nabla_h \times \boldsymbol{\phi}\right\}\!\!\right\} \mathrm{d}A - \int_{\mathcal{F}_h} \left\{\!\!\left\{\nabla_h \times \boldsymbol{E}_h\right\}\!\!\right\} \cdot [\![\boldsymbol{\phi}]\!]_T \, \mathrm{d}A \\ + \sum_{F \in \mathcal{F}_h} (\eta_F + n_f) (\mathcal{R}_F([\![\boldsymbol{E}]\!]_T), \mathcal{R}_F([\![\boldsymbol{\phi}]\!]_T))_\Omega \,, \end{aligned} \tag{4.23}$$

and

$$\begin{aligned}\mathcal{J}_h^{br}(\boldsymbol{\phi}) = (\boldsymbol{J}, \boldsymbol{\phi})_\Omega \\ - \int_{\mathcal{F}_h^b} \boldsymbol{g} \cdot (\nabla_h \times \boldsymbol{\phi}) \, \mathrm{d}A + \sum_{F \in \mathcal{F}_h^b} (\eta_F + n_f) (\mathcal{R}_F(\boldsymbol{g}), \mathcal{R}_F(\boldsymbol{n} \times \boldsymbol{\phi}))_\Omega \,, \end{aligned} \tag{4.24}$$

respectively, then the discrete formulation for the time-harmonic Maxwell equations can be written as follows. Find $\boldsymbol{E}_h \in \Sigma_h^p$ such that for all $\boldsymbol{\phi} \in \Sigma_h^p$ the relation

$$\mathcal{B}_h^{br}(\boldsymbol{E}_h, \boldsymbol{\phi}) = \mathcal{J}_h^{br}(\boldsymbol{\phi}) \tag{4.25}$$

is satisfied.

The discrete counterparts of the eigenvalue problem (4.2) for the IP and Brezzi type DG methods naturally follow from (4.20) and (4.25), i.e. find $k^2 \in \mathbb{R}_0^+$ such that for some $\boldsymbol{E}_h \in \Sigma_h^p$, respectively, $B_h^{ip}(\boldsymbol{E}_h, \boldsymbol{\phi}) = 0$ and $B_h^{br}(\boldsymbol{E}_h, \boldsymbol{\phi}) = 0$ are satisfied for all $\boldsymbol{\phi} \in \Sigma_h^p$.

4.4 Explicit parameter and error estimates

Both the IP and the Brezzi type DG formulations, given respectively by (4.20) and (4.25), contain parameters that need to be set to ensure stability. In this section, we provide explicit formulations for these parameters. First, we present an accurate lower bound for the lifting operator $\mathcal{R}_F$ on tetrahedral elements, extending the proof in [69] for hexahedra. Next, we recall the statements in [46], which are necessary for the convergence proof and keep track of all constant terms. Using these results we provide optimal penalty parameter for both the IP and the Brezzi type DG method. We also point out that these conditions are sufficient for a spurious-free convergence for the associated eigenvalue problems, discussed in [14].

In the consecutive estimates K^L and K^R denote the adjacent elements to the face $F \in \mathcal{F}_h$ and we introduce

$$M_F = \max\left\{ \frac{S(F)}{V(K^L)}, \frac{S(F)}{V(K^R)} \right\},$$

where S and V denote the surface and volume, respectively.

4.4.1 Bounds for the lifting operator

Lemma 1. *For an arbitrary face F_K of $K \in \mathcal{T}_h$ any $\mathbf{v} \in \Sigma_h^p$ satisfies the inequality*

$$\frac{2}{3}\frac{p^2}{F^2(p)}\frac{S(F_K)}{V(K)} \| [\![\mathbf{v}]\!]_T \|_{0,F_K}^2 \leq \| \mathcal{R}_F([\![\mathbf{v}]\!]_T) \|_K^2, \tag{4.26}$$

where $F^2(p) = 8\sum_{i=\frac{p}{2}}^{p} \frac{1}{2i+3}$ if p is even and $F^2(p) = \frac{8p^2}{(p+1)^2}\sum_{i=\frac{p-1}{2}+1}^{p} \frac{1}{2i+3}$ if p is odd.

Proof: The proof is divided into three steps.
Step 1 – Extension operator on the reference tetrahedron. We first consider a reference tetrahedron $\hat{K}$ with vertices $(1,1,1), (-1,1,1), (1,-1,1), (1,1,-1)$ and define an extension operator corresponding to the face $\hat{F}$ opposite to $(1,1,1)$. Let

Δ_s denote a triangle with vertices $(s,1,1),(1,s,1),(1,1,s)$. An arbitrary point (ξ,η,ζ) can be represented as

$$(\xi,\eta,\zeta) = (1,s,1) + u(0,1-s,s-1) + v(s-1,1-s,0), \tag{4.27}$$

where $0 \le u,v,u+v \le 1$ and $-1 \le s \le 1$, hence $\hat{F} = \Delta_{-1}$. The Jacobian of the mapping $(\xi,\eta,\zeta) \to (u,v,s)$ is

$$\begin{pmatrix} 0 & s-1 & v \\ 1-s & 1-s & 1-u-v \\ s-1 & 0 & u \end{pmatrix} \tag{4.28}$$

with the determinant $(1-s)^2$ and under this transformation the face $\hat{F}$ is mapped to the face $\tilde{F}$.

We now define the extension of the polynomial $\tilde{\phi}\colon \tilde{F} \to \mathbb{R}$, which is given in terms of the local coordinates (u,v). Note that the transformation $(\xi,\eta,\zeta) \to (u,v,s)$ is linear from $\hat{F}$ to $\tilde{F}$ and therefore

$$\int_{\tilde{F}} |\tilde{\phi}|^2 = \frac{S(\tilde{F})}{S(\hat{F})} \int_{\hat{F}} |\hat{\phi}|^2 = \frac{1}{4\sqrt{3}} \int_{\hat{F}} |\hat{\phi}|^2. \tag{4.29}$$

If the order p of the polynomial $\tilde{\phi}$ is even, the extension $\hat{E}(\tilde{\phi})$ is defined as

$$\hat{E}(\tilde{\phi})(u,v,s) = \frac{2}{p} \sum_{j=\frac{p}{2}+1}^{p} P_j^{(0,2)}(-s)\tilde{\phi}(u,v), \tag{4.30}$$

where $P_j^{(0,2)}$ denotes the jth-order Jacobi polynomial on $(-1,1)$ with the weight function $w(x) = (1+x)^2$ and $P_j^{(0,2)}(1) = 1$. It is also known that

$$\int_{-1}^{1} (1+x)^2 P_i^{(0,2)}(x) P_j^{(0,2)}(x)\,\mathrm{d}x = \frac{2^3 \cdot \Gamma(j+3)\Gamma(j+1)}{j! \cdot (2j+3)\Gamma(j+3)}\, 8\,\delta_{ij} = \frac{8\,\delta_{ij}}{2j+3}.$$

The identity in (4.30) gives that $\hat{E}(\tilde{\phi})(u,v,-1) = \tilde{\phi}(u,v)$. In terms of ξ,η,ζ, we have, using (4.27) with $\tilde{\phi}(u,v) = \hat{\phi}(\xi,\eta)$ that

$$\hat{E}(\hat{\phi})(\xi,\eta,\zeta) = \hat{\phi}(\xi,\eta) \qquad \text{at } \hat{F},$$

hence $\hat{E}(\hat{\phi})$ is in fact an extension of $\hat{\phi}$. Using (4.28), (4.30) and (4.29), we have

$$\begin{aligned} \int_{\hat{K}} |\hat{E}(\hat{\phi})(\xi,\eta,\zeta)|^2 &= \int_{-1}^{1}\int_{0}^{1}\int_{0}^{1-v} |\hat{E}(\tilde{\phi})(u,v,s)|^2 (1-s)^2\,\mathrm{d}u\,\mathrm{d}v\,\mathrm{d}s \\ &= \int_{-1}^{1}\int_{0}^{1}\int_{0}^{1-v} \left| \frac{2}{p} \sum_{i=\frac{p}{2}+1}^{p} P_i^{(0,2)}(-s)\tilde{\phi}(u,v) \right|^2 (1-s)^2\,\mathrm{d}u\,\mathrm{d}v\,\mathrm{d}s \end{aligned} \tag{4.31}$$

$$= \frac{4}{p^2} \int_{-1}^{1} \sum_{i=\frac{p}{2}+1}^{p} \sum_{j=\frac{p}{2}+1}^{p} P_i^{(0,2)}(-s) P_j^{(0,2)}(-s)(1-s)^2 \,\mathrm{d}s \int_0^1 \int_0^{1-v} |\tilde{\phi}(u,v)|^2 \,\mathrm{d}u \,\mathrm{d}v$$

$$= \frac{4}{p^2} \sum_{i=\frac{p}{2}+1}^{p} \frac{8}{2i+3} \int_{\tilde{F}} |\tilde{\phi}|^2 = \frac{1}{p^2} \sum_{i=\frac{p}{2}+1}^{p} \frac{8}{\sqrt{3}} \frac{1}{2i+3} \int_{\hat{F}} |\hat{\phi}|^2.$$

As a result, we obtain the relation

$$\|\hat{E}(\hat{\phi})\|_{0,\hat{K}} = \frac{1}{\sqrt[4]{3}} \frac{1}{p} F(p) \|\hat{\phi}\|_{0,\hat{F}}, \tag{4.32}$$

where

$$F^2(p) = 8 \sum_{i=\frac{p}{2}+1}^{p} \frac{1}{2i+3} \qquad \text{if } p \text{ is even.} \tag{4.33}$$

Analogously, for odd p we define the extension as

$$\hat{E}(\tilde{\phi})(u,v,s) = \frac{2}{p+1} \sum_{i=\frac{p-1}{2}+1}^{p} P_i^{(0,2)}(-s) \tilde{\phi}(u,v)$$

and the same derivation as in (4.31) gives that

$$\|\hat{E}(\hat{\phi})\|_{0,\hat{K}}^2 = \frac{1}{\sqrt{3}(p+1)^2} \sum_{i=\frac{p-1}{2}+1}^{p} \frac{8}{2i+3} \int_{\hat{F}} |\hat{\phi}|^2 = \frac{1}{\sqrt{3}} \frac{F^2(p)}{p^2} \|\hat{\phi}\|_{0,\hat{F}}^2, \tag{4.34}$$

such that we have

$$F^2(p) = \frac{8p^2}{(p+1)^2} \sum_{i=\frac{p-1}{2}+1}^{p} \frac{1}{2i+3} \qquad \text{if } p \text{ is odd.} \tag{4.35}$$

For computing the norm of the extension operator $\hat{E}$, both for odd and even p, we use the estimates

$$\sum_{i=\frac{p}{2}+1}^{p} \frac{1}{2i+3} \le \int_{\frac{p}{2}}^{p} \frac{1}{2t+3} \,\mathrm{d}t = \frac{1}{2} \ln\left(\frac{2p+3}{p+3}\right) \le \frac{1}{2} \ln 2$$

and

$$\sum_{i=\frac{p-1}{2}+1}^{p} \frac{1}{2i+3} \le \int_{\frac{p-1}{2}}^{p} \frac{1}{2t+3} \,\mathrm{d}t = \frac{1}{2} \ln\left(\frac{2p+3}{p+2}\right) \le \frac{1}{2} \ln 2$$

and obtain the simple estimate

$$F^2(p) \le 4 \ln 2. \tag{4.36}$$

The estimate in (4.36) is sharp as $\lim p \to \infty$.

Step 2 – Extension operator on a general tetrahedron. For an arbitrary tetrahedron K with a face F_K we define the affine transformation $T_K : \hat{K} \to K$ as

$$T_K(\hat{\mathbf{x}}) = J_K\hat{\mathbf{x}} + \boldsymbol{b}, \quad \text{where} \quad \boldsymbol{b} \in \mathbb{R}^3, J_K \in \mathbb{R}^{3\times 3} \text{ and } T_K(\hat{F}) = F_K.$$

The extension E of a function $\phi : F_K \to \mathbb{R}$ is given then as follows:

- We define the function $\hat{\phi} : \hat{F} \to \mathbb{R}$ with

$$\hat{\phi}(\hat{\mathbf{x}}) := \phi(T_K\hat{\mathbf{x}}).$$

- We extend $\hat{\phi}$ to $\hat{E}(\hat{\phi})$ using the method in *Step 1.*
- The extension to K is given by

$$E(\phi)(\mathbf{x}) := \hat{E}(\hat{\phi})(T_K^{-1}\mathbf{x}).$$

As J_K is linear, we can apply a simple change of variables $\mathbf{x} = T_K(\hat{\mathbf{x}})$ for computing the integral of any $g \in L_1(K)$:

$$\int_K g(\mathbf{x}) = |\det J_K| \int_{\hat{K}} \hat{g}(\hat{\mathbf{x}}) = \frac{V(K)}{V(\hat{K})} \int_{\hat{K}} \hat{g}(\hat{\mathbf{x}}). \tag{4.37}$$

Since the restriction of J_K to the face F_K of K remains affine, we also have, as in (4.29), that

$$\int_{F_K} g(\mathbf{x}) = \frac{S(F_K)}{S(\hat{F})} \int_{\hat{F}} \hat{g}(\hat{\mathbf{x}}). \tag{4.38}$$

Using (4.37) with the relations (4.32), (4.34) and (4.38) we obtain

$$\begin{aligned} \|E(\phi)\|_{0,K}^2 &= \frac{V(K)}{V(\hat{K})}\|\hat{E}(\hat{\phi})\|_{0,\hat{K}}^2 = \frac{V(K)}{V(\hat{K})}\frac{1}{\sqrt{3}}\frac{F^2(p)}{p^2}\|\hat{\phi}\|_{0,\hat{F}}^2 \\ &= \frac{V(K)}{V(\hat{K})}\frac{1}{\sqrt{3}}\frac{S(\hat{F})}{S(F_K)}\frac{F^2(p)}{p^2}\|\phi\|_{0,F_K}^2 = \frac{S(\hat{F})}{V(\hat{K})}\frac{V(K)}{S(F_K)}\frac{1}{\sqrt{3}}\frac{F^2(p)}{p^2}\|\phi\|_{0,F_K}^2. \end{aligned} \tag{4.39}$$

On the reference tetrahedron $\hat{K}$ we extended $\hat{\phi}$ from the face $\hat{F}$ with $S(\hat{F}) = 2\sqrt{3}$ and we have $V(\hat{K}) = \frac{4}{3}$, therefore (4.39) reduces to

$$\|E(\phi)\|_{0,K}^2 = \frac{3}{2}\frac{V(K)}{S(F_K)}\frac{F^2(p)}{p^2}\|\phi\|_{0,F_K}^2. \tag{4.40}$$

Step 3 – The inequality for the jump term. Using the estimate in (4.40), the definition of $\mathcal{R}_F$ in (4.7) with the fact that $E([\![\mathbf{v}]\!]_T)$ is continuous on ∂K we obtain

$$\begin{aligned} \| [\![\mathbf{v}]\!]_T \|_{0,F_K}^2 &= \int_F [\![\mathbf{v}]\!]_T \cdot E([\![\mathbf{v}]\!]_T) = \int_K \mathcal{R}_F([\![\mathbf{v}]\!]_T) \cdot E([\![\mathbf{v}]\!]_T) \\ &\le \|\mathcal{R}_F([\![\mathbf{v}]\!]_T)\|_{0,K} \left(\frac{3}{2}\frac{V(K)}{S(F_K)}\frac{F^2(p)}{p^2}\right)^{\frac{1}{2}} \| [\![\mathbf{v}]\!]_T \|_{0,F_K}, \end{aligned}$$

which gives the desired inequality. □

Remark: Since K is an arbitrary element adjacent to F_K, we can rewrite the estimate in Lemma 1 as

$$\frac{2}{3}M_F\frac{p^2}{F^2(p)}\| [\![\mathbf{v}]\!]_T \|^2_{0,F} \leq \|\mathcal{R}_F [\![\mathbf{v}]\!]_T \|^2_{0,K}. \tag{4.41}$$

In the following lemma, we will make use of the inverse trace inequality on an arbitrary face F of the element K

$$\|\mathbf{w}\|^2_{0,F} \leq \frac{(p+1)(p+3)}{3}\frac{S(F)}{V(K)}\|\mathbf{w}\|^2_{0,K} \tag{4.42}$$

in Σ_h^p, which is proved in Theorem 4 in [82].

Lemma 2. *For every face $F \in \mathcal{F}_h$ and every $\mathbf{v} \in \Sigma_h^p$ we have the inequality*

$$\|\mathcal{R}_F([\![\mathbf{v}]\!]_T)\|_0 \leq \sqrt{\frac{M_F(p+1)(p+3)}{6}}\| [\![\mathbf{v}]\!]_T \|_{0,F}. \tag{4.43}$$

Proof. The definition of the $[L_2(\Omega)]^3$ norm and the trace inequality in (4.42) give that for an arbitrary $\mathbf{v} \in \Sigma_h^p$

$$\begin{aligned}
\|\mathcal{R}_F([\![\mathbf{v}]\!]_T)\|_0 &= \sup_{\mathbf{w}\in\Sigma_h^p}\frac{\int_\Omega \mathcal{R}_F([\![\mathbf{v}]\!]_T)\cdot\mathbf{w}}{\|\mathbf{w}\|_0} = \sup_{\mathbf{w}\in\Sigma_h^p}\frac{\int_F [\![\mathbf{v}]\!]_T\cdot\{\!\{\mathbf{w}\}\!\}}{\|\mathbf{w}\|_0}\\
&\leq \sup_{\mathbf{w}\in\Sigma_h^p}\frac{\| [\![\mathbf{v}]\!]_T \|_{0,F}\left(\int_F\left(\frac{\mathbf{w}|_{\partial K^L}+\mathbf{w}|_{\partial K^R}}{2}\right)^2\right)^{\frac{1}{2}}}{\|\mathbf{w}\|_0}\\
&\leq \sup_{\mathbf{w}\in\Sigma_h^p}\frac{\| [\![\mathbf{v}]\!]_T \|_{0,F}\left(\frac{1}{2}(\|\mathbf{w}\|^2_{\partial K^L}+\|\mathbf{w}\|^2_{\partial K^R})\right)^{\frac{1}{2}}}{\|\mathbf{w}\|_0}\\
&\leq \sup_{\mathbf{w}\in\Sigma_h^p}\frac{\sqrt{M_F\frac{(p+1)(p+3)}{3}}\| [\![\mathbf{v}]\!]_T \|_{0,F}}{\|\mathbf{w}\|_0}\\
&\cdot\left(\frac{1}{2}\left(\frac{V(K^L)}{S(F)}\frac{3}{(p+1)(p+3)}\|\mathbf{w}\|^2_{\partial K^L}+\frac{V(K^R)}{S(F)}\frac{3}{(p+1)(p+3)}\|\mathbf{w}\|^2_{\partial K^R}\right)\right)^{\frac{1}{2}}\\
&\leq \sup_{\mathbf{w}\in\Sigma_h^p}\frac{\sqrt{M_F\frac{(p+1)(p+3)}{3}}\| [\![\mathbf{v}]\!]_T \|_{0,F}\left(\frac{1}{2}\left(\|\mathbf{w}\|^2_{0,K^L}+\|\mathbf{w}\|^2_{0,K^R}\right)\right)^{\frac{1}{2}}}{\|\mathbf{w}\|_0}\\
&\leq \sup_{\mathbf{w}\in\Sigma_h^p}\frac{\sqrt{M_F\frac{(p+1)(p+3)}{6}}\| [\![\mathbf{v}]\!]_T \|_{0,F}\|\mathbf{w}\|_0}{\|\mathbf{w}\|_0} = \sqrt{\frac{M_F(p+1)(p+3)}{6}}\| [\![\mathbf{v}]\!]_T \|_{0,F},
\end{aligned}$$

as stated. □

4.4.2 Gårding inequalities and continuity estimates

We begin by proving the Gårding inequality for the bilinear form of the Brezzi type DG formulation (4.25).

Lemma 3. *There exist constants $\{\eta_{F,0}\}_{F\in\mathcal{F}_h}$, independent of the discretisation parameter $h = \max_{K\in\mathcal{T}_h} \operatorname{diam} K$ and the wave number k, such that for all $\mathbf{v} \in \Sigma_h^p$ and all parameters $\eta_F \geq \eta_{F,0}$ we have the following inequality*

$$\mathcal{B}_h^{br}(\mathbf{v},\mathbf{v}) \geq \beta^2 \|\mathbf{v}\|_{\mathrm{DG}}^2 - (k^2+\beta^2)\|\mathbf{v}\|_0^2. \tag{4.44}$$

Proof. The right hand side of (4.44) can be rewritten as

$$\beta^2(\|\nabla_h \times \mathbf{v}\|_0^2 + \|\mathtt{h}^{-\frac{1}{2}} [\![\mathbf{v}]\!]_T \|_{0,\mathcal{F}_h}^2) - k^2\|\mathbf{v}\|_0^2.$$

Therefore, using (4.23) it is sufficient to prove that

$$\begin{aligned}\|\nabla_h \times \mathbf{v}\|_0^2 - 2\int_{\mathcal{F}_h} [\![\mathbf{v}]\!]_T \cdot \{\!\{\nabla_h \times \mathbf{v}\}\!\}\, \mathrm{d}A + \sum_{F\in\mathcal{F}_h} (n_f+\eta_F)\|\mathcal{R}_F([\![\mathbf{v}]\!]_T)\|_0^2 \\ \geq \beta^2(\|\nabla_h \times \mathbf{v}\|_0^2 + \|\mathtt{h}^{-\frac{1}{2}} [\![\mathbf{v}]\!]_T \|_{0,\mathcal{F}_h}^2).\end{aligned} \tag{4.45}$$

The second term on the left hand side can be estimated with any positive C_{K^L} and C_{K^R} as,

$$\begin{aligned}
&2\int_{\mathcal{F}_h} [\![\mathbf{v}]\!]_T \cdot \{\!\{\nabla_h \times \mathbf{v}\}\!\}\, \mathrm{d}A \\
&= \sum_{F\in\mathcal{F}_h} \int_F \frac{2}{\sqrt{1-\beta^2}} h_F^{-\frac{1}{2}} C_{K^L}^{-1} [\![\mathbf{v}]\!]_T \cdot C_{K^L} \frac{\sqrt{1-\beta^2}}{2} h_F^{\frac{1}{2}} \nabla_h \times \mathbf{v}^L|_F \\
&+ \frac{2}{\sqrt{1-\beta^2}} h_F^{-\frac{1}{2}} C_{K^R}^{-1} [\![\mathbf{v}]\!]_T \cdot C_{K^R} \frac{\sqrt{1-\beta^2}}{2} h_F^{\frac{1}{2}} \nabla_h \times \mathbf{v}^R|_F \,\mathrm{d}A \\
&\leq \frac{1}{1-\beta^2} \sum_{F\in\mathcal{F}_h} h_F^{-1} C_{K^L}^{-2} \| [\![\mathbf{v}]\!]_T \|_{0,F}^2 + \frac{1-\beta^2}{4} \sum_{F\in\mathcal{F}_h} h_F C_{K^L}^2 \|\nabla_h \times \mathbf{v}^L\|_{0,F}^2 \\
&+ \frac{1}{1-\beta^2} \sum_{F\in\mathcal{F}_h} h_F^{-1} C_{K^R}^{-2} \| [\![\mathbf{v}]\!]_T \|_{0,F}^2 + \frac{1-\beta^2}{4} \sum_{F\in\mathcal{F}_h} h_F C_{K^R}^2 \|\nabla_h \times \mathbf{v}^R\|_{0,F}^2.
\end{aligned} \tag{4.46}$$

Applying (4.42) to the curl terms on the right-hand side of (4.46), we obtain

$$\begin{aligned}\frac{1-\beta^2}{4} h_F C_{K^L}^2 \|\nabla_h \times \mathbf{v}^L\|_{0,F}^2 \\ \leq \frac{1-\beta^2}{4} h_F C_{K^L}^2 \frac{(p+1)(p+3)}{3} \frac{S(F)}{V(K^L)} \|\nabla_h \times \mathbf{v}^L\|_{0,K^L}^2,\end{aligned} \tag{4.47}$$

and in the same way

$$\frac{1-\beta^2}{4}h_F C_{K^R}^2 \|\nabla_h \times \mathbf{v}^R\|_{0,F}^2 \leq \frac{1-\beta^2}{4}h_F C_{K^R}^2 \frac{(p+1)(p+3)}{3}\frac{S(F)}{V(K^R)}\|\nabla_h \times \mathbf{v}^R\|_{0,K^R}^2. \quad (4.48)$$

For the jump terms, using (4.26), we obtain

$$C_{K^L}^{-2}h_F^{-1}\| [\![\mathbf{v}]\!]_T \|_{0,F}^2 \leq C_{K^L}^{-2}h_F^{-1}\frac{3}{2}\frac{V(K^L)}{S(F)}\frac{F^2(p)}{p^2}\|\mathcal{R}_F([\![\mathbf{v}]\!]_T)\|_0^2, \quad (4.49)$$

and in the same way

$$C_{K^R}^{-2}h_F^{-1}\| [\![\mathbf{v}]\!]_T \|_{0,F}^2 \leq C_{K^R}^{-2}h_F^{-1}\frac{3}{2}\frac{V(K^R)}{S(F)}\frac{F^2(p)}{p^2}\|\mathcal{R}_F([\![\mathbf{v}]\!]_T)\|_0^2. \quad (4.50)$$

Choosing

$$C_{K^L} = \sqrt{\frac{3\cdot V(K^L)}{h_F(p+1)(p+3)\cdot S(F)}} \quad \text{and} \quad C_{K^R} = \sqrt{\frac{3\cdot V(K^R)}{h_F(p+1)(p+3)\cdot S(F)}}$$

respectively, and summation of the inequalities in (4.47) and (4.48) (for all of the four faces of all tetrahedra) gives that

$$\frac{1-\beta^2}{4}\sum_{F\in\mathcal{F}_h} h_F C_{K^L}^2\|\nabla_h \times \mathbf{v}^L\|_{0,F}^2 + \frac{1-\beta^2}{4}\sum_{F\in\mathcal{F}_h} h_F C_{K^R}^2\|\nabla_h \times \mathbf{v}^R\|_{0,F}^2 \leq (1-\beta^2)\|\nabla_h \times \mathbf{v}\|_0^2 \quad (4.51)$$

and similarly, summation of (4.49) and (4.50) gives that

$$\frac{1}{1-\beta^2}\sum_{F\in\mathcal{F}_h} C_{K^L}^{-2}h_F^{-1}\| [\![\mathbf{v}]\!]_T \|_{0,F}^2 + \frac{1}{1-\beta^2}\sum_{F\in\mathcal{F}_h} C_{K^R}^{-2}h_F^{-1}\| [\![\mathbf{v}]\!]_T \|_{0,F}^2 \leq \frac{1}{1-\beta^2}\frac{F^2(p)(p+1)(p+3)}{p^2}\|\mathcal{R}([\![\mathbf{v}]\!]_T)\|_0^2. \quad (4.52)$$

Using estimates (4.51) and (4.52) in (4.46) we obtain that

$$\begin{aligned} &2\int_{\mathcal{F}_h} [\![\mathbf{v}]\!]_T \cdot \{\!\{\nabla_h \times \mathbf{v}\}\!\}\, \mathrm{d}A \\ &\leq \frac{1}{1-\beta^2}\frac{F^2(p)(p+1)(p+3)}{p^2}\|\mathcal{R}([\![\mathbf{v}]\!]_T)\|_0^2 + (1-\beta^2)\|\nabla_h \times \mathbf{v}\|_0^2. \end{aligned} \quad (4.53)$$

Therefore, using also (4.41) we can estimate the left hand side of (4.45) as

$$\|\nabla_h \times \mathbf{v}\|_0^2 - 2\int_{\mathcal{F}_h} [\![\mathbf{v}]\!]_T \cdot \{\!\{\nabla_h \times \mathbf{v}\}\!\}\, \mathrm{d}A + \sum_{F\in\mathcal{F}_h}(n_f+\eta_F)\|\mathcal{R}_F([\![\mathbf{v}]\!]_T)\|_0^2$$

$$\geq \beta^2 \|\nabla_h \times \mathbf{v}\|_0^2 + \sum_{F\in\mathcal{F}_h} \left(n_f + \eta_F - \frac{1}{1-\beta^2} \frac{F^2(p)(p+1)(p+3)}{p^2} \right) \|\mathcal{R}_F([\![\mathbf{v}]\!]_T)\|_0^2$$

$$\geq \beta^2 \|\nabla_h \times \mathbf{v}\|_0^2 + \sum_{F\in\mathcal{F}_h} h_F \left(n_f + \eta_F - \frac{1}{1-\beta^2} \frac{F^2(p)(p+1)(p+3)}{p^2} \right)$$

$$\cdot \frac{p^2}{F^2(p)} \frac{2}{3} M_F h_F^{-1} \| [\![\mathbf{v}]\!]_T \|_0^2. \quad (4.54)$$

Therefore, we have to choose η_F such that

$$h_F M_F \cdot \left(n_f + \eta_F - \frac{1}{1-\beta^2} \frac{F^2(p)(p+1)(p+3)}{p^2} \right) \frac{2}{3} \frac{p^2}{F^2(p)} \geq \beta^2 \quad (4.55)$$

and with this (4.45) is satisfied. □

Remarks:

1. Given that $n_f = 4$ for tetrahedra we can make the condition for η_F explicit,

$$\eta_{F,0} = \frac{F^2(p)}{p^2} \left(\frac{3\beta^2}{2h_F M_F} + \frac{(p+1)(p+3)}{1-\beta^2} \right) - 4. \quad (4.56)$$

2. The coercivity constant β is, however, still undefined. Using the a priori error analysis, which will be discussed in the next section, we can find an optimal value for $\eta_{F,0}$.
3. A straightforward estimation gives that $\frac{F^2(p)(p+1)(p+3)}{p^2} \geq 1$, which together with (4.55) gives that

$$n_f + \eta_F \geq 1 \qquad \text{if} \quad 0 \leq \beta^2 < 1. \quad (4.57)$$

Observe that for an arbitrary K we have $\operatorname{diam} K = h_F \geq m_F$, where F is a face of K and m_F is the height corresponding to F. Hence,

$$S(F) h_F \geq S(F) m_F = 3V(K)$$

and therefore,

$$\max_{F\in\mathcal{F}_h} h_F M_F \geq \max_{F\in\mathcal{F}_h} h_F \max \left\{ \frac{S(F)}{V(K^L)}, \frac{S(F)}{V(K^R)} \right\} \geq 3. \quad (4.58)$$

Using the method in Lemma 3 we can also obtain a bound for the penalty parameter in the interior penalty (IP) method (4.18) such that the Gårding inequality is valid.

Lemma 4. *There exist constants $\mathtt{a}_{F,0}$, independent of the discretisation parameter $h = \max_{K\in\mathcal{T}_h} \operatorname{diam} K$ and the wave number k, such that for all $\mathbf{v} \in \Sigma_h^p$ and all parameters $\mathtt{a}_F \geq \mathtt{a}_{F,0}$ we have the following inequality*

$$\mathcal{B}_h^{ip}(\mathbf{v}, \mathbf{v}) \geq \beta^2 \|\mathbf{v}\|_{\mathrm{DG}}^2 - (k^2 + \beta^2) \|\mathbf{v}\|_0^2. \quad (4.59)$$

Proof. According to the proof of Lemma 3 it is sufficient to prove that

$$\|\nabla_h \times \mathbf{v}\|_0^2 - 2\int_{\mathcal{F}_h} [\![\mathbf{v}]\!]_T \cdot \{\!\{\nabla_h \times \mathbf{v}\}\!\}\, \mathrm{d}A + \sum_{F\in\mathcal{F}_h} \mathtt{a}_F \| [\![\mathbf{v}]\!]_T \|_{0,F}^2 \geq \beta^2(\|\nabla_h \times \mathbf{v}\|_0^2 + \|\mathtt{h}^{-\frac{1}{2}} [\![\mathbf{v}]\!]_T \|_{0,\mathcal{F}_h}^2). \quad (4.60)$$

With the same choice of coefficients C_{K^L} and C_{K^R} as in Lemma 3, we obtain the inequality

$$\begin{aligned} &2\int_{\mathcal{F}_h} [\![\mathbf{v}]\!]_T \cdot \{\!\{\nabla_h \times \mathbf{v}\}\!\}\, \mathrm{d}A \\ &\quad \leq \frac{1}{1-\beta^2} \sum_{F\in\mathcal{F}_h} C_{K^L}^{-2} h_F^{-1} \| [\![\mathbf{v}]\!]_T \|_{0,F}^2 + \frac{1}{1-\beta^2} \sum_{F\in\mathcal{F}_h} C_{K^R}^{-2} h_F^{-1} \| [\![\mathbf{v}]\!]_T \|_{0,F}^2 \\ &\quad \leq \frac{1}{1-\beta^2} \frac{(p+1)(p+3)}{3} \left(\frac{S(F)}{V(K^L)} + \frac{S(F)}{V(K^R)} \right) \| [\![\mathbf{v}]\!]_T \|_{\mathcal{F}_h,0}^2. \end{aligned}$$

Substituting (4.46) into the right-hand side and using also (4.51), the left hand side of (4.60) is estimated as

$$\begin{aligned} &\|\nabla_h \times \mathbf{v}\|_0^2 - 2\int_{\mathcal{F}_h} [\![\mathbf{v}]\!]_T \cdot \{\!\{\nabla_h \times \mathbf{v}\}\!\}\, \mathrm{d}A + \sum_{F\in\mathcal{F}_h} \mathtt{a}_F \| [\![\mathbf{v}]\!]_T \|_0^2 \\ &\geq \beta^2 \|\nabla_h \times \mathbf{v}\|_0^2 \qquad (4.61) \\ &+ \sum_{F\in\mathcal{F}_h} h_F \left(\mathtt{a}_F - \frac{1}{1-\beta^2} \frac{(p+1)(p+3)}{3} \left(\frac{S(F)}{V(K^L)} + \frac{S(F)}{V(K^R)} \right) \right) h_F^{-1} \| [\![\mathbf{v}]\!]_T \|_{\mathcal{F}_h,0}^2. \end{aligned}$$

We have to choose then the parameter $\mathtt{a}_F$ on the face F such that

$$h_F \left(\mathtt{a}_F - \frac{1}{1-\beta^2} \frac{1}{3} (p+1)(p+3) \left(\frac{S(F)}{V(K^L)} + \frac{S(F)}{V(K^R)} \right) \right) \geq \beta^2,$$

which gives the explicit bound

$$\mathtt{a}_{F,0} \geq \frac{\beta^2}{h_F} + \frac{1}{1-\beta^2} \frac{1}{3} (p+1)(p+3) \left(\frac{S(F)}{V(K^L)} + \frac{S(F)}{V(K^R)} \right). \quad (4.62)$$

This proves the lemma. □

In the error analysis one has to consider (see [46]) the extended (cf. (4.23)) bilinear form

$$\mathcal{B}^{br} : (H_0(\mathrm{curl}, \Omega) + \Sigma_h^p) \times (H_0(\mathrm{curl}, \Omega) + \Sigma_h^p) \to \mathbb{R},$$

which is given as

$$\mathcal{B}^{br}(\mathbf{u}, \mathbf{v}) = (\nabla_h \times \mathbf{u}, \nabla_h \times \mathbf{v})_\Omega - k^2 (\mathbf{u}, \mathbf{v})_\Omega - \sum_{F\in\mathcal{F}_h} (\mathcal{R}_F([\![\mathbf{u}]\!]_T), \nabla_h \times \mathbf{v})_\Omega$$

$$- (\mathcal{R}_F([\![\mathbf{v}]\!]_T), \nabla_h \times \mathbf{u})_\Omega + \sum_{F\in\mathcal{F}_h} (n_f + \eta_F)(\mathcal{R}_F([\![\mathbf{u}]\!]_T), \mathcal{R}_F([\![\mathbf{v}]\!]_T))_\Omega$$

and the linear form $\mathcal{J}_h : H_0(\mathrm{curl}, \Omega) + \Sigma_h^p \to \mathbb{R}$, defined as

$$\mathcal{J}_h(\mathbf{v}) = (\mathbf{J}, \mathbf{v})_\Omega$$

when zero boundary conditions are considered. In following two lemmas we use the notation

$$\mathcal{M} = \max_{F\in\mathcal{F}_h} \sqrt{h_F M_F \frac{(p+1)(p+3)}{6}}.$$

Using (4.58) for $p \geq 1$ we have that $\mathcal{M} \geq 2$.

Using the inverse trace inequality (4.42) we also have that

$$\begin{aligned}\|\nabla_h \times \mathbf{u}^L\|_{0,F}^2 &\leq \frac{(p+1)(p+3)}{3} \frac{S(F)}{V(K^L)} \|\nabla_h \times \mathbf{u}\|_{0,K^L}^2 \\ &\leq h_F^{-1} \max_{F\in\mathcal{F}_h} M_F h_F \frac{(p+1)(p+3)}{3} \|\nabla_h \times \mathbf{u}\|_{0,K^L}^2 \\ &\leq 2h_F^{-1}\mathcal{M}^2 \|\nabla_h \times \mathbf{u}\|_{0,K^L}^2 \quad (4.63)\end{aligned}$$

and a similar estimate holds for the neighboring element K^R.

Lemma 5. *The bilinear form $\mathcal{B}^{br}$ is continuous on $(H_0(\mathrm{curl}, \Omega) + \Sigma_h^p) \times (H_0(\mathrm{curl}, \Omega) + \Sigma_h^p)$ with respect to the DG norm, i.e. the following inequality holds for all $\mathbf{u} = \mathbf{u}_0 + \mathbf{u}_h$ and $\mathbf{v} = \mathbf{v}_0 + \mathbf{v}_h$ with $\mathbf{u}_0, \mathbf{v}_0 \in H_0(\mathrm{curl}, \Omega)$ and $\mathbf{u}_h, \mathbf{v}_h \in \Sigma_h^p$:*

$$\mathcal{B}^{br}(\mathbf{u}, \mathbf{v}) \leq C \|\mathbf{u}\|_{\mathrm{DG}} \|\mathbf{v}\|_{\mathrm{DG}}, \quad (4.64)$$

where

$$C = \max_{F\in\mathcal{F}_h} \left\{ k^2, \frac{5}{4}\mathcal{M}^2 (n_f + \eta_F) \right\}.$$

Proof. Using the triangle inequality, Lemma 2, the result of the eigenvalue problem discussed in the Appendix, the estimate $\mathcal{M} \geq 2$ and (4.57) we obtain that

$$\begin{aligned}\mathcal{B}^{br}(\mathbf{u}, \mathbf{v}) &\leq |(\nabla_h \times \mathbf{u}, \nabla_h \times \mathbf{v})_\Omega| + k^2 |(\mathbf{u}, \mathbf{v})_\Omega| + \sum_{F\in\mathcal{F}_h} |(\mathcal{R}_F([\![\mathbf{u}]\!]_T), \nabla_h \times \mathbf{v})_\Omega| \\ &+ |(\mathcal{R}_F([\![\mathbf{v}]\!]_T), \nabla_h \times \mathbf{u})_\Omega| + |\sum_{F\in\mathcal{F}_h} (n_f + \eta_F)(\mathcal{R}_F([\![\mathbf{u}]\!]_T), \mathcal{R}_F([\![\mathbf{v}]\!]_T))_\Omega| \\ &\leq \|\nabla_h \times \mathbf{u}\|_0 \|\nabla_h \times \mathbf{v}\|_0 + k^2 \|\mathbf{u}\|_0 \|\mathbf{v}\|_0 + \sum_{F\in\mathcal{F}_h} \|\mathcal{R}_F([\![\mathbf{u}]\!]_T)\|_0 \|\nabla_h \times \mathbf{v}\|_0 \\ &+ \sum_{F\in\mathcal{F}_h} \|\mathcal{R}_F([\![\mathbf{v}]\!]_T)\|_0 \|\nabla_h \times \mathbf{u}\|_0 + |\sum_{F\in\mathcal{F}_h} (n_f + \eta_F) \|\mathcal{R}_F([\![\mathbf{u}]\!]_T)\|_0 \|\mathcal{R}_F([\![\mathbf{v}]\!]_T)\|_0 \\ &\leq \|\nabla_h \times \mathbf{u}\|_0 \|\nabla_h \times \mathbf{v}\|_0 + k^2 \|\mathbf{u}\|_0 \|\mathbf{v}\|_0\end{aligned}$$

$$
\begin{aligned}
&+ \mathcal{M}\|\mathtt{h}_F^{-\frac{1}{2}} [\![\mathbf{u}]\!]_T \|_{0,\mathcal{F}_h} \|\nabla_h \times \mathbf{v}\|_0 + \mathcal{M}\|\mathtt{h}_F^{-\frac{1}{2}} [\![\mathbf{v}]\!]_T \|_{0,\mathcal{F}_h} \|\nabla_h \times \mathbf{u}\|_0 \\
&+ \mathcal{M}^2 \|\mathtt{h}_F^{-\frac{1}{2}} [\![\mathbf{u}]\!]_T \|_{0,\mathcal{F}_h} \|\mathtt{h}_F^{-\frac{1}{2}} [\![\mathbf{v}]\!]_T \|_{0,\mathcal{F}_h} \max_F (n_f + \eta_F) \\
&\leq \max_{F\in\mathcal{F}_h} \left\{k^2, 1 + \mathcal{M}^2 (n_f + \eta_F)\right\} \|\mathbf{u}\|_{\mathrm{DG}} \|\mathbf{v}\|_{\mathrm{DG}} \\
&\leq \max_{F\in\mathcal{F}_h} \left\{k^2, \frac{5}{4}\mathcal{M}^2 (n_f + \eta_F)\right\} \|\mathbf{u}\|_{\mathrm{DG}} \|\mathbf{v}\|_{\mathrm{DG}},
\end{aligned}
$$

which was stated in the lemma. □

The fourth inequality in the previous lemma is obtained by solving a simple eigenvalue problem, relegated to the Appendix for the sake of readability.

A similar result can be proved for the IP method. In the analysis of the IP method one uses use extension of the discretisation operator to $(H_0(\mathrm{curl},\Omega) + \Sigma_h^p) \times (H_0(\mathrm{curl},\Omega) + \Sigma_h^p) \to \mathbb{R}$, see [46]. For this the following estimate is valid.

Lemma 6. *The bilinear form $\mathcal{B}^{ip}$ is continuous on $(H_0(\mathrm{curl},\Omega) + \Sigma_h^p) \times (H_0(\mathrm{curl},\Omega) + \Sigma_h^p)$ with respect to the DG norm, i.e. the following inequality holds for all $\mathbf{u} = \mathbf{u}_0 + \mathbf{u}_h$ and $\mathbf{v} = \mathbf{v}_0 + \mathbf{v}_h$ with $\mathbf{u}_0, \mathbf{v}_0 \in H_0(\mathrm{curl},\Omega)$ and $\mathbf{u}_h, \mathbf{v}_h \in \Sigma_h^p$:*

$$
\mathcal{B}^{ip}(\mathbf{u},\mathbf{v}) \leq C \|\mathbf{u}\|_{\mathrm{DG}} \|\mathbf{v}\|_{\mathrm{DG}},
$$

where

$$
C = \max_{F\in\mathcal{F}_h} \left\{k^2, h_F \mathtt{a}_F + \frac{3}{2}\mathcal{M}\right\}. \tag{4.65}
$$

Proof. Using the triangle inequality and (4.63), we obtain that

$$
\begin{aligned}
&\mathcal{B}^{ip}(\mathbf{u},\mathbf{v}) \leq \\
&\leq \|\nabla_h \times \mathbf{u}\|_0 \|\nabla_h \times \mathbf{v}\|_0 + k^2 \|\mathbf{u}\|_0 \|\mathbf{v}\|_0 \\
&+ \sum_{F\in\mathcal{F}_h} \|[\![\mathbf{u}]\!]_T\|_{0,F} \|\frac{1}{2}(\nabla_h \times \mathbf{v}^L + \nabla_h \times \mathbf{v}^R)\|_{0,F} \\
&+ \sum_{F\in\mathcal{F}_h} \|[\![\mathbf{v}]\!]_T\|_{0,F} \|\frac{1}{2}(\nabla_h \times \mathbf{u}^L + \nabla_h \times \mathbf{u}^R)\|_{0,F} + \sum_{F\in\mathcal{F}_h} \mathtt{a}_F \|[\![\mathbf{u}]\!]_T\|_{0,F} \|[\![\mathbf{v}]\!]_T\|_{0,F} \\
&\leq \|\nabla_h \times \mathbf{u}\|_0 \|\nabla_h \times \mathbf{v}\|_0 + k^2 \|\mathbf{u}\|_0 \|\mathbf{v}\|_0 \\
&+ \|\mathtt{h}^{-\frac{1}{2}} [\![\mathbf{u}]\!]_T\|_{0,\mathcal{F}_h} \left(\sum_{F\in\mathcal{F}_h} h_F \|\frac{1}{2}(\nabla_h \times \mathbf{u}^L + \nabla_h \times \mathbf{u}^R)\|_{0,F}^2 \right)^{\frac{1}{2}} \\
&+ \|\mathtt{h}^{-\frac{1}{2}} [\![\mathbf{v}]\!]_T\|_{0,\mathcal{F}_h} \left(\sum_{F\in\mathcal{F}_h} h_F \|\frac{1}{2}(\nabla_h \times \mathbf{v}^L + \nabla_h \times \mathbf{v}^R)\|_{0,F}^2 \right)^{\frac{1}{2}} \\
&+ \sum_{F\in\mathcal{F}_h} h_F \mathtt{a}_F \cdot h_F^{-\frac{1}{2}} \|[\![\mathbf{u}]\!]_T\|_{0,F} \cdot h_F^{-\frac{1}{2}} \|[\![\mathbf{v}]\!]_T\|_{0,F} \\
&\leq \|\nabla_h \times \mathbf{u}\|_0 \|\nabla_h \times \mathbf{v}\|_0 + k^2 \|\mathbf{u}\|_0 \|\mathbf{v}\|_0
\end{aligned}
$$

$$
\begin{aligned}
&+ \|\mathtt{h}^{-\frac{1}{2}} [\![\mathbf{u}]\!]_T \|_{0,\mathcal{F}_h} \left(\sum_{F\in\mathcal{F}_h} \frac{h_F}{2} \|\nabla_h \times \mathbf{u}^L\|^2_{0,F} + \frac{h_F}{2} \|\nabla_h \times \mathbf{u}^R\|^2_{0,F} \right)^{\frac{1}{2}} \\
&+ \|\mathtt{h}^{-\frac{1}{2}} [\![\mathbf{v}]\!]_T \|_{0,\mathcal{F}_h} \left(\sum_{F\in\mathcal{F}_h} \frac{h_F}{2} \|\nabla_h \times \mathbf{v}^L\|^2_{0,F} + \frac{h_F}{2} \|\nabla_h \times \mathbf{v}^R\|^2_{0,F} \right)^{\frac{1}{2}} \\
&+ \sum_{F\in\mathcal{F}_h} h_F \mathtt{a}_F \cdot h_F^{-\frac{1}{2}} \| [\![\mathbf{u}]\!]_T \|_{0,F} \cdot h_F^{-\frac{1}{2}} \| [\![\mathbf{v}]\!]_T \|_{0,F} \\
&\leq \|\nabla_h \times \mathbf{u}\|_0 \|\nabla_h \times \mathbf{v}\|_0 + k^2 \|\mathbf{u}\|_0 \|\mathbf{v}\|_0 \\
&+ \|\mathtt{h}^{-\frac{1}{2}} [\![\mathbf{u}]\!]_T \|_{0,\mathcal{F}_h} \left(\sum_{F\in\mathcal{F}_h} \mathcal{M}^2 \|\nabla_h \times \mathbf{u}\|^2_{0,K^L} + \mathcal{M}^2 \|\nabla_h \times \mathbf{u}\|^2_{0,K^R} \right)^{\frac{1}{2}} \\
&+ \|\mathtt{h}^{-\frac{1}{2}} [\![\mathbf{v}]\!]_T \|_{0,\mathcal{F}_h} \left(\sum_{F\in\mathcal{F}_h} \mathcal{M}^2 \|\nabla_h \times \mathbf{v}\|^2_{0,K^L} + \mathcal{M}^2 \|\nabla_h \times \mathbf{v}\|^2_{0,K^R} \right)^{\frac{1}{2}} \\
&+ \max_{F\in\mathcal{F}_h} h_F \mathtt{a}_F \sum_{F\in\mathcal{F}_h} h_F^{-\frac{1}{2}} \| [\![\mathbf{u}]\!]_T \|_{0,F} \cdot h_F^{-\frac{1}{2}} \| [\![\mathbf{v}]\!]_T \|_{0,F} \\
&\leq \|\nabla_h \times \mathbf{u}\|_0 \|\nabla_h \times \mathbf{v}\|_0 + k^2 \|\mathbf{u}\|_0 \|\mathbf{v}\|_0 + 2\mathcal{M} \|\mathtt{h}^{-\frac{1}{2}} [\![\mathbf{u}]\!]_T \|_{0,\mathcal{F}_h} \|\nabla_h \times \mathbf{u}\|_0 \\
&+ 2\mathcal{M} \|\mathtt{h}^{-\frac{1}{2}} [\![\mathbf{v}]\!]_T \|_{0,\mathcal{F}_h} \|\nabla_h \times \mathbf{v}\|_0 + \max_{F\in\mathcal{F}_h} h_F \mathtt{a}_F \|\mathtt{h}^{-\frac{1}{2}} [\![\mathbf{u}]\!]_T \|_{0,\mathcal{F}_h} \|\mathtt{h}^{-\frac{1}{2}} [\![\mathbf{v}]\!]_T \|_{0,\mathcal{F}_h} \\
&\leq \max_{F\in\mathcal{F}_h} \left\{ k^2, h_F \mathtt{a}_F + \frac{3}{2}\mathcal{M} \right\} \|\mathbf{u}\|_{\mathrm{DG}} \|\mathbf{v}\|_{\mathrm{DG}}.
\end{aligned}
$$

as stated in the lemma. □

The penultimate inequality is, again, the consequence of the a simple eigenvalue problem - see the Appendix - and the estimate $\max_{F\in\mathcal{F}} h_F \mathtt{a}_F \geq 1$, which can be proved using (4.62) with $\mathcal{M} \geq 2$.

Use now the Gårding inequality and the boundedness of $\mathcal{B}^{br}(\mathbf{u}, \mathbf{v})$ to obtain the following expression for the error,

$$
\begin{aligned}
\beta^2 \|\mathbf{E} - \mathbf{E}_h\|^2_{\mathrm{DG}} &\leq \mathcal{B}^{br}_h(\mathbf{E} - \mathbf{E}_h, \mathbf{E} - \mathbf{E}_h) + \left(k^2 + \beta^2\right) \|\mathbf{E} - \mathbf{E}_h\|^2_{0,\Omega} \\
&= \mathcal{B}^{br}_h(\mathbf{E} - \mathbf{E}_h, \mathbf{E} - \mathbf{v}) + \left(k^2 + \beta^2\right) \|\mathbf{E} - \mathbf{E}^2_h\|^2_{0,\Omega} \qquad (4.66) \\
&\leq \max_{F\in\mathcal{F}_h} \left\{ k^2, \frac{5}{4}\mathcal{M}^2 \left(n_f + \eta_F\right) \right\} \cdot \|\mathbf{E} - \mathbf{E}_h\|_{\mathrm{DG}} \|\mathbf{E} - \mathbf{v}\|_{\mathrm{DG}} \\
&+ (k^2 + \beta^2) \|\mathbf{E} - \mathbf{E}_h\|^2_{0,\Omega},
\end{aligned}
$$

where in the second line the orthogonality relation with $\mathbf{v} \in \Sigma^p_h$ was used. From this we can arrive at the estimate

$$
\begin{aligned}
\|\mathbf{E} - \mathbf{E}_h\|^2_{\mathrm{DG}} \leq \frac{1}{\beta^2} \max_{F\in\mathcal{F}_h} \left\{ k^2, \frac{5}{4}\mathcal{M}^2 \left(n_f + \eta_F\right) \right\} \inf_{\mathbf{v}\in\Sigma^p_h} \|\mathbf{E} - \mathbf{v}\|^2_{\mathrm{DG}} \\
+ \frac{k^2 + \beta^2}{\beta^2} \|\mathbf{E} - \mathbf{E}_h\|^2_{0,\Omega}. \qquad (4.67)
\end{aligned}
$$

Note that the coefficient

$$M_F h_F = \max\left\{\frac{S(F)}{V(K^L)}, \frac{S(F)}{V(K^R)}\right\} h_F = \mathcal{O}(1),$$

so the error depends on k^2, the polynomial order p and the interpolation error, which in turn depends on h and p. In addition, the coercivity constant β plays an important part too and its value is related to the penalty parameter.

4.4.3 Optimal value for the penalty parameters

The penalty parameter η_F in the Brezzi DG formulation (4.25) and the coercivity constant β in the Gårding inequality are related by (4.55) through

$$\eta_F \geq \frac{3F^2(p)}{2p^2 h_F M_F}\beta^2 + \frac{1}{1-\beta^2}\frac{(p+1)(p+3)}{p^2}F^2(p) - n_f, \tag{4.68}$$

while according to (4.67) optimal accuracy requires a minimal coefficient

$$\frac{1}{\beta^2}\max_{F\in\mathcal{F}_h}\left\{k^2, \frac{5}{4}\mathcal{M}^2\left(n_f+\eta_F\right)\right\}. \tag{4.69}$$

Take now the minimum value for η_F in (4.68) and use this in (4.69). For an optimal stabilisation, hence with a minimal effect on accuracy and efficiency, we need to minimise the second term in (4.69), i.e. the following quantity:

$$(n_f+\eta_F)\frac{(p+1)(p+3)}{6\beta^2}M_F h_F = \left(\frac{3F^2(p)}{2p^2 h_F M_F}\beta^2 + \frac{1}{1-\beta^2}\frac{(p+1)(p+3)}{p^2}F^2(p)\right)\cdot\frac{(p+1)(p+3)}{6\beta^2}M_F h_F.$$

For this we can leave all constants and find β that minimises the following

$$\frac{1}{\beta^2(1-\beta^2)}\frac{(p+1)(p+3)F^2(p)}{p^2}.$$

An elementary calculation gives that $\beta^2 = \frac{1}{2}$ such that using (4.56) we obtain the optimal value of η_F in $\mathcal{B}_h^{br}$,

$$\eta_{F,0} = \frac{F^2(p)}{p^2}\left(\frac{3}{4h_F M_F} + 2(p+1)(p+3)\right) - 4. \tag{4.70}$$

Analogously to the analysis for η_F we can find an optimal value of $\mathtt{a}_F$ using the relations (cf. (4.62))

$$\mathtt{a}_F \geq \frac{\beta^2}{h_F} + \frac{1}{1-\beta^2}\frac{1}{3}(p+1)(p+3)\left(\frac{S(F)}{V(K^L)} + \frac{S(F)}{V(K^R)}\right) \tag{4.71}$$

and minimise $\frac{h_F \mathtt{a}_F}{\beta^2}$ in (4.65) with an appropriate β. Using (4.71) we obtain

$$\frac{1}{2h_F} + \frac{1}{\beta^2(1-\beta^2)}\frac{1}{3}(p+1)(p+3)\left(\frac{S(F)}{V(K^L)} + \frac{S(F)}{V(K^R)}\right)$$

and again, we have a minimal value at $\beta^2 = \frac{1}{2}$. The optimal value of $\mathtt{a}_F$ is thus

$$\mathtt{a}_{F,0} = \frac{1}{2h_F} + \frac{2}{3}(p+1)(p+3)\left(\frac{S(F)}{V(K^L)} + \frac{S(F)}{V(K^R)}\right). \tag{4.72}$$

Note an interesting difference between the approximations using $\mathcal{B}^{ip}$ and $\mathcal{B}^{br}$ is that $\mathtt{a}_F$ in the IP-DG method needs to be increased quadratically with the polynomial order, whereas in the DG method of the Brezzi type formulation

$$\lim_{p\to\infty} \eta_F = 8\ln 2 - 4.$$

4.4.4 Convergence of the Brezzi type DG method

Using Lemma 3 and Lemma 5 one can see that with obvious modifications the analysis in [46] can be carried out for the Brezzi type bilinear form and accordingly, we obtain the following:

Theorem 1. *Assume that η_F satisfies the condition in Lemma 3 and for some parameter $s > \frac{1}{2}$ the exact solution of (4.1) satisfies*

$$\mathbf{E} \in H^s(\Omega) \qquad \text{and} \qquad \nabla \times \mathbf{E} \in H^s(\Omega).$$

Then using a full polynomial finite element space of order p with a mesh size h sufficiently small, we have the following error bound

$$\|\mathbf{E} - \mathbf{E}_h\|_{DG} \le \beta^{-2} k^2 C h^{\min\{p,s\}} \left(\|\mathbf{E}\|_{s,0} + \|\nabla \times \mathbf{E}\|_{s,0}\right), \tag{4.73}$$

where the constant C does not depend on h and k. □

Remarks:

1. The k and β dependence of the constants can be obtained in the same way as in Proposition 5.1 in [46].
2. The constant C in (4.73) depends on the coefficients in interpolation estimates, which can again depend on the geometry of the mesh and the polynomial order of the finite elements.

The results in [46] have been extended in [14], where a general framework is laid down to investigate the asymptotic spectral correctness of any DG discretisation of (4.2). Also, if a DG discretisation of (4.2) is spectrally correct (i.e. free of spurious modes), then the existence and uniqueness of the solution for the indefinite problem (4.1) is guaranteed. In order prove asymptotic spectral correctness, one only needs to check a set of conditions. These were proved for the symmetric IP-DG method in [14] on tetrahedral meshes and the results trivially extend to some other symmetric DG discretisations, including the Brezzi type considered in this chapter.

4.5 Numerical experiments

The numerical examples in this section serve two purposes. First, they intend to show how sharp the parameter estimates are in the previous section. We will see how the L^2-error and the number of iterations (i.e. computational work) changes as a function of the penalty parameter for both the IP-DG method and the method of Brezzi et al. [12]. Second, we provide asymptotic convergence tests for both methods. Although we have little to add to the theoretical results in [46, 14], our three-dimensional computations complement those results as they have so far been only verified on two-dimensional meshes [13, 15].

As a test example, we consider the Maxwell equations (4.1) with $k^2 = 1$ in the domain $\Omega = (0,1)^3$ and assume the boundary to be a perfect electric conductor (PEC), i.e. $\boldsymbol{g} = \boldsymbol{0}$ in (4.1). The source term is given as

$$\boldsymbol{J}(x,y,z) = \left(2\pi^2 - 1\right) \begin{pmatrix} \sin(\pi y)\sin(\pi z) \\ \sin(\pi z)\sin(\pi x) \\ \sin(\pi x)\sin(\pi y) \end{pmatrix}, \tag{4.74}$$

so we have the exact solution

$$\boldsymbol{E}(x,y,z) = \begin{pmatrix} \sin(\pi y)\sin(\pi z) \\ \sin(\pi z)\sin(\pi x) \\ \sin(\pi x)\sin(\pi y) \end{pmatrix}. \tag{4.75}$$

For all computations, a hierarchic construction of $H(\mathrm{curl})$-conforming vector-valued basis functions is used [2, 74]. The first six of the basis functions constitute the first-order first-family of Nédélec elements [58]. The first twelve of the basis functions used here are not the same as those that form the first-order second-family of Nédélec elements. However, they span exactly the same space and have the same approximation properties as those, defined in [59].

All numerical computations have been carried out in the framework of *hp*GEM [62], a software environment for DG discretisations suitable for a variety of physical problems. To solve the linear system that results from the DG discretisations, we use PETSc [5] and opt for `MINRES` as a suitable linear solver with incomplete Cholesky factorisation (ICC)[1] as preconditioners.

4.5.1 Sharpness of the parameter estimates

In this example, we demonstrate the sharpness of the estimates (4.70) and (4.72). A range of different values of η_F and $\mathtt{a}_F$ are used on two different meshes. One is a structured mesh of 320 tetrahedra and the other is an unstructured mesh of 432 tetrahedra. A tolerance of $\mathtt{tol} = 10^{-8}$ is used in `MINRES`, but the linear solver is stopped after 10^5 iterations even if that tolerance is not achieved.

[1] We note that ICC is not, in general, guaranteed to work for the discretisations considered here since the linear system is indefinite and Cholesky factorisation requires a positive definite matrix. However, it is successful in the following examples precisely because the factorisation is now incomplete.

For the DG method using the Brezzi formulation, we show the results on the structured mesh in Figure 4.1 and on the unstructured mesh in Figure 4.2. For the IP-DG method, Figure 4.3 depicts the results on the structured mesh and Figure 4.4 for the unstructured one. The critical parameter value is clearly visible in the plots for both methods: this is the point where the error as well as the iteration count drop dramatically. From here the error increases slightly as it converges to the error of the $H(\mathrm{curl})$-conforming discretisation – where the tangential continuity is enforced strongly through the basis function rather than weakly through the penalty term – of the same order. This convergence behaviour is a direct consequence of the theoretical and numerical study on the Maxwell eigenvalue problem in [81].

In contrast, the number of iterations increases indefinitely as the penalty parameters grow, resulting in excess computational cost. The increase is markedly steeper on the unstructured mesh than on the structured one. In each plot, bullet points indicate the theoretical estimates (4.70) and (4.72), shown to be the optimal choice in the previous section. The theoretical estimates provide a clearly stable solution with computational cost no more than two times higher than the numerically established minimum. The estimate for the penalty parameter a_F of the IP-DG method is somewhat sharper than for the penalty parameter η_F of the DG method with the Brezzi formulation. For both DG methods, the estimates for the higher-order polynomials, $p = 3$ and, especially, $p = 4$, are noticeably sharper. It is noteworthy that the estimate for a_F of the IP-DG method grows as we increase the polynomial order whereas for the DG method with the Brezzi formulation it is approximately constant. These properties are also reflected in the numerically established stability criterion.

4.5.2 Asymptotic convergence

The theoretical framework for determining the asymptotic convergence rates of DG discretisations of the Maxwell equations is fairly complete in [14], albeit for conformal meshes. However, those theoretical results have so far been accompanied by two-dimensional computations only [46, 13, 15]. We now provide numerical three-dimensional convergence results for both DG methods discussed in this work.

The computations are performed on two different sequences of meshes. The first are highly structured meshes and constructed as follows. The domain $\Omega = (0,1)^3$ is divided into $n \times n \times n$ number of congruent subcubes, with integer $n = 2^m$ and nonnegative integer m. We then divide each of these subcubes into five tetrahedra, four of which are congruent and have volume one-sixth of the original cube. The fifth has volume one-third of the original cube. Although the mesh is not uniform, this has proved to be a simple and convenient way of measuring convergence, as each time we refine the mesh, the maximum of the face diameter h_F will be exactly half of that of the previous mesh. The convergence results on structured meshes are shown in Table I for the IP-DG method and in Table III for the DG method using the Brezzi formulation.

We have also run the same example on a sequence of unstructured meshes. The

meshes were generated by CentaurSoft (`http://www.centaursoft.com`), a package suitable for generating a variety of hybrid meshes with complex geometries. In this sequence of meshes, we begin with a coarse mesh of 54 tetrahedra. Then we divide each tetrahedron into eight smaller tetrahedra to get the next (finer) mesh. The convergence results on unstructured meshes are depicted in Table II for the IP-DG discretisation and in Table IV for the DG method using the Brezzi formulation.

Based on the analysis in [46] and [13], the optimal convergence rate for this example is $\mathcal{O}(h^{p+1})$ in the $L^2(\Omega)$-norm and $\mathcal{O}(h^p)$ in the DG norm. We can see that, for both methods on structured meshes, the optimal convergence rate is achieved in the $L^2(\Omega)$-norm, and higher-than-optimal convergence rates are observed in the DG norm. On unstructured meshes, we only have an estimated convergence rate with $h \sim N_{\text{el}}^{-\frac{1}{3}}$. Here the convergence rates are slightly suboptimal, in part because we have to estimate the rates of convergence, and in part because we are still in the pre-asymptotic regime.

As a second example of asymptotic convergence, we solve the discrete eigenvalue problem that results from the DG approximation of (4.1) when the Brezzi type DG method (4.23) is used. All the eigenvalues of (4.1), corresponding to smooth eigenfunctions, are known to be

$$\omega^2 = \pi^2 \left(l^2 + m^2 + n^2\right)$$

where l, m and n are non-negative integers such that $lm + ln + nm > 0$. When $lmn > 0$, there are two identical eigenvalues associated with linearly independent eigenfunctions. Again, the analysis in [14] provides a theoretical estimate for the convergence rate of the eigenvalues. That rate is $\mathcal{O}(h^{2p})$ for both methods described here, since the eigenspaces are smooth and the discretisations symmetric. Tables V–VIII show on a sequence of uniform meshes the first twenty exact and approximate eigenvalues, representing five different values because of the multiplicity. All eigenvalues are clearly free of spurious modes in this part of the spectrum. Actually, all eigenvalues whose eigenfunctions are reasonably well-resolved (e.g. relative L^2-error of 0.1 at most) are in the 'clear' spectrum for the parameter estimates derived in Section 4.4. The approximated eigenvalues converge asymptotically at a rate predicted by the theoretical results [14] and found in two-dimensional experiments [13].

4.6 Concluding remarks and outlook

We have derived optimal penalty parameters and error estimates for symmetric discontinuous Galerkin discretisations of the time-harmonic Maxwell equations. The penalty parameters are given so that the geometric information of the mesh and the polynomial order are taken into account and therefore they are valid in the pre-asymptotic regime. This contrasts earlier results in the same field, which focused mainly on the asymptotic behaviour of the schemes. It is important that both the theoretical results and the ensuing numerical simulations consider finite

Table I: *Convergence of the IP-DG method on structured meshes*

	$p = 1$			
	$\|\mathbf{E} - \mathbf{E}_h\|_0$	Order	$\|\mathbf{E} - \mathbf{E}_h\|_{\mathrm{DG}}$	Order
$N_{\mathrm{el}} = 5$	2.5854E-01	—	4.5133E-01	—
$N_{\mathrm{el}} = 40$	2.5686E-01	0.01	3.9962E-01	0.18
$N_{\mathrm{el}} = 320$	5.8863E-02	2.13	1.1723E-01	1.78
$N_{\mathrm{el}} = 2560$	1.4605E-02	2.01	4.5535E-02	1.36
$N_{\mathrm{el}} = 20480$	3.6754E-03	1.99	2.0669E-02	1.14
	$p = 2$			
	$\|\mathbf{E} - \mathbf{E}_h\|_0$	Order	$\|\mathbf{E} - \mathbf{E}_h\|_{\mathrm{DG}}$	Order
$N_{\mathrm{el}} = 5$	2.8524E-01	—	4.1467E-01	—
$N_{\mathrm{el}} = 40$	3.1044E-02	3.20	5.0040E-02	3.05
$N_{\mathrm{el}} = 320$	3.7101E-03	3.06	8.2802E-03	2.60
$N_{\mathrm{el}} = 2560$	4.6444E-04	3.00	1.7224E-03	2.27
	$p = 3$			
	$\|\mathbf{E} - \mathbf{E}_h\|_0$	Order	$\|\mathbf{E} - \mathbf{E}_h\|_{\mathrm{DG}}$	Order
$N_{\mathrm{el}} = 5$	5.7244E-02	—	8.5302E-02	—
$N_{\mathrm{el}} = 40$	4.5008E-03	3.67	7.1218E-03	3.58
$N_{\mathrm{el}} = 320$	2.3366E-04	4.27	5.0151e-04	3.83
	$p = 4$			
	$\|\mathbf{E} - \mathbf{E}_h\|_0$	Order	$\|\mathbf{E} - \mathbf{E}_h\|_{\mathrm{DG}}$	Order
$N_{\mathrm{el}} = 5$	2.3057E-02	—	3.2834E-02	—
$N_{\mathrm{el}} = 40$	5.3477E-04	5.43	8.1995E-04	5.32
$N_{\mathrm{el}} = 320$	1.5714E-05	5.09	3.0315E-05	4.75
	$p = 5$			
	$\|\mathbf{E} - \mathbf{E}_h\|_0$	Order	$\|\mathbf{E} - \mathbf{E}_h\|_{\mathrm{DG}}$	Order
$N_{\mathrm{el}} = 5$	4.4752E-03	—	6.4666E-03	—
$N_{\mathrm{el}} = 40$	1.4442E-04	4.95	2.0711E-04	4.96
$N_{\mathrm{el}} = 320$	1.1092E-06	7.02	1.8604E-06	6.80

mesh sizes in three dimensions, because in practice three-dimensional simulations are rarely asymptotic.

The numerical examples we have presented show that the theoretical estimates are sharper for higher-order polynomials in terms of computational work, and even in the worst case they are no more than 2-3 times more expensive than the best value that we found numerically. Finally, numerical convergence results are also provided to complement the existing theoretical and lower-dimensional numerical results in literature.

Table II: *Convergence of the IP-DG on unstructured meshes*

	$p=1$			
	$\|\mathbf{E}-\mathbf{E}_h\|_0$	Order	$\|\mathbf{E}-\mathbf{E}_h\|_{\mathrm{DG}}$	Order
$N_{\mathrm{el}}=54$	2.2548E-01	—	3.6943E-01	—
$N_{\mathrm{el}}=432$	7.1925E-02	1.65	1.4363E-01	1.36
$N_{\mathrm{el}}=3456$	2.1031E-02	1.77	6.1771E-02	1.22
$N_{\mathrm{el}}=27648$	6.2947E-03	1.74	3.8283E-02	0.69
	$p=2$			
	$\|\mathbf{E}-\mathbf{E}_h\|_0$	Order	$\|\mathbf{E}-\mathbf{E}_h\|_{\mathrm{DG}}$	Order
$N_{\mathrm{el}}=54$	3.0435E-02	—	4.9090E-02	—
$N_{\mathrm{el}}=432$	4.9945E-03	2.61	1.0397E-02	2.24
$N_{\mathrm{el}}=3456$	7.2720E-04	2.78	2.4843E-03	2.07
	$p=3$			
	$\|\mathbf{E}-\mathbf{E}_h\|_0$	Order	$\|\mathbf{E}-\mathbf{E}_h\|_{\mathrm{DG}}$	Order
$N_{\mathrm{el}}=54$	4.8645E-03	—	7.9219E-03	—
$N_{\mathrm{el}}=432$	4.9752E-04	3.29	9.8238E-04	3.01
$N_{\mathrm{el}}=3456$	4.1326E-05	3.60	1.2622E-04	2.96
	$p=4$			
	$\|\mathbf{E}-\mathbf{E}_h\|_0$	Order	$\|\mathbf{E}-\mathbf{E}_h\|_{\mathrm{DG}}$	Order
$N_{\mathrm{el}}=54$	5.4669E-04	—	8.2955E-04	—
$N_{\mathrm{el}}=432$	3.7641E-05	3.86	6.3357E-05	3.71
	$p=5$			
	$\|\mathbf{E}-\mathbf{E}_h\|_0$	Order	$\|\mathbf{E}-\mathbf{E}_h\|_{\mathrm{DG}}$	Order
$N_{\mathrm{el}}=54$	1.4740E-04	—	2.1325E-04	—
$N_{\mathrm{el}}=432$	6.0287E-06	4.61	9.2191E-06	4.53

Appendix

To obtain the fourth inequalities in Lemma 5, we rewrite the expression on the left-hand side of that inequality as

$$(A\underline{\mathrm{u}},\underline{\mathrm{v}})\,,$$

where

$$\begin{pmatrix} k^2 & 0 & 0 \\ 0 & 1 & \mathcal{M} \\ 0 & \mathcal{M} & \max\limits_{F\in\mathcal{F}_h}(n_f+\eta_F)\,\mathcal{M}^2 \end{pmatrix},$$

and

$$\underline{\mathrm{u}}=\left(\|\mathbf{u}\|_0,\|\nabla_h\times\mathbf{u}\|_0,\|h_F^{-\frac{1}{2}}[\![\mathbf{u}]\!]_T\|_{0,\mathcal{F}_h}\right)^T \quad\Longrightarrow\quad |\underline{\mathrm{u}}|=\|\mathbf{u}\|_{\mathrm{DG}},$$

Table III: *Convergence of the method of DG method using the Brezzi formulation on structured meshes*

	$p=1$			
	$\|\mathbf{E}-\mathbf{E}_h\|_0$	Order	$\|\mathbf{E}-\mathbf{E}_h\|_{\mathrm{DG}}$	Order
$N_{\mathrm{el}}=5$	5.2216E-01	—	7.4201E-01	—
$N_{\mathrm{el}}=40$	3.0615E-01	0.77	4.3594E-01	0.77
$N_{\mathrm{el}}=320$	7.1871E-02	2.09	1.0625E-01	2.04
$N_{\mathrm{el}}=2560$	1.7673E-02	2.02	2.9920E-02	1.83
$N_{\mathrm{el}}=20480$	4.4003E-03	2.01	1.0473E-02	1.51
	$p=2$			
	$\|\mathbf{E}-\mathbf{E}_h\|_0$	Order	$\|\mathbf{E}-\mathbf{E}_h\|_{\mathrm{DG}}$	Order
$N_{\mathrm{el}}=5$	3.0892E-01	—	4.3901E-01	—
$N_{\mathrm{el}}=40$	3.3887E-02	3.19	4.9367E-02	3.15
$N_{\mathrm{el}}=320$	4.0850E-03	3.05	6.7364E-03	2.87
$N_{\mathrm{el}}=2560$	5.0782E-04	3.01	1.1718E-03	2.52
	$p=3$			
	$\|\mathbf{E}-\mathbf{E}_h\|_0$	Order	$\|\mathbf{E}-\mathbf{E}_h\|_{\mathrm{DG}}$	Order
$N_{\mathrm{el}}=5$	6.4391E-02	—	9.1864E-02	—
$N_{\mathrm{el}}=40$	4.7730E-03	3.75	6.9565E-03	3.72
$N_{\mathrm{el}}=320$	2.4716E-04	4.27	4.3197E-04	4.01
	$p=4$			
	$\|\mathbf{E}-\mathbf{E}_h\|_0$	Order	$\|\mathbf{E}-\mathbf{E}_h\|_{\mathrm{DG}}$	Order
$N_{\mathrm{el}}=5$	2.3335E-02	—	3.3088E-02	—
$N_{\mathrm{el}}=40$	5.5087E-04	5.40	8.1681E-04	5.34
$N_{\mathrm{el}}=320$	1.6179E-05	5.09	2.8348E-05	4.85
	$p=5$			
	$\|\mathbf{E}-\mathbf{E}_h\|_0$	Order	$\|\mathbf{E}-\mathbf{E}_h\|_{\mathrm{DG}}$	Order
$N_{\mathrm{el}}=5$	4.3251E-03	—	6.1734E-03	—
$N_{\mathrm{el}}=40$	1.4449E-04	4.90	2.0586E-04	4.91
$N_{\mathrm{el}}=320$	1.1041E-06	7.03	1.8247E-06	6.82

$$\underline{\mathrm{v}} = \left(\|\mathbf{v}\|_0, \|\nabla_h \times \mathbf{v}\|_0, \|h_F^{-\frac{1}{2}} [\![\mathbf{v}]\!]_T \|_{0,\mathcal{F}_h} \right)^T \quad \Longrightarrow \quad |\underline{\mathrm{v}}| = \|\mathbf{v}\|_{\mathrm{DG}}.$$

Since A is symmetric we have

$$(A\underline{\mathrm{u}}, \underline{\mathrm{v}}) \leq \max_{\lambda \in \mathrm{eig}(A)} |\lambda| \, |\underline{\mathrm{u}}| \, |\underline{\mathrm{v}}| = \max_{\lambda \in \mathrm{eig}(A)} |\lambda| \, \|\mathbf{u}\| \, \|\mathbf{v}\| \,,$$

from which a straightforward computation gives

$$\max_{\lambda \in \mathrm{eig}(A)} |\lambda| = \max \left\{ k^2, \lambda_2 \right\},$$

Table IV: *Convergence of the DG method using the Brezzi formulation on unstructured meshes*

	$p=1$			
	$\|\mathbf{E}-\mathbf{E}_h\|_0$	Order	$\|\mathbf{E}-\mathbf{E}_h\|_{\mathrm{DG}}$	Order
$N_{\mathrm{el}}=54$	2.9871E-01	—	4.2626E-01	—
$N_{\mathrm{el}}=432$	9.4108E-02	1.67	1.3758E-01	1.63
$N_{\mathrm{el}}=3456$	2.7543E-02	1.77	4.3294E-02	1.67
$N_{\mathrm{el}}=27648$	8.3263E-03	1.73	1.5441E-02	1.49
	$p=2$			
	$\|\mathbf{E}-\mathbf{E}_h\|_0$	Order	$\|\mathbf{E}-\mathbf{E}_h\|_{\mathrm{DG}}$	Order
$N_{\mathrm{el}}=54$	3.3293E-02	—	4.8203E-02	—
$N_{\mathrm{el}}=432$	5.4652E-03	2.61	8.4958E-03	2.50
$N_{\mathrm{el}}=3456$	7.9569E-04	2.78	1.5428E-03	2.46
	$p=3$			
	$\|\mathbf{E}-\mathbf{E}_h\|_0$	Order	$\|\mathbf{E}-\mathbf{E}_h\|_{\mathrm{DG}}$	Order
$N_{\mathrm{el}}=54$	5.2936E-03	—	7.7574E-03	—
$N_{\mathrm{el}}=432$	5.2925E-04	3.32	8.3911E-04	3.21
$N_{\mathrm{el}}=3456$	4.3710E-05	3.60	8.7359E-05	3.26
	$p=4$			
	$\|\mathbf{E}-\mathbf{E}_h\|_0$	Order	$\|\mathbf{E}-\mathbf{E}_h\|_{\mathrm{DG}}$	Order
$N_{\mathrm{el}}=54$	5.6374E-04	—	8.2022E-04	—
$N_{\mathrm{el}}=432$	3.8520E-05	3.87	5.8694E-05	3.80
	$p=5$			
	$\|\mathbf{E}-\mathbf{E}_h\|_0$	Order	$\|\mathbf{E}-\mathbf{E}_h\|_{\mathrm{DG}}$	Order
$N_{\mathrm{el}}=54$	1.4759E-04	—	2.1091E-04	—
$N_{\mathrm{el}}=432$	6.0329E-06	4.61	8.8707E-06	4.57

where λ_2 is the solution of the equation

$$0=(1-\lambda_2)(\max_{F\in\mathcal{F}_h}(n_f+\eta_F)\,\mathcal{M}^2-\lambda_2)-\mathcal{M}^2.$$

Using (4.57) the larger solution can be estimated as

$$\begin{aligned}&\frac{1}{2}\max_{F\in\mathcal{F}_h}(n_f+\eta_F)\,\mathcal{M}^2+1+\\&\frac{1}{2}\sqrt{(\max_{F\in\mathcal{F}_h}(n_f+\eta_F)\,\mathcal{M}^2+1)^2-4\max_{F\in\mathcal{F}_h}(n_f+\eta_F)\,\mathcal{M}^2+4\mathcal{M}^2}\\&\leq\frac{1}{2}\left(\max_{F\in\mathcal{F}_h}(n_f+\eta_F)\,\mathcal{M}^2+1+\sqrt{(\max_{F\in\mathcal{F}_h}(n_f+\eta_F)\,\mathcal{M}^2+1)^2-4\mathcal{M}^2+4\mathcal{M}^2}\right)\end{aligned}\tag{4.76}$$

Table V: *Eigenvalues (divided by π^2) obtained on uniform meshes with $p = 1$*

h	$h/2$	$h/4$	$h/8$	$h/16$	Expected
3.1339	2.2578	2.0747	2.0192	2.0048	2.0000
3.1339	2.2578	2.0747	2.0192	2.0048	2.0000
3.1339	2.2578	2.0747	2.0192	2.0048	2.0000
5.2780	3.7951	3.1682	3.0431	3.0108	3.0000
5.7352	3.7951	3.1682	3.0431	3.0108	3.0000
5.7352	5.5034	5.4426	5.1182	5.0300	5.0000
8.5813	5.5034	5.4426	5.1182	5.0300	5.0000
8.5813	5.5034	5.4426	5.1182	5.0300	5.0000
8.5813	7.8215	5.4426	5.1182	5.0300	5.0000
9.7578	7.8215	5.4426	5.1182	5.0300	5.0000
9.7578	7.8215	5.4426	5.1182	5.0300	5.0000
9.7578	8.2393	6.6343	6.1695	6.0430	6.0000
11.7469	8.2393	6.6343	6.1695	6.0430	6.0000
11.7469	8.2393	6.6343	6.1695	6.0430	6.0000
11.7469	9.1638	6.6442	6.1707	6.0432	6.0000
13.3385	9.1638	6.6442	6.1707	6.0432	6.0000
17.2489	9.1638	6.6442	6.1707	6.0432	6.0000
17.2489	12.2453	9.0311	8.2990	8.0768	8.0000
17.2489	12.2453	9.0311	8.2990	8.0768	8.0000
17.4002	12.2453	9.0311	8.2990	8.0768	8.0000

$$= \max_{F \in \mathcal{F}_h} (n_f + \eta_F)\, \mathcal{M}^2 + 1,$$

hence

$$\max_{\lambda \in \mathrm{eig}(A)} |\lambda| \leq \max_{F \in \mathcal{F}_h} \left\{ k^2, (n_f + \eta_F)\, \mathcal{M}^2 + 1 \right\}.$$

Similarly, the last inequality in Lemma 6 is obtained through defining the matrix

$$\begin{pmatrix} k^2 & 0 & 0 \\ 0 & 1 & 2\mathcal{M} \\ 0 & 2\mathcal{M} & \max\limits_{F \in \mathcal{F}_h} h_F \mathtt{a}_F \end{pmatrix},$$

after which a simple calculation and using the inequality $h_F \mathtt{a}_F \geq 1$ yields

$$\max_{\lambda \in \mathrm{eig}(A)} |\lambda| = \max_{F \in \mathcal{F}_h} \left\{ k^2, h_F \mathtt{a}_F + \frac{3}{2}\mathcal{M} \right\}.$$

Table VI: *Eigenvalues (divided by π^2) obtained on uniform meshes with $p = 2$*

h	$h/2$	$h/4$	$h/8$	Expected
2.1270	2.0197	2.0014	2.0001	2.0000
2.1270	2.0197	2.0014	2.0001	2.0000
2.1270	2.0197	2.0014	2.0001	2.0000
3.8664	3.0219	3.0047	3.0003	3.0000
3.8664	3.0219	3.0047	3.0003	3.0000
5.9288	5.1348	5.0196	5.0013	5.0000
5.9288	5.1348	5.0196	5.0013	5.0000
5.9288	5.1348	5.0196	5.0013	5.0000
6.8030	5.2479	5.0196	5.0013	5.0000
6.8030	5.2479	5.0196	5.0013	5.0000
6.8030	5.2479	5.0196	5.0013	5.0000
8.8557	6.3128	6.0335	6.0023	6.0000
9.2855	6.3128	6.0335	6.0023	6.0000
9.2855	6.3128	6.0335	6.0023	6.0000
9.2855	6.4152	6.0352	6.0024	6.0000
11.5504	6.4152	6.0352	6.0024	6.0000
11.5504	6.4152	6.0352	6.0024	6.0000
11.5504	8.5082	8.0789	8.0056	8.0000
14.8586	8.5082	8.0789	8.0056	8.0000
14.8586	8.5082	8.0789	8.0056	8.0000

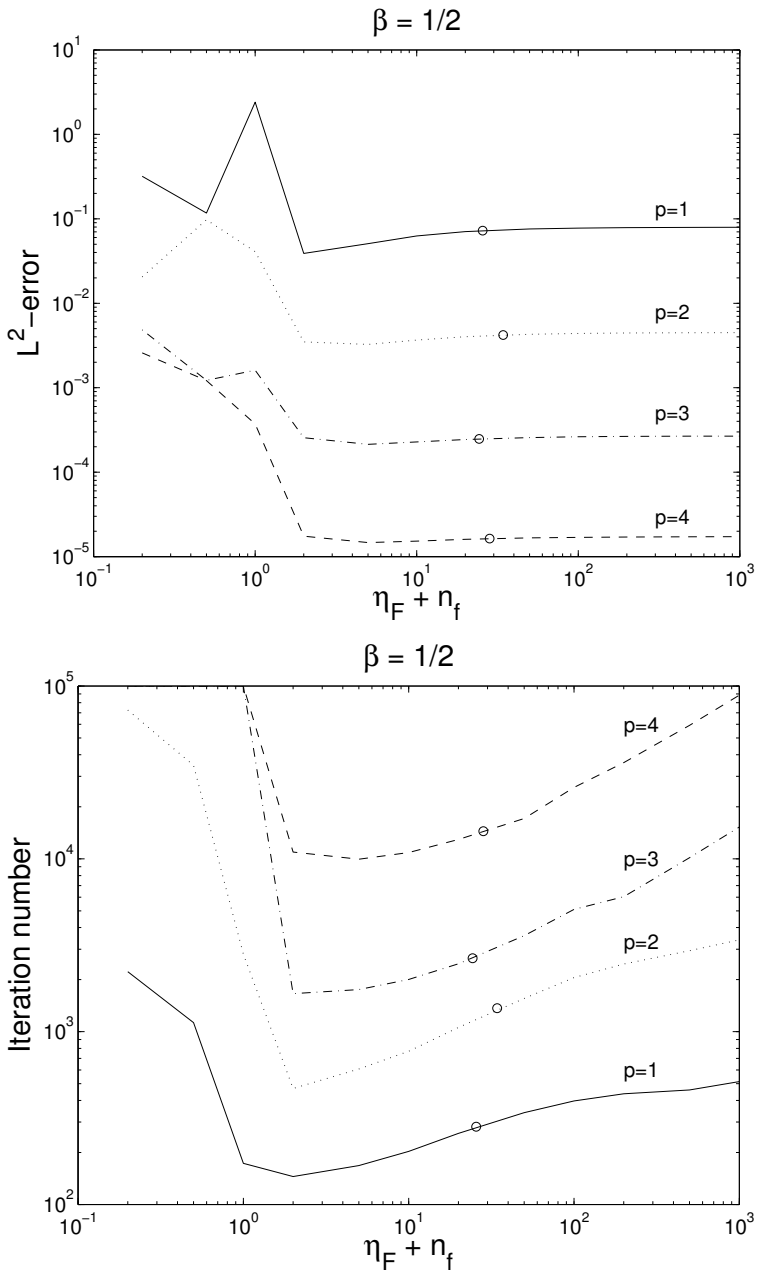


Figure 4.1: *L^2-error (left) and the number of* **MINRES** *iterations (right) as a function of the penalty parameter $\eta_F + n_f$ in the DG formulation of Brezzi. A structured mesh of 320 tetrahedra and coercivity constant $\beta = \frac{1}{2}$ are used.*

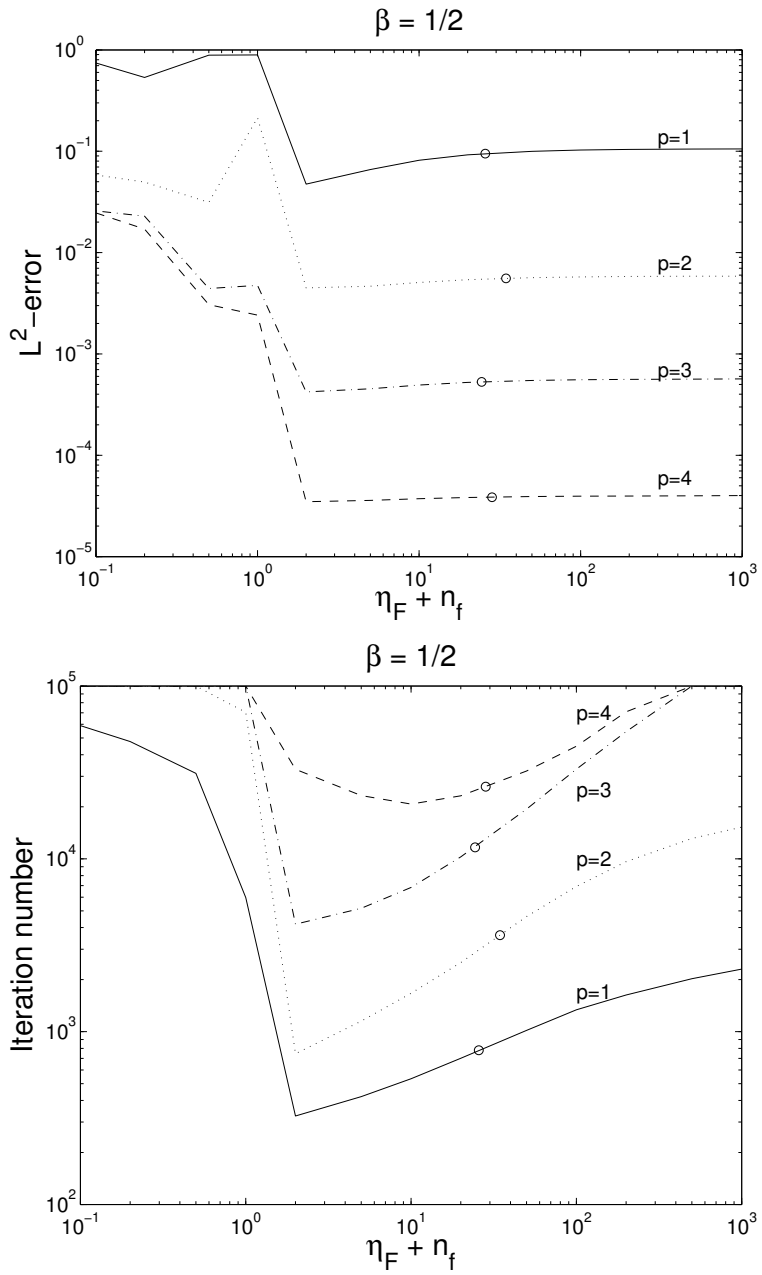


Figure 4.2: *L^2-error (left) and the number of* MINRES *iterations (right) as a function of the penalty parameter $\eta_F + n_f$ in the DG formulation of Brezzi. An unstructured mesh of 432 tetrahedra and coercivity constant $\beta = \frac{1}{2}$ are used.*

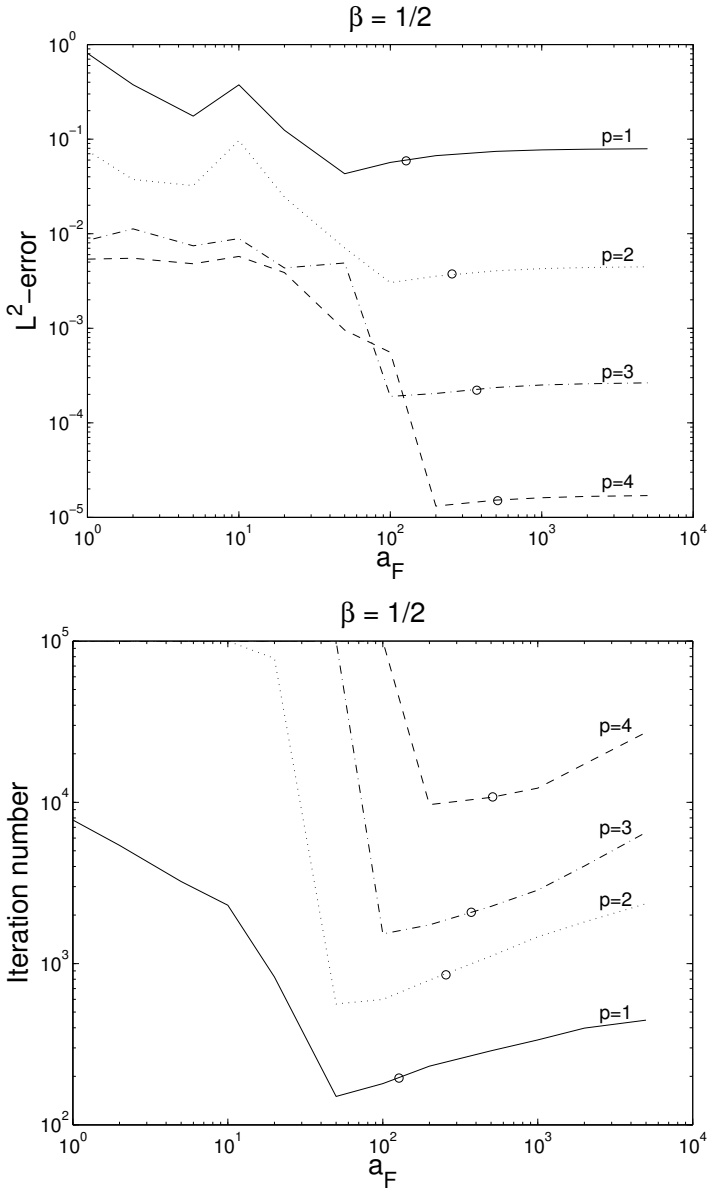


Figure 4.3: *L^2-error (left) and the number of* `MINRES` *iterations (right) as a function of the penalty parameter* a_F *in the IP-DG method. A structured mesh of 320 tetrahedra and coercivity constant* $\beta = \frac{1}{2}$ *are used.*

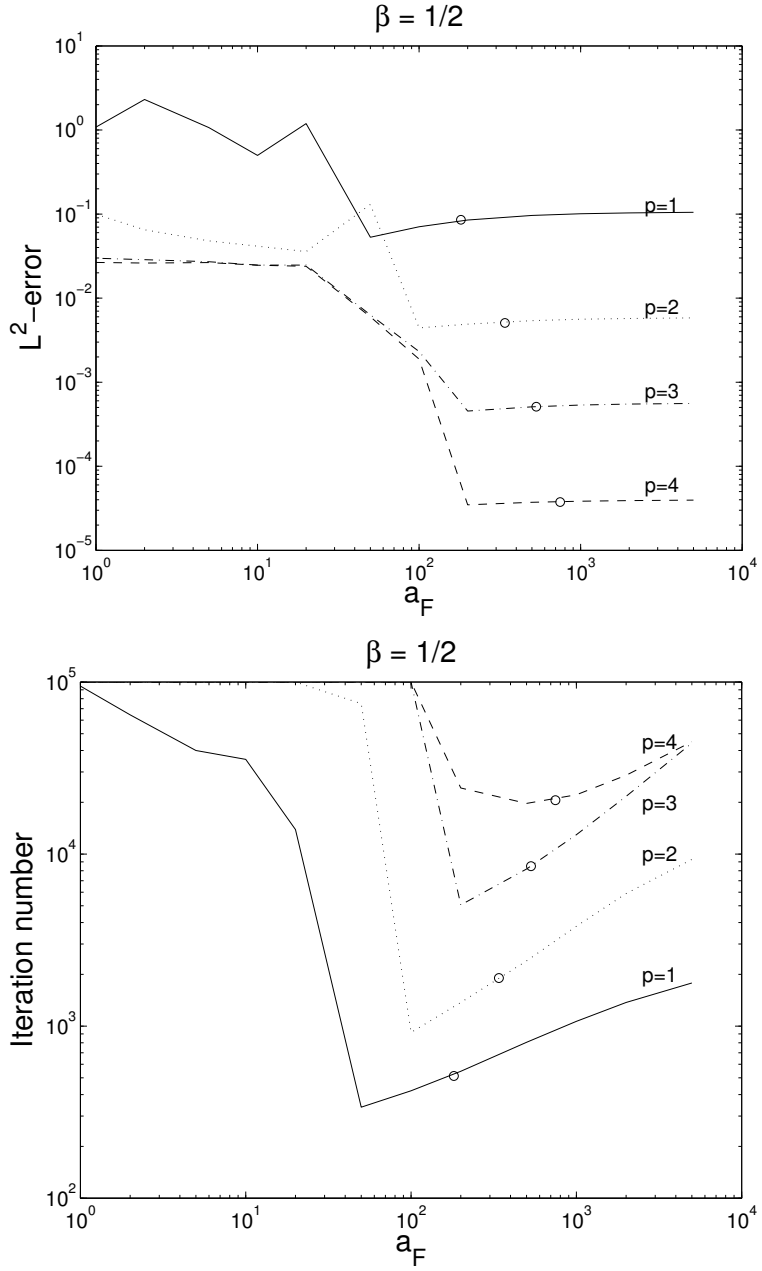


Figure 4.4: *L^2-error (left) and the number of* MINRES *iterations (right) as a function of the penalty parameter a_F in the IP-DG method. An unstructured mesh of 432 tetrahedra and coercivity constant $\beta = \frac{1}{2}$ are used.*

Table VII: *Eigenvalues (divided by π^2) obtained on uniform meshes with $p = 3$*

h	$h/2$	$h/4$	Expected
2.0482	2.0008	2.0000	2.0000
2.0482	2.0008	2.0000	2.0000
2.0482	2.0008	2.0000	2.0000
3.1833	3.0067	3.0001	3.0000
3.1833	3.0067	3.0001	3.0000
5.2151	5.0236	5.0005	5.0000
5.2151	5.0236	5.0005	5.0000
5.2151	5.0236	5.0005	5.0000
5.4621	5.0252	5.0005	5.0000
5.4621	5.0252	5.0005	5.0000
5.4621	5.0252	5.0005	5.0000
6.6602	6.0446	6.0010	6.0000
6.6602	6.0446	6.0010	6.0000
6.6602	6.0446	6.0010	6.0000
7.2220	6.0471	6.0011	6.0000
7.2220	6.0471	6.0011	6.0000
7.2220	6.0471	6.0011	6.0000
8.9659	8.1927	8.0032	8.0000
8.9659	8.1927	8.0032	8.0000
10.4743	8.1927	8.0032	8.0000

Table VIII: *Eigenvalues (divided by π^2) obtained on uniform meshes with $p = 4$*

h	$h/2$	$h/4$	Expected
2.0013	2.0000	2.0000	2.0000
2.0013	2.0000	2.0000	2.0000
2.0013	2.0000	2.0000	2.0000
3.0118	3.0000	3.0000	3.0000
3.0118	3.0000	3.0000	3.0000
5.1823	5.0013	5.0000	5.0000
5.1823	5.0013	5.0000	5.0000
5.1823	5.0013	5.0000	5.0000
5.1912	5.0014	5.0000	5.0000
5.1912	5.0014	5.0000	5.0000
5.1912	5.0014	5.0000	5.0000
6.3525	6.0031	6.0000	6.0000
6.3525	6.0031	6.0000	6.0000
6.3525	6.0031	6.0000	6.0000
6.4912	6.0037	6.0000	6.0000
6.4912	6.0037	6.0000	6.0000
6.4912	6.0037	6.0000	6.0000
8.2632	8.0050	8.0001	8.0000
8.8368	8.0050	8.0001	8.0000
8.8368	8.0050	8.0001	8.0000

CHAPTER 5

DG VS NÉDÉLEC IN THE FINITE ELEMENT DISCRETISATIONS OF THE SECOND-ORDER TIME-DOMAIN MAXWELL EQUATION

5.1 Introduction

High-order finite element methods (FEM) are an increasingly important technology in large-scale electromagnetic simulations thanks to their ability to effectively model complex geometrical structures and long-time wave propagation. It has long been known that the standard H^1-conforming FEM for electromagnetic waves may result in non-physical, spurious solutions. Instead, one may naturally opt for the $H(\text{curl})$-conforming FEM pioneered by Nédélec [58, 59] and Bossavit [8, 9]. It has the advantage of mimicking the geometrical properties of the Maxwell equations at the discrete level. However, in time-domain computations it requires solving linear systems with mass matrices even if an explicit time-integration method is used. One attractive alternative – also free of spurious solutions under certain conditions – is the discontinuous Galerkin FEM (DG-FEM) [22, 42, 44], where the resulting mass matrix is block-diagonal and therefore the computational cost of its inversion is negligible. But this additional flexibility comes at a cost. The number of degrees of freedom in DG discretisations is higher than that in the $H(\text{curl})$-conforming discretisation, although the difference decreases as the polynomial order in the spatial discretisation grows. As an illustration, Figure 5.1 shows the sparsity patterns of the mass matrices for both methods when a mesh of 320 tetrahedra and third-order polynomials are used.

So there appears to be a trade-off between the two methods in time-domain computations. In general, the $H(\text{curl})$-conforming approach is more efficient with low-order polynomials and DG-FEM with high-order ones. The expected break-even point depends on a number of factors, such as the conditioning and sparsity

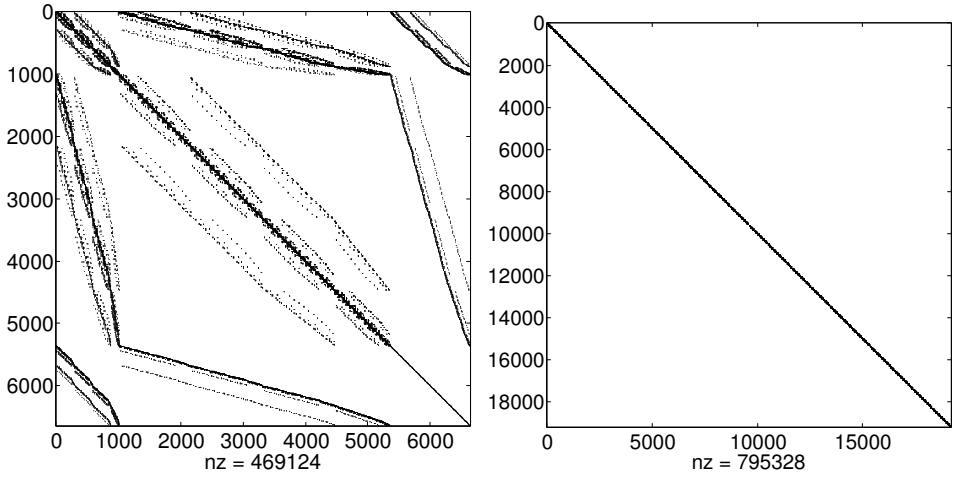


Figure 5.1: *Sparsity pattern of the mass matrix for $H(\mathrm{curl})$-conforming FEM (left) and DG-FEM (right) for a mesh with 320 elements. Third-order polynomials are used, which means that the size of the blocks in the right plot is 60×60. Note the difference in size between the two matrices.*

of the mass and stiffness matrices in the resulting semi-discrete systems. As a novelty, the focus of this chapter is to provide a comparison of the computational performance of the two methods when hierarchic $H(\mathrm{curl})$-conforming basis functions [2, 74] are used on tetrahedral meshes. The motivation behind this choice is that these basis functions play an ever more important role in the development of p- and hp-adaptive methods [23] for the Maxwell equation.

Throughout the chapter, the different discretisation techniques are applied to the three-dimensional Maxwell equations in the second-order time-dependent form,

$$\varepsilon_{\mathrm{r}} \frac{\partial^2 \mathbf{E}}{\partial t^2} + \sigma \frac{\partial \mathbf{E}}{\partial t} + \nabla \times \left(\mu_{\mathrm{r}}^{-1} \nabla \times \mathbf{E} \right) = -\frac{\partial \mathbf{J}}{\partial t}, \tag{5.1}$$

with homogeneous boundary conditions $\mathbf{n} \times \mathbf{E} = \mathbf{0}$. All quantities are dimensionless[1] in (5.1), where $\mathbf{E}$ is the electric field and $\mathbf{J}$ is the electric current density. The values σ, ε_{r} and μ_{r} are assumed to be time-independent constant scalars, and they respectively denote conductivity, relative dielectric permittivity and relative magnetic permeability. If the domain is filled with nonconductive material, the damping term $\sigma \frac{\partial \mathbf{E}}{\partial t}$ is absent. If, in addition, the source term $-\partial_t \mathbf{J}$ is also taken to be zero, we have the conservative Maxwell wave equation.

[1] We can derive the dimensionless form by using the scalings $\mathbf{x} = \tilde{\mathbf{x}}/\tilde{L}$, $t = \tilde{t}/(\tilde{L}/\tilde{c}_0)$, $\mathbf{E} = \tilde{\mathbf{E}}/(\tilde{Z}_0 \tilde{H}_0)$, $\boldsymbol{H} = \tilde{\boldsymbol{H}}/\tilde{H}_0$, $\mathbf{J} = \tilde{\mathbf{J}}/(\tilde{H}_0/\tilde{L})$ and $\sigma = \tilde{\mathbf{J}}\tilde{L}\tilde{Z}_0/\tilde{\mathbf{E}}$, with tilde denoting the dimensional quantities. Here $\tilde{L}$ is the reference length, $\tilde{c}_0 = (\tilde{\mu}_0 \tilde{\varepsilon}_0)^{-1/2}$ is the speed of light in vacuum, $\boldsymbol{H}$ is the magnetic field (eliminated in (5.1)), $\tilde{H}_0$ is the reference magnetic field strength and $\tilde{Z}_0 = (\mathrm{i}\tilde{\omega}\tilde{\mu}_0/(\tilde{\sigma} + \mathrm{i}\tilde{\omega}\tilde{\varepsilon}_0))^{1/2}$ is the intrinsic impedance, with $\tilde{\omega}$ being the angular frequency and i the imaginary unit.

Following the method of lines, we first discretise the spatial operators, using the $H(\mathrm{curl})$-conforming FEM or the DG-FEM. In either case, we arrive at a semi-discrete system in the form of second-order ordinary differential equations (ODEs) in $\mathbb{R}^n$,

$$M_\varepsilon u'' + M_\sigma u' + S_\mu u = j, \tag{5.2}$$

where u is the unknown vector of N scalar coefficients associated with the approximation of the electric field $\mathbf{E}$. The source term j is the projection of $-\partial_t \mathbf{J}$ onto the finite-element space and in general may also contain boundary data. For simplicity, however, we restrict ourselves to the homogeneous Dirichlet boundary condition, $\mathbf{n} \times \mathbf{E} = \mathbf{0}$, in this chapter. Each term in the left-hand side of (5.2) corresponds to the respective term in the left-hand side of (5.1). The mass matrix M_ε is symmetric positive definite and the conductivity matrix M_σ is symmetric positive semi-definite. In addition, for constant scalars σ and ε_r the matrices M_ε and M_σ are identical up to a constant. The stiffness matrix S_μ is the discretisation of the wave term and is symmetric positive semi-definite.

Convergence results for the $H(\mathrm{curl})$-conforming semi-discrete approximation (5.2) are relatively well-established [45, 56]. Results on the semi-discrete DG discretisation are more recent: energy-norm estimates [33] and L^2-estimates [34] have been derived for the Maxwell equations; optimal error estimates for the fully-discrete second-order scalar wave equation have been provided in [35]; and a promising energy-conserving local-time stepping scheme has been developed in [24].

A vital feature of (5.1) and (5.2) from the point of view of time integration is that it includes the conductivity σ. Even moderate values of σ may result in a prohibitively small time step for many of the popular time-integration schemes. Therefore, we pay special attention to time-integration methods that treat the conductivity mass matrix M_σ in an implicit way. Many of such methods and others discussed in this chapter have been previously studied in [11] for the system of first-order Maxwell equations discretised by the lowest-order $H(\mathrm{curl})$-conforming elements. See also [64] for more details on composition methods for the conduction-free Maxwell equations.

The semi-discrete system (5.2) conserves (discrete) energy for the spatial discretisations discussed here, since these are both symmetric. Hence, using an energy-conservative time-integration method results in a conservative fully-discrete scheme. We investigate the dispersion and dissipation error of the schemes in two steps. First, we determine the dispersion error of the semi-discrete scheme by solving the time-harmonic eigenvalue problem corresponding to the semi-discrete system. Second, we can then apply any given time-integration scheme to a simple, but equivalent, model problem that includes the information of the semi-discrete numerical frequency, and thus define the dispersion (and, if there is any, dissipation) error of the time-integration method. This approach shows if the dispersion error is dominated by the spatial or temporal discretisation – a piece of information that may prove useful in deciding whether or not to go for high-order time-integration schemes.

The computational performance of the $H(\mathrm{curl})$-conforming method hinges to a

great degree on efficiently solving the linear system with the mass matrix. A number of advanced techniques have been proposed recently, including mass lumping [7, 27, 85], the explicit computation of an approximate sparse inverse mass matrix [38], or the construction of special preconditioners. These approaches, however, do not in their current states provide a general framework and therefore cannot be extended to high-order discretisations in a straightforward manner. That is the reason why in this chapter we resort to standard preconditioners. It is of course also possible to use sparse direct solvers but in test problems we found that they are too memory demanding for large-scale three-dimensional computations.

The remaining part of the chapter is organised as follows. The weak formulations of the $H(\mathrm{curl})$-conforming FEM and the DG-FEM are given in Section 5.2. The semi-discrete system arising from either of the spatial discretisations is analysed in Section 5.3, while we briefly describe a number of the most widely-used time-integration methods in Section 5.4. Numerical examples that compare the computational performance of the two finite element approaches are presented in Section 5.5, where we test both low-order and high-order approximations. Section 5.6 concludes the chapter with final remarks.

5.2 The weak formulation

Before we present the weak formulations that result from the $H(\mathrm{curl})$-conforming and the DG discretisations, we introduce the tessellation $\mathcal{T}_h$ that partitions the polyhedral domain $\Omega \subset \mathbb{R}^3$ into a set of tetrahedra $\{K\}$. Throughout the chapter we assume that the mesh is shape-regular and that each tetrahedron is straight-sided. The notations $\mathcal{F}_h$, $\mathcal{F}_h^i$ and $\mathcal{F}_h^b$ stand respectively for the set of all faces $\{F\}$, the set of all internal faces, and the set of all boundary faces.

On the computational domain Ω, we define the spaces

$$H(\mathrm{curl};\Omega) := \left\{ \mathbf{u} \in \left[L^2(\Omega)\right]^3 : \nabla \times \mathbf{u} \in \left[L^2(\Omega)\right]^3 \right\},$$
$$H_0(\mathrm{curl};\Omega) := \left\{ \mathbf{u} \in H(\mathrm{curl};\Omega) \;\middle|\; \mathbf{n} \times \mathbf{u} = \mathbf{0} \text{ on } \partial\Omega \right\},$$

and the L^2 inner product $(\cdot,\cdot)$

$$(\mathbf{u},\mathbf{v}) = \int_\Omega \mathbf{u}\cdot\mathbf{v}\,\mathrm{d}V.$$

The continuous weak formulation of (5.1) now reads as follows: Find $\mathbf{E} \in H_0(\mathrm{curl},\Omega)$ such that $\forall \mathbf{w} \in H_0(\mathrm{curl},\Omega)$ the relation

$$\frac{\partial^2}{\partial t^2}\left(\varepsilon_\mathrm{r}\mathbf{E},\mathbf{w}\right) + \frac{\partial}{\partial t}\left(\sigma\mathbf{E},\mathbf{w}\right) + \left(\mu_\mathrm{r}^{-1}\nabla\times\mathbf{E},\nabla\times\mathbf{w}\right) = -\left(\frac{\partial\mathbf{J}}{\partial t},\mathbf{w}\right) \tag{5.3}$$

is satisfied. See e.g. [51, 56].

5.2.1 Weak formulation of the $H(\mathrm{curl})$-conforming discretisation

In order to discretise (5.3), we first introduce the finite element space associated with the tessellation $\mathcal{T}_h$. Let $\mathcal{P}_p(K)$ be the space of polynomials of degree at most $p \geq 1$ on $K \in \mathcal{T}_h$. Over each element K the $H(\mathrm{curl})$-conforming polynomial space is defined as

$$Q^p = \left\{ \boldsymbol{u} \in [\mathcal{P}_p(K)]^3 ;\ \boldsymbol{u}_T|_{F_i^K} \in \left[\mathcal{P}_p(F_i^K)\right]^2 ;\ \boldsymbol{u} \cdot \boldsymbol{\tau}_j|_{e_j^K} \in \mathcal{P}_p(e_j^K) \right\}, \tag{5.4}$$

where F_i^K, $i = 1, 2, 3, 4$ are the faces of the element; e_j^K, $j = 1, 2, 3, 4, 5, 6$ are the edges of the element; $\mathbf{u}_T$ is the tangential component of $\mathbf{u}$; and $\boldsymbol{\tau}_j$ is the directed tangential vector on edge e_j^K. For the construction of Q^p, we use a set of $H(\mathrm{curl})$-conforming hierarchic basis functions [2, 74].

Next, we introduce the discrete space of globally $H(\mathrm{curl})$-conforming functions

$$\Upsilon_h^p := \left\{ \boldsymbol{v} \in [H_0(\mathrm{curl}, \Omega)]^3 \;\middle|\; \boldsymbol{v}|_K \in Q^p,\ \forall K \in \mathcal{T}_h \right\},$$

and let the set of basis functions $\{\boldsymbol{\psi}_i\}$ span the space Υ_h^p. See [56] for a detailed discussion on both continuous and discrete $H(\mathrm{curl})$-conforming spaces. We can then approximate the electric field $\mathbf{E}$ as

$$\mathbf{E} \approx \mathbf{E}_h = \sum_i u_i(t) \boldsymbol{\psi}_i(x), \tag{5.5}$$

from which the discrete weak formulation reads as follows: Find $\mathbf{E}_h \in \Upsilon_h^p$ such that $\forall \boldsymbol{\phi} \in \Upsilon_h^p$ the relation

$$\frac{\partial^2}{\partial t^2} \left(\varepsilon_\mathrm{r} \mathbf{E}_h, \boldsymbol{\phi}\right) + \frac{\partial}{\partial t} \left(\sigma \mathbf{E}_h, \boldsymbol{\phi}\right) + \left(\mu_\mathrm{r}^{-1} \nabla \times \mathbf{E}_h, \nabla \times \boldsymbol{\phi}\right) = -\left(\frac{\partial \mathbf{J}}{\partial t}, \boldsymbol{\phi}\right) \tag{5.6}$$

is satisfied. Note that (5.6) is satisfied if and only if it is satisfied for every basis function $\boldsymbol{\psi}_i, i = 1, \ldots, N$, with N being the global number of degrees of freedom. As a result, substitution of (5.5) into (5.6) yields the semi-discrete system (5.2) with

$$[M_\varepsilon]_{ij} = \left(\varepsilon_\mathrm{r} \boldsymbol{\psi}_i, \boldsymbol{\psi}_j\right), \quad [S_\mu]_{ij} = \left(\mu_\mathrm{r}^{-1} \nabla \times \boldsymbol{\psi}_i, \nabla \times \boldsymbol{\psi}_j\right),$$
$$[M_\sigma]_{ij} = \left(\sigma \boldsymbol{\psi}_i, \boldsymbol{\psi}_j\right), \quad [j]_i = -\left(\frac{\partial \mathbf{J}}{\partial t}, \boldsymbol{\psi}_i\right).$$

Each of the above matrices – M_ε, M_σ and S_μ – has a large number entries far off the diagonal, increasing the computational cost for both explicit and implicit time-integration methods.

5.2.2 Weak formulation of DG-FEM

In contrast to the $H(\mathrm{curl})$-conforming discretisation, in DG-FEM we are looking for the discrete solution in the space

$$\Sigma_h^p := \left\{ \boldsymbol{\sigma} \in \left[L^2(\Omega)\right]^3 \;\middle|\; \boldsymbol{\sigma}|_K \in Q^p,\ \forall K \in \mathcal{T}_h \right\}.$$

That is, we allow the polynomial functions to be fully discontinuous across element interfaces and assume that the set of basis functions $\{\boldsymbol{\psi}_i\}$ now span the space Σ_h^p. Instead of enforcing continuity of the tangential components, the information between elements is now coupled through the numerical flux [22, 4, 44]. Before we can define the numerical flux and formulate the discretisation for DG-FEM, we first need to introduce more notation.

Consider an interface $F \in \mathcal{F}_h$ between element K^L and element K^R, and let $\boldsymbol{n}^L$ and $\boldsymbol{n}^R$ represent their respective outward pointing normal vectors. We define the tangential jump and the average of the quantity $\boldsymbol{u}$ across interface F as

$$[\![\boldsymbol{u}]\!]_T = \boldsymbol{n}^L \times \boldsymbol{u}^L + \boldsymbol{n}^R \times \boldsymbol{u}^R \quad \text{and} \quad \{\!\{\boldsymbol{u}\}\!\} = \left(\boldsymbol{u}^L + \boldsymbol{u}^R\right)/2,$$

respectively. Here $\boldsymbol{u}^L$ and $\boldsymbol{u}^R$ are the values of the trace of $\boldsymbol{u}$ at ∂K^L and ∂K^R, respectively. At the boundary Γ, we set $\{\!\{\boldsymbol{u}\}\!\} = \boldsymbol{u}$ and $[\![\boldsymbol{u}]\!]_T = \boldsymbol{n} \times \boldsymbol{u}$. We furthermore introduce the global lifting operator $\mathcal{R}(\boldsymbol{u}) : \left[L^2(\mathcal{F}_h)\right]^3 \to \Sigma_h^p$ as

$$(\mathcal{R}(\boldsymbol{u}), \boldsymbol{v})_\Omega = \int_{\mathcal{F}_h} \boldsymbol{u} \cdot \{\!\{\boldsymbol{v}\}\!\} \, \mathrm{d}A, \quad \forall \boldsymbol{v} \in \Sigma_h^p, \tag{5.7}$$

and, for a given face $F \in \mathcal{F}_h$, the local lifting operator $\mathcal{R}_F(\boldsymbol{u}) : \left[L^2(F)\right]^3 \to \Sigma_h^p$ as

$$(\mathcal{R}_F(\boldsymbol{u}), \boldsymbol{v})_\Omega = \int_F \boldsymbol{u} \cdot \{\!\{\boldsymbol{v}\}\!\} \, \mathrm{d}A, \quad \forall \boldsymbol{v} \in \Sigma_h^p. \tag{5.8}$$

Note that $\mathcal{R}_F(\boldsymbol{u})$ vanishes outside the elements connected to the face F so that for a given element $K \in \mathcal{T}_h$ we have the relation

$$\mathcal{R}(\boldsymbol{u}) = \sum_{F \in \mathcal{F}_h} \mathcal{R}_F(\boldsymbol{u}), \quad \forall \boldsymbol{u} \in \left[L^2(\mathcal{F}_h)\right]^3. \tag{5.9}$$

The discrete weak formulation for DG-FEM now reads as follows [68]: Find $\mathbf{E}_h \in \Sigma_h^p$ such that $\forall \boldsymbol{\phi} \in \Sigma_h^p$ the relation

$$\begin{aligned} &\frac{\partial^2}{\partial t^2}\left(\varepsilon_\mathrm{r}\mathbf{E}_h, \boldsymbol{\phi}\right) + \frac{\partial}{\partial t}\left(\sigma\mathbf{E}_h, \boldsymbol{\phi}\right) + \left(\mu_\mathrm{r}^{-1}\nabla_h \times \mathbf{E}_h, \nabla_h \times \boldsymbol{\phi}\right) \\ &\quad - \int_{\mathcal{F}_h} [\![\boldsymbol{E}_h]\!]_T \cdot \{\!\{\nabla_h \times \boldsymbol{\phi}\}\!\} \, \mathrm{d}A - \int_{\mathcal{F}_h} \{\!\{\nabla_h \times \boldsymbol{E}_h\}\!\} \cdot [\![\boldsymbol{\phi}]\!]_T \, \mathrm{d}A \\ &\quad + \sum_{F \in \mathcal{F}_h} C_F \left(\mathcal{R}_F([\![\boldsymbol{E}]\!]_T), \mathcal{R}_F([\![\boldsymbol{\phi}]\!]_T)\right)_\Omega = -\left(\frac{\partial \mathbf{J}}{\partial t}, \boldsymbol{\phi}\right) \end{aligned} \tag{5.10}$$

is satisfied, where the operator ∇_h denotes the elementwise application of ∇. For stability, the constant C_F has to satisfy the condition [68]

$$C_F \geq n_f C_1 + \min\left\{\frac{1}{2}, \frac{1}{C_2}\right\},$$

where C_1 and C_2 are positive constants and n_f denotes the number of sides of each element, that is for tetrahedra $n_f = 4$. As a consequence, the constant C_F is independent of both the polynomial order and the mesh size.

Again, (5.10) is satisfied if and only if it is satisfied for every basis function $\boldsymbol{\psi}_i, i = 1, \ldots, N$, with N being the global number of degrees of freedom. Substitution of $\mathbf{E} \approx \mathbf{E}_h = \sum_i u_i(t)\boldsymbol{\psi}_i(x)$ into (5.10) yields the semi-discrete system (5.2) with

$$
\begin{aligned}
[M_\varepsilon]_{ij} &= \left(\varepsilon_\mathrm{r}\boldsymbol{\psi}_i, \boldsymbol{\psi}_j\right), \quad [M_\sigma]_{ij} = \left(\sigma\boldsymbol{\psi}_i, \boldsymbol{\psi}_j\right), \quad [j]_i = -\left(\frac{\partial \mathbf{J}}{\partial t}, \boldsymbol{\psi}_i\right), \\
[S_\mu]_{ij} &= \left(\mu_\mathrm{r}^{-1}\nabla_h \times \boldsymbol{\psi}_i, \nabla_h \times \boldsymbol{\psi}_j\right) - \int_{\mathcal{F}_h} [\![\boldsymbol{\psi}_i]\!]_T \cdot \{\!\{\nabla_h \times \boldsymbol{\psi}_j\}\!\} \,\mathrm{d}A \\
&\quad - \int_{\mathcal{F}_h} \{\!\{\nabla_h \times \boldsymbol{\psi}_i\}\!\} \cdot [\![[\![\boldsymbol{\psi}_j]\!]_T]\!] \,\mathrm{d}A + \sum_{F \in \mathcal{F}_h} C_F \left(\mathcal{R}_F([\![\boldsymbol{\psi}_i]\!]_T), \mathcal{R}_F([\![[\![\boldsymbol{\psi}_j]\!]_T]\!])\right)_\Omega .
\end{aligned}
$$

The matrices M_ε and M_σ are now block-diagonal with the elementwise matrices being the blocks. However, the stiffness matrix S_μ has still many entries far off the diagonal because of the face integrals in its construction. That is why, DG in general warrants the use of explicit time-integration schemes but not implicit ones.

We emphasise that (5.10) is only one of many possible formulations of DG-FEM, depending on the numerical flux one chooses to use. The one we have introduced here is based on the numerical flux from [12] (see also [6]), and was analysed in detail for the time-harmonic Maxwell equations in [68]. See also [4] for an overview of DG-FEM methods for elliptic problems and for a large number of possible choices for the numerical flux.

5.2.3 The energy norm

Convergence results for FEMs are generally derived not only in the L^2-norm but also in a norm associated with the discrete energy of the approximation [33, 35]. These are defined for the $H(\mathrm{curl})$-conforming and DG discretisations as

$$
\|\mathbf{v}\|^2_{H(\mathrm{curl})} = \|\mathbf{v}\|^2 + \|\nabla \times \mathbf{v}\|^2
$$

and

$$
\|\mathbf{v}\|^2_{\mathrm{DG}} = \|\mathbf{v}\|^2 + \|\nabla_h \times \mathbf{v}\|^2 + \|\mathtt{h}^{-\frac{1}{2}} [\![\mathbf{v}]\!]_T \|^2_{\mathcal{F}_h},
$$

respectively. In the above definition, $\|\cdot\|_{\mathcal{F}_h}$ denotes the $L^2(\mathcal{F})$ norm and $\mathtt{h}(\boldsymbol{x}) = h_F$ is the diameter of face F containing $\boldsymbol{x}$. We note that the two definitions of the energy norm are actually identical as ∇_h becomes ∇ and $[\![\cdot]\!]_T$ vanishes if $H(\mathrm{curl})$-conforming discretisation is used.

5.3 Stability of the semi-discrete system

To carry out a basic stability analysis, we first transform (5.2) into a first-order system of ODEs,

$$\begin{aligned} u' &= v, \\ M_\varepsilon v' + M_\sigma v + S_\mu u &= j. \end{aligned} \tag{5.11}$$

Recall that S_μ is symmetric and therefore – using the inner-product notation for discrete vectors – we have the property

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\left(v^T M_\varepsilon v + u^T S_\mu u\right) = \frac{\mathrm{d}v^T}{\mathrm{d}t} M_\varepsilon v + v^T M_\varepsilon \frac{\mathrm{d}v}{\mathrm{d}t} + \frac{\mathrm{d}u^T}{\mathrm{d}t} S_\mu u + u^T S_\mu \frac{\mathrm{d}u}{\mathrm{d}t} = \\ 2v^T\left(-M_\sigma v - S_\mu u + j\right) + 2v^T S_\mu u = 2v^T j - 2v^T M_\sigma v. \end{aligned} \tag{5.12}$$

If $j = 0$, this entails stability, that is

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(v^T M_\varepsilon v + u^T S_\mu u\right) = -2v^T M_\sigma v \leq 0,$$

since, for constant σ, the matrix M_σ is positive definite if $\sigma > 0$ and $M_\sigma = 0$ if $\sigma = 0$. Therefore, if $\sigma = 0$ in addition to $j = 0$, (5.12) shows conservation.

In order to use a stability test model introduced later in this section, we transform (5.11) to an equivalent explicit form. To do so, we multiply the first equation in (5.11) with M_ε and introduce the Cholesky factorisation $LL^T = M_\varepsilon$. The new variables $\tilde{v} = L^T v$ and $\tilde{u} = L^T u$ then satisfy the system

$$\begin{pmatrix} \tilde{u}' \\ \tilde{v}' \end{pmatrix} = \begin{pmatrix} 0 & I \\ -\tilde{S}_\mu & -\tilde{M}_\sigma \end{pmatrix} \begin{pmatrix} \tilde{u} \\ \tilde{v} \end{pmatrix} + \begin{pmatrix} 0 \\ \tilde{j} \end{pmatrix}, \tag{5.13}$$

where

$$\tilde{j} = L^{-1} j, \quad \tilde{S}_\mu = L^{-1} S_\mu L^{-T}, \quad \tilde{M}_\sigma = L^{-1} M_\sigma L^{-T}.$$

Since both the conductivity coefficient σ and the permittivity coefficient ε_r are constant scalars in (5.1), the matrix $\tilde{M}_\sigma$ in (5.13) is the constant diagonal matrix

$$\tilde{M}_\sigma = \gamma I, \quad \gamma = \frac{\sigma}{\varepsilon_\mathrm{r}}.$$

From this we can derive a two-by-two system through which stability of time-integration methods for (5.11) can be examined.

The matrix $\tilde{S}_\mu$ is symmetric positive semi-definite so it can be decomposed as $\tilde{S}_\mu = U\Lambda U^T$, where Λ is a diagonal matrix with the eigenvalues of $\tilde{S}_\mu$ on its diagonal

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r \geq \lambda_{r+1} = \lambda_{r+1} = \ldots = \lambda_n = 0,$$

where r is the rank of the matrix. The matrix U is orthogonal and its columns are the eigenvectors of $\tilde{S}_\mu$. Using a permutation matrix $\mathcal{P}$, we have

$$\mathcal{A} = \begin{pmatrix} 0 & I \\ -\tilde{S}_\mu & -\tilde{M}_\sigma \end{pmatrix} = \begin{pmatrix} 0 & UU^T \\ -U\Lambda U^T & -\gamma I \end{pmatrix} = \begin{pmatrix} U & 0 \\ 0 & U \end{pmatrix} \begin{pmatrix} 0 & I \\ -\Lambda & -\gamma I \end{pmatrix} \begin{pmatrix} U^T & 0 \\ 0 & U^T \end{pmatrix} = \begin{pmatrix} U & 0 \\ 0 & U \end{pmatrix} \mathcal{P}\Lambda_{\mathcal{P}}\mathcal{P}^T \begin{pmatrix} U^T & 0 \\ 0 & U^T \end{pmatrix}, \quad (5.14)$$

where $\Lambda_{\mathcal{P}}$ is a block-diagonal matrix with two-by-two blocks

$$\begin{pmatrix} 0 & 1 \\ -\lambda_k & -\gamma \end{pmatrix}, \qquad k = 1, \ldots, N. \quad (5.15)$$

This allows us to state the following proposition.

Proposition 1. *Assume that σ and ε_{r} are scalar and $\gamma = \sigma/\varepsilon_{\mathrm{r}}$. Then the matrix $\mathcal{A}$ has*

(i) $n - r$ *zero eigenvalues,*

(ii) $n - r$ *eigenvalues which equal* $-\gamma$*,*

(iii) $2r$ *eigenvalues which are*

$$\frac{-\gamma \pm \sqrt{\gamma^2 - 4\lambda_k}}{2}, \quad k = 1, \ldots, r.$$

Thus, the orthogonal transformation $V \equiv \left(\begin{smallmatrix} U & 0 \\ 0 & U \end{smallmatrix}\right)\mathcal{P}$ decouples (5.13) into r two-by-two systems

$$\begin{pmatrix} \hat{u}' \\ \hat{v}' \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -\lambda & -\gamma \end{pmatrix} \begin{pmatrix} \hat{u} \\ \hat{v} \end{pmatrix} + \begin{pmatrix} 0 \\ \hat{j} \end{pmatrix},$$

with $\lambda = \lambda_k > 0$, $k = 1, \ldots, r$, and $n - r$ two-by-two systems

$$\begin{pmatrix} \hat{u}' \\ \hat{v}' \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & -\gamma \end{pmatrix} \begin{pmatrix} \hat{u} \\ \hat{v} \end{pmatrix} + \begin{pmatrix} 0 \\ \hat{j} \end{pmatrix}.$$

For the stability analysis, we may neglect the source term and thus arrive at the two-by-two stability test model

$$\begin{pmatrix} \hat{u}' \\ \hat{v}' \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -\lambda & -\gamma \end{pmatrix} \begin{pmatrix} \hat{u} \\ \hat{v} \end{pmatrix}, \qquad \lambda \geq 0, \quad \gamma \geq 0. \quad (5.16)$$

The attractive feature of this formulation is that stability for the test model (5.16) induces stability for (5.11) in the norm generated by the inner product in (5.12).

Useful though equation (5.16) is, it is important to emphasise that the derivation of (5.16) requires constant scalars in the coefficients ε_{r} and σ, thus limiting the generality of this approach.

5.4 Time-integration methods

Probably the most popular time-integration methods to use in combination with high-order DG methods are high-order Runge-Kutta methods, giving rise to what are collectively called the Runge-Kutta DG (RKDG) methods [22]. For continuous and $H(\mathrm{curl})$-conforming FEMs geometric integrators are also widely used thanks to their ability to conserve symplecticity[2] at the discrete level [64]. In this section, we briefly recall the construction of these two families of methods and we also discuss local and global Richardson extrapolation.

The highest-order polynomial we use within the finite element methods is $p = 3$. For both the DG and the $H(\mathrm{curl})$-conforming methods, this corresponds to fourth-order convergence for the semi-discrete system (5.2) provided that the solution is smooth [33, 34, 45, 56]. Therefore, we now only discuss time-integration methods that are also at most fourth-order accurate. Extension to higher order, however, is usually straightforward.

For investigating the properties of any given time-integration method, let τ denote the time-step size and introduce $z_\lambda = \tau\sqrt{\lambda}$ and $z_\gamma = \tau\gamma$.[3] The stability of the time-integration method can then, in general, be best inspected through the (numerically determined) stability region

$$\mathcal{S} = \{(z_\lambda, z_\gamma) : z_\lambda, z_\gamma \geq 0 \ \text{ with } \ |\mu| < 1, \ \mu \text{ eigenvalues of the amplification operator}\}$$

associated with the test model (5.16).

5.4.1 Runge-Kutta methods

Out of the many different types of Runge-Kutta methods, strong-stability-preserving Runge-Kutta methods (SSPRK) [22] are particularly well suited for the time integration of semi-discrete hyperbolic problems.

With the definition of the initial values $U_0 = u_n$ and $V_0 = v_n$ for the time step from t_n to t_{n+1}, the general s-stage SSPRK scheme for (5.11) reads

$$\begin{aligned} U_k &= \sum_{l=0}^{k-1} \left(\alpha_{kl} U_l + \tau\beta_{kl} V_l\right), \\ M_\varepsilon V_k &= \sum_{l=0}^{k-1} \left(\alpha_{kl} V_l + \tau\beta_{kl}\left(-S_\mu U_l - M_\sigma V_l + j(t_l)\right)\right), \\ u_{n+1} &= U_s, \\ v_{n+1} &= V_s, \end{aligned} \tag{5.17}$$

[2] The preservation of symplecticity is important because it is related to energy. More precisely, for symplectic integrators the error in total energy will remain within a certain margin throughout the entire time integration.

[3] These values appear in a natural way in the amplification matrices of most time-integration methods described later in this section

where $k = 1, \ldots, s$ while α_{kl} and β_{kl} are the coefficients in the SSPRK method. Applying (5.17) to the test equation (5.16), the amplification operator $\mathcal{M}^s_{\text{ssp}}$ of an s-stage SSPRK method is

$$\mathcal{M}^k_{\text{ssp}} = \sum_{l=1}^{k-1} \mathcal{B}_{kl} \mathcal{M}^{l-1}_{\text{ssp}} \qquad \text{with} \qquad \mathcal{B}_{kl} = \alpha_{kl} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \beta_{kl} \begin{pmatrix} 0 & 1 \\ z_\lambda^2 & -z_\gamma \end{pmatrix}, \quad (5.18)$$

where, again, $k = 1, \ldots, s$ and $\mathcal{M}^0_{\text{ssp}}$ is the identity. We show the stability regions of several SSPRK schemes in Figure 5.2, where we refer to an s-stage pth-order SSPRK method as SSPRK(s,p). It is important to emphasise that any (standard as well as 'nonstandard' such as SSP) explicit s-stage pth-order RK methods with $s = p$ has the same amplification matrix. In those cases the choice of coefficients only determines the SSP property and not the shape of the stability region. For the cases when $s = p+1$, we display the stability regions of the methods that were derived and analysed in [75, 66, 29]. The plots suggest that increasing the number of stages, while keeping the polynomial order fixed, results in a more favourable time-step restriction for the conduction part – one that more than offsets the cost of introducing an additional stage at each time step. This is in line with known results for the linear advection equation [75, 66]. Nevertheless, explicit SSPRK methods treat the conduction term, as well as the wave term, explicitly, which entails a time-step condition that is too restrictive even for moderately conductive materials (see next section). Note that the second-order methods are only stable for $z_\gamma > 0$, i.e. for $\sigma > 0$.

5.4.2 Composition methods

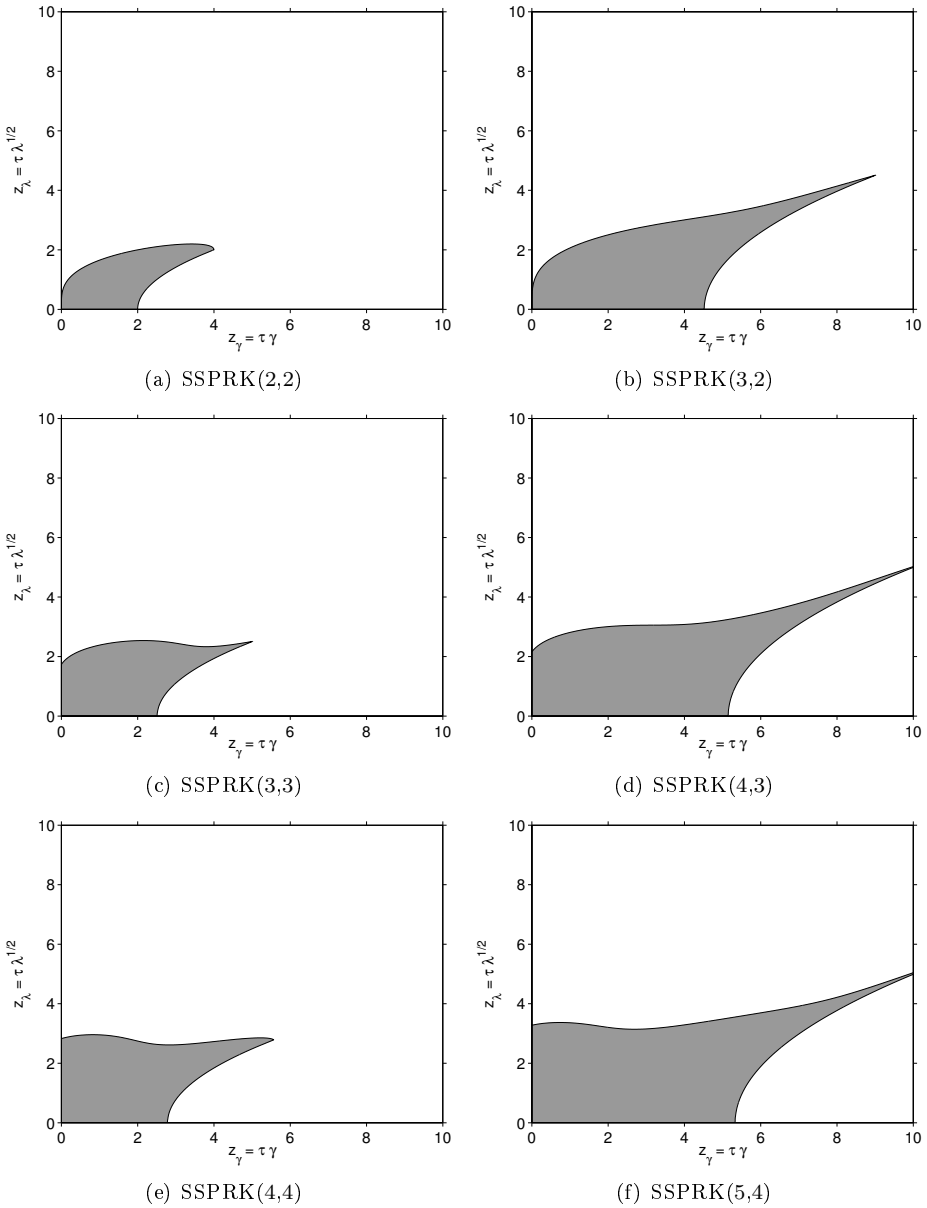
Composition methods [37, 54, 67] are especially suitable for geometric integration [64] and thus for the time integration of first-order Hamiltonian systems. Our description of the composition methods here strictly follows that in [11] and we refer to that work for more details.

The second-order composition method for (5.11) is defined as

$$\begin{aligned} \frac{u_{n+1/2} - u_n}{\tau} &= \frac{1}{2} v_n, \\ M_\varepsilon \frac{v_{n+1} - v_n}{\tau} &= -S_\mu u_{n+1/2} - \frac{1}{2} M_\sigma (v_n + v_{n+1}) + \frac{1}{2} \left(j(t_n) + j(t_{n+1}) \right), \\ \frac{u_{n+1} - u_{n+1/2}}{\tau} &= \frac{1}{2} v_{n+1}, \end{aligned} \quad (5.19)$$

which is akin to the ubiquitous leapfrog scheme, with the only difference being in the treatment of the source term (cf. [65]). If applied to the test model (5.16), it has the amplification matrix

$$\mathcal{M}_{\text{co2}} = \begin{pmatrix} 1 - \frac{1}{2} z_\lambda^2 + \frac{1}{2} z_\gamma & 1 - \frac{1}{4} z_\lambda^2 \\ -z_\lambda^2 & 1 - \frac{1}{2} z_\lambda^2 - \frac{1}{2} z_\gamma \end{pmatrix}, \quad (5.20)$$



(a) SSPRK(2,2)

(b) SSPRK(3,2)

(c) SSPRK(3,3)

(d) SSPRK(4,3)

(e) SSPRK(4,4)

(f) SSPRK(5,4)

Figure 5.2: *Stability regions (shaded areas) for several explicit SSPRK(s,p) methods, where s is the number of stages and p is the order of the method. Note that all explicit RK methods with $s = p$ have the same stability regions as SSPRK(s,p) with $s = p$ (left column).*

which entails the stability properties: $z_\lambda \leq 2$ if $z_\gamma = 0$ and $z_\lambda < 2$ if $z_\gamma > 0$. An attractive feature of this method over explicit RK methods is that it is unconditionally stable with respect to the conduction term.

In principle, it is possible to construct an arbitrary high-order composition method [37]. In this chapter, however, we are only interested in at most fourth-order accurate methods so we will now only discuss the fourth-order composition method. We define the initial values for the inner time step as $U_0 = u_n$ and $V_0 = v_n$, time levels t^u, t^v for u, v and coefficients

$$\begin{aligned} &\beta_0 = \alpha_0 = 0, \quad \beta_1 = \alpha_5 = \tfrac{14-\sqrt{19}}{108}, \quad \beta_2 = \alpha_4 = \tfrac{-23-20\sqrt{19}}{270}, \\ &\beta_3 = \alpha_3 = \tfrac{1}{5}, \quad \beta_4 = \alpha_2 = \tfrac{-2+10\sqrt{19}}{135}, \quad \beta_5 = \alpha_1 = \tfrac{146+5\sqrt{19}}{540}. \end{aligned}$$

The fourth-order composition method [37, 54, 11] for (5.11) now reads

$$\begin{aligned} \frac{U_k - U_{k-1}}{\tau} &= (\beta_k + \alpha_{k-1})\, V_{k-1}, \\ M_\varepsilon \frac{V_k - V_{k-1}}{\tau} &= \beta_k \left(-S_\mu U_k - M_\sigma V_{k-1} + j(t^v_{k-1})\right) + \alpha_k \left(-S_\mu U_k - M_\sigma V_k + j(t^v_k)\right), \\ v_{n+1} &= V_s, \\ u_{n+1} &= U_s + \alpha_s \tau V_s, \end{aligned} \tag{5.21}$$

where $k = 1, \ldots, s$, $s = 5$ is the number of internal time levels, and $t^v_k = t_n + (\tilde{\alpha}_k + \tilde{\beta}_k)\tau$ and $t^u_k = t_n + (\tilde{\alpha}_{k-1} + \tilde{\beta}_k)\tau$ with the coefficients $\tilde{\alpha}_k = \alpha_1 + \cdots + \alpha_k$ and $\tilde{\beta}_k = \beta_1 + \cdots + \beta_k$.

The amplification operator of (5.21) when applied to (5.16) is then

$$\prod_{k=5}^{1} \frac{1}{1 + \alpha_k z_\gamma} \begin{pmatrix} 1 + \alpha_k z_\gamma & (1 + \alpha_k z_\gamma)(\alpha_{k-1} + \beta_k) \\ -(\alpha_k + \beta_k) z_\lambda^2 & 1 - \beta_k z_\gamma - (\beta_k + \alpha_{k-1})(\beta_k + \alpha_k) z_\lambda^2 \end{pmatrix}. \tag{5.22}$$

An important property of any fourth-order composition method is that it inevitably contains a negative coefficient, which in our case is $\alpha_4 = \beta_2$. This entails a stability restriction that is conditional even for an implicitly treated conduction term. This is illustrated in Figure 5.3, where parts of the upper right half of the stability region for (5.21) is shown. Stability is guaranteed as long as $z_\gamma < 2.4$ and $z_\lambda < 3$, or equivalently, if $\tau < 2.4/\gamma$ and $\tau < 3/\sqrt{\lambda}$.

5.4.3 Fourth-order global Richardson extrapolation

As already mentioned in the previous section, when $\sigma > 0$ the stability condition may be very restrictive even for moderately conductive materials. In these cases, high-order composition methods and SSPRK methods are not competitive. Instead, one would prefer to use explicit methods which treat the conduction term in an unconditionally stable manner. Since the second-order composition method is such a method, extending it to higher order through Richardson extrapolation is an obvious alternative. We refer to [11] for a detailed discussion on the stability

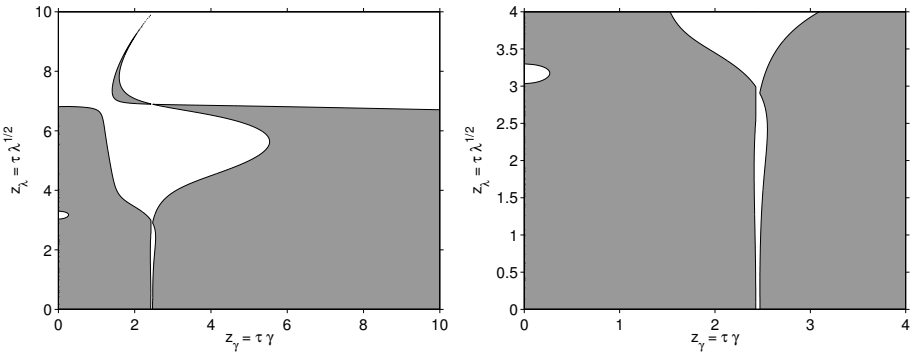


Figure 5.3: *Stability region (shaded area) of the fourth-order composition method. The right plot zooms in on the region where stability is guaranteed. (Cf. Figure 5.1 in [11].)*

properties of the fourth-order local and global versions of the Richardson extrapolation. Here we first recall the construction of the fourth-order global Richardson extrapolation (GEX4)

$$u_\tau^{\mathrm{gex4}} = \frac{4}{3}u_{\frac{\tau}{2}}^{\mathrm{co2}} - \frac{1}{3}u_\tau^{\mathrm{co2}}, \tag{5.23}$$

where $u_{\frac{\tau}{2}}^{\mathrm{co2}}$ and u_τ^{co2} denote the results at final time computed by the second-order composition method with time steps $\frac{\tau}{2}$ and τ, respectively. Since extrapolation only takes place once at the final time of the integration, this method has the same stability properties as the second-order composition method. Note that it only needs three times as much computational work per time step.

For long time integration and in the absence of damping, global extrapolation may not be sufficiently effective in annihilating leading error terms. In these cases, the local version of Richardson extrapolation – when the extrapolation is performed at each time step – is usually more beneficial. The local version of (5.23) is, however, not unconditionally stable with respect to z_γ. Instead, we can use the fourth-order local extrapolation (LEX4) defined as

$$u_\tau^{\mathrm{lex4}} = \frac{9}{8}u_{\tau/3}^{\mathrm{co2}} - \frac{1}{8}u_\tau^{\mathrm{co2}}, \tag{5.24}$$

where the work per time step is approximately four times as much as that of CO2. The amplification operator of LEX4 for the test model (5.16) reads

$$\frac{9}{8}\mathcal{M}_{\mathrm{co2}}^3(z_\lambda/3, z_\gamma/3) - \frac{1}{8}\mathcal{M}_{\mathrm{co2}}(z_\lambda, z_\gamma), \tag{5.25}$$

where $\mathcal{M}_{\mathrm{co2}}(z_\lambda, z_\gamma)$ denotes the amplification operator (5.20) of CO2. Figure 5.4 shows the associated stability region $\mathcal{S}$, which indicates an approximate stability interval $0 \le z_\lambda \le 2.85$ and unconditional stability for z_γ.

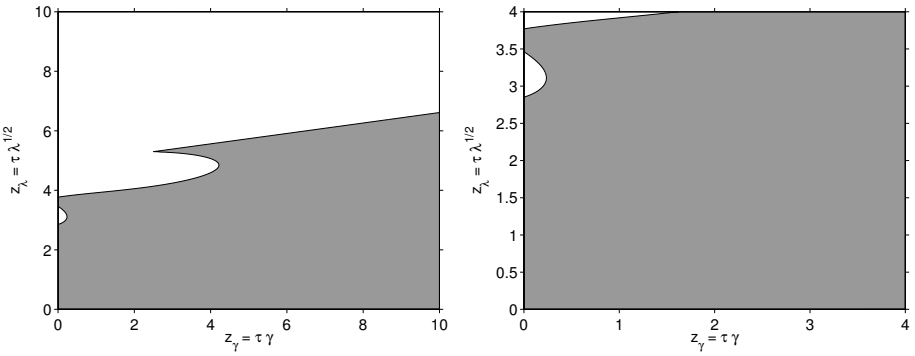


Figure 5.4: *Stability region (shaded area) of the fourth-order local Richardson extrapolation (5.24). The right plot zooms in on the region where stability for wave term z_λ is guaranteed. Stability for the conduction term z_γ is unconditional.*

5.5 Numerical experiments

In this section, we perform numerical tests to establish the convergence rates of the fully discrete systems that result from the spatial and time discretisations described in the previous two sections. We also carry out a numerical dispersion analysis of the semi- and fully discrete system with DG spatial discretisation. This is done in the following way: *i)* solve the time-harmonic eigenvalue problem, which corresponds to the semi-discrete system with Fourier mode initial conditions; *ii)* apply a chosen time-integration method to the test model (5.16) with the computed semi-discrete numerical frequency. This approach has two main advantages over simply solving the eigenvalue problem that results from applying the amplification matrix directly to (5.11). First, it is more efficient because we solve an eigenvalue problem that is smaller and always symmetric. Second, it makes it possible to study the dispersion (and dissipation) properties of the time-integration scheme separately from those of the semi-discrete scheme.

5.5.1 Convergence and comparison of performance

We use a simple test example to illustrate the numerical performance of the two spatial discretisation techniques described in Section 5.2. For both methods, the predicted convergence rate of the semi-discrete system is $\mathcal{O}(h^{p+1})$ in the $L^2(\Omega)$ norm and $\mathcal{O}(h^p)$ in the energy norm for smooth solutions; see for example [56, 46, 68, 33, 34]. It is thus natural to choose the time-integration method such that it guarantees at least the same order of convergence. Therefore, if the polynomial order in the FEM is at most one we use the second-order composition method; if the polynomial order is two or three we apply one of the possible fourth-order methods described in Section 5.4.

The numerical tests are implemented in *hp*GEM[4] [62], a general finite element package suitable for solving a variety of physical problems in fluid dynamics and electromagnetism. To integrate the semi-discrete system in time we use PETSc [5], which is particularly efficient in computing matrix-vector multiplications and solving linear systems for large sparse matrices. As a stopping criterion in the linear solver for the $H(\mathrm{curl})$-conforming method, we set the tolerance at $\texttt{tol} = 10^{-8}$.

In the example, we consider (5.1) in the cubic domain $\Omega = (0,1)^3$. We define the time-independent field

$$\bar{\mathbf{E}}(x,y,z) = \begin{pmatrix} \sin(\pi y)\sin(\pi z) \\ \sin(\pi z)\sin(\pi x) \\ \sin(\pi x)\sin(\pi y) \end{pmatrix}$$

and choose the source term to be

$$-\frac{\partial \mathbf{J}}{\partial t} = \left(\varepsilon_{\mathrm{r}}\eta''(t) + \sigma\eta'(t) + 2\pi^2\eta(t)\right)\bar{\mathbf{E}}(x,y,z).$$

Thus the exact solution reads

$$\mathbf{E}(t,x,y,z) = \eta(t)\bar{\mathbf{E}}(x,y,z), \quad \eta(t) = \sum_{k=1}^{3} \cos\omega_k t, \tag{5.26}$$

with $\omega_1 = 1$, $\omega_2 = 1/2$, $\omega_3 = 1/3$, $\varepsilon_{\mathrm{r}} = 1$. We either set $\sigma = 0$ or $\sigma = 60\pi$, and we integrate until final time $T_{\mathrm{end}} = 12\pi$ (exactly one time period).

When the globally $H(\mathrm{curl})$-conforming discretisation [56, 45] is used, we need to solve a linear system at each time step. Since the matrix M_ε is positive definite, a natural choice of linear solver is the preconditioned conjugate gradient (PCG) method. For simplicity, we apply the incomplete Cholesky preconditioner for all meshes and polynomial orders. We emphasise that preconditioning is not an issue for the DG method since we can simply invert the block-diagonal mass matrix at negligible cost.

As a first example, we run (5.26) with $\sigma = 0$ and final time $T_{\mathrm{end}} = 12\pi$, on a sequence of structured meshes with $N_{\mathrm{el}} = 5, 40, 320, 2560, 20480, 163840$ elements. In each mesh the largest face diameter h is exactly half that of the previous mesh. We plot the convergence rates in Figure 5.5 in both the $L^2(\Omega)$-norm and the energy norm for polynomial orders $p = 1, 2, 3$. Note that the convergence is shown as a function of degrees of freedom, which is equivalent to showing the convergence as function of $1/h$ in the DG case. However, in the $H(\mathrm{curl})$-conforming case there is some difference between the two, as the number of degrees of freedom generally increase slightly more than eightfold when h is halved. Nevertheless, we can see that the expected convergence rates are achieved asymptotically for both the DG and the $H(\mathrm{curl})$-conforming methods. We can also observe that it takes fewer degrees of freedom for the $H(\mathrm{curl})$-conforming discretisation to reach a given accuracy.

[4] The software is available at `http://wwwhome.math.utwente.nl/~hpgemdev` free of charge

Furthermore, we can confirm the well-established observation that the use of high-order approximations pays off (at least for smooth solutions) in terms of accuracy per degrees of freedom.

To gain further insight into the computational costs of the time integration, we show the performance of the DG method in Tables I and III; and that of the $H(\mathrm{curl})$-conforming method in Tables II and IV. In this particular example, we use a structured mesh with 320 elements and an unstructured one with 432 elements.[5] Although the accuracy of the two methods is comparable, the computational costs are not and the pattern changes dramatically as the order increases. The total number of matrix-vector multiplications (matvecs) needed to integrate until T_{end} is always higher for the $H(\mathrm{curl})$-conforming case than for the DG method. This is not surprising given that at each time step a linear system has to be solved. However, this seemingly unfavourable property does not manifests itself in longer computational time for $p = 1$ and $p = 2$ on structured meshes, thanks in part to the smaller size of the system and in part to a weaker time-step restriction in the $H(\mathrm{curl})$-conforming FEM. The situation is different for $p = 3$. Here, the increased number of matvecs translates readily into more CPU time on both structured and unstructured meshes. This is partly because of a trade-off between the conditioning of the mass matrix and the use of the hierarchic basis. Mass matrices based on hierarchic bases tend to be relatively badly conditioned. This does not influence the performance of the DG method. But it renders the $H(\mathrm{curl})$-conforming method less effective because the number of iterations in solving the linear system at each time step grows significantly with the polynomial order. This effect is even more pronounced on unstructured meshes, where DG performs slightly better for $p = 2$ already and where the $H(\mathrm{curl})$-conforming computation for $p = 3$ is excessively long – which is one reason why we only completed one of them.

The choice of the time-integration method does not influence the computational results much in this example. Nevertheless, LEX4 appears to be the most efficient thanks to the balance between the allowable time-step size and the computational work needed per time step. We also note that for this particular mesh the use of fourth-order time-integration methods may not be necessary even for $p = 2, 3$. This is solely because the spatial error is not yet in the asymptotic regime and therefore dominates. We take a closer look at this shortly in terms of numerical dispersion and dissipation.

On structured meshes, we repeat example (5.26) with conductivity $\sigma = 60\pi$, which corresponds to the dimensional value $\tilde{\sigma} = 0.5\,\mathrm{S\,m^{-1}}$, typical of the human abdomen. The convergence results are shown in Figure 5.6, from which it appears that they are similar to the nonconductive case except that optimal rates of convergence are reached sooner. On unstructured meshes, the example is repeated with conductivity $\sigma = 450\pi$, a value more typical of seawater. See Table V for the conductivity of a small selection of materials[6].

[5] A mesh of 320 or 432 tetrahedra is sufficient to compare the different methods from the point of view of accuracy and computational work. A finer mesh would naturally give a more accurate solution but the relative performance of the methods would remain the same.

[6] Source: `en.wikipedia.org/wiki/Electrical_conductivity`

Table I: *Computational costs of the DG method for example (5.26) with $\sigma = 0$. A structured mesh of 320 elements is used with spatial polynomial orders $p = 1, 2, 3$.*

	method	# DoF	$L^2(\Omega)$ error	# matvecs	τ	CPU time
$p = 1$	CO2	3840	1.2174e-01	4526	0.0167	8s
$p = 2$	CO2	9600	1.1696e-02	7542	0.0100	114s
$p = 2$	GEX4	9600	1.2303e-02	22624	0.0100	342s
$p = 3$	CO4	19200	7.0432e-04	35192	0.0107	2013s
$p = 3$	GEX4	19200	9.0148e-04	31672	0.0071	1863s
$p = 3$	LEX4	19200	6.1762e-04	28154	0.0107	1623s

Table II: *Computational costs of the H(curl)-conforming method for example (5.26) with $\sigma = 0$. A structured mesh of 320 elements is used with spatial polynomial orders $p = 1, 2, 3$.*

	method	# DoF	$L^2(\Omega)$ error	# matvecs	τ	CPU time
$p = 1$	CO2	504	2.7283e-01	7783	0.0417	2s
$p = 2$	CO2	2388	1.3642e-02	59201	0.0250	87s
$p = 2$	GEX4	2388	1.2942e-02	180067	0.0250	264s
$p = 3$	CO4	6640	7.4117e-04	817880	0.0268	19683s
$p = 3$	GEX4	6640	7.7523e-04	736276	0.0179	17492s
$p = 3$	LEX4	6640	8.5234e-04	653611	0.0268	15510s

Table III: *Computational costs of the DG method for example (5.26) with $\sigma = 0$. An unstructured mesh of 432 elements is used with spatial polynomial orders $p = 1, 2, 3$.*

	method	# DoF	$L^2(\Omega)$ error	# matvecs	τ	CPU time
$p = 1$	CO2	5184	1.9583e-01	11878	0.00635	41s
$p = 2$	CO2	12960	1.3396e-02	19796	0.00381	429s
$p = 2$	GEX4	12960	1.4316e-02	59384	0.00381	1263s
$p = 3$	CO4	25920	1.4311e-03	92372	0.00408	7585s
$p = 3$	GEX4	25920	1.5558e-03	83134	0.00272	6749s
$p = 3$	LEX4	25920	1.4038e-03	73898	0.00408	5909s

The computational work, depicted in Tables VI–IX, also shows a similar pattern to the conduction-free case, except when the fourth-order composition method is used. In that case, the conduction term poses a stricter time-step size than the wave term and increases the number of time steps and thus the computational cost. On the structured mesh with 320 elements and $\sigma = 60\pi$, this only affects the H(curl)-conforming discretisation because the stiffness matrix in the DG method has a significantly larger spectral radius (and therefore it still determines the stability condition). On the unstructured mesh with 432 elements and $\sigma = 450\pi$, however, it already affects the DG discretisation too. This indicates that large val-

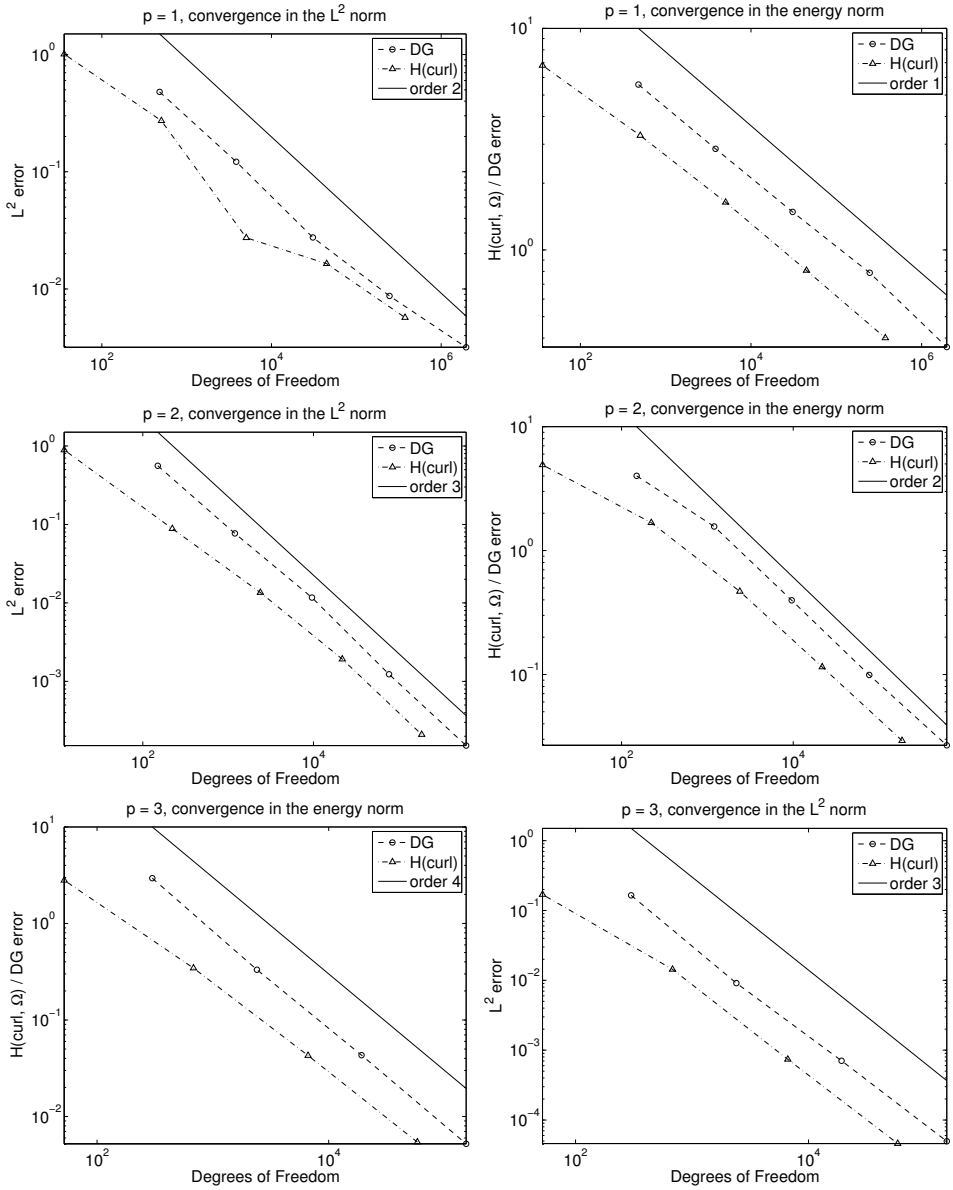


Figure 5.5: *Convergence plots in the L^2-norm (left column) and in the energy norm (right column) for test example (5.26) with $\sigma = 0$. In each plot the convergence rates of the DG method and the $H(\mathrm{curl})$-conforming method are shown along with the expected order of convergence.*

Table IV: *Computational costs of the $H(\mathrm{curl})$-conforming method for example (5.26) with $\sigma = 0$. An unstructured mesh of 432 elements is used with spatial polynomial orders $p = 1, 2, 3$.*

	method	# DoF	$L^2(\Omega)$ error	# matvecs	τ	CPU time
$p = 1$	CO2	744	1.9113e-01	33691	0.02380	18s
$p = 2$	CO2	3420	1.7294e-02	169129	0.01429	963s
$p = 2$	GEX4	3420	1.8860e-02	406784	0.01429	2778s
$p = 3$	CO4	9360	—	>1e+07	0.01530	>5e+05s
$p = 3$	GEX4	9360	1.9676e-03	21337490	0.01020	782647s
$p = 3$	LEX4	9360	—	>1e+07	0.01530	>5e+05s

Table V: *Electrical conductivity of some materials measured in Siemens per metre ($\mathrm{S\,m^{-1}}$). For the dimensionless value a multiplication by 120π is needed. Source:* ***en.wikipedia.org/wiki/Electrical_conductivity****.*

Material	Conductivity ($\mathrm{S\,m^{-1}}$)	Note
Silver	63.0e+06	Best electrical conductor
Copper	59.6e+06	
Gold	45.2e+06	Commonly used in electrical contacts
Aluminium	37.8e+06	
Seawater	4.8	For average salinity of 35 g/kg
Human Body	0.006–1.5	Varies from bone to spinal fluids
Drinking water	0.0005–0.05	
Deionised water	5.5e-06	Lowest value, with monoatomic gases
Air	5e-15	Varies slightly depending on humidity

ues of σ prohibit the use of fourth-order (or, indeed, any high-order) composition methods, as well as explicit RK methods, such as SSPRK. Instead, Richardson extrapolation based on the second-order composition method may be used since they are unconditionally stable with respect to the conductivity term. Similarly to the conduction-free case we killed the $H(\mathrm{curl})$-conforming computations after almost six days – already significantly more than what the DG computations take.

5.5.2 Numerical dispersion analysis

To investigate the dispersion and dissipation properties of the fully discrete schemes, we consider the semi-discrete system (5.11) with $\sigma = 0$ and $j = 0$,

$$\begin{pmatrix} u' \\ v' \end{pmatrix} = \mathcal{A} \begin{pmatrix} u \\ v \end{pmatrix} \quad \text{with} \quad \mathcal{A} = \begin{pmatrix} 0 & I \\ -M_\varepsilon^{-1} S_\mu & 0 \end{pmatrix}, \tag{5.27}$$

and assume a plane wave exact solution

$$\mathbf{E}(\mathbf{x}, t) = \hat{\boldsymbol{E}} \exp(-\mathrm{i}\omega t) \exp(\mathrm{i}\boldsymbol{k} \cdot \boldsymbol{x}) \tag{5.28}$$

Table VI: *Computational costs of the DG method for example (5.26) with $\sigma = 60\pi$. A mesh of 320 elements is used with spatial polynomial orders $p = 1, 2, 3$.*

	method	# DoF	$L^2(\Omega)$ error	# matvecs	τ	CPU time
$p = 1$	CO2	3840	6.6817e-02	4526	0.0167	8s
$p = 2$	CO2	9600	8.4244e-03	7542	0.0100	113s
$p = 2$	GEX4	9600	8.4243e-03	22624	0.0100	341s
$p = 3$	CO4	19200	5.5619e-04	35192	0.0107	2012s
$p = 3$	GEX4	19200	5.5612e-04	31672	0.0071	1864s
$p = 3$	LEX4	19200	5.5612e-04	28154	0.0107	1623s

Table VII: *Computational costs of the H(curl)-conforming method for example (5.26) with $\sigma = 60\pi$. A mesh of 320 elements is used with spatial polynomial orders $p = 1, 2, 3$.*

	method	# DoF	$L^2(\Omega)$ error	# matvecs	τ	CPU time
$p = 1$	CO2	504	1.1789e-01	6253	0.0417	1s
$p = 2$	CO2	2388	1.2315e-02	56301	0.0250	82s
$p = 2$	GEX4	2388	1.2314e-02	166303	0.0250	247s
$p = 3$	CO4	6640	7.3357e-04	1717157	0.0127	40862s
$p = 3$	GEX4	6640	7.3358e-04	734732	0.0179	17472s
$p = 3$	LEX4	6640	7.3358e-04	653024	0.0268	15498s

Table VIII: *Computational costs of the DG method for example (5.26) with $\sigma = 450\pi$. An unstructured mesh of 432 elements is used with spatial polynomial orders $p = 1, 2, 3$.*

	method	# DoF	$L^2(\Omega)$ error	# matvecs	τ	CPU time
$p = 1$	CO2	5184	4.6168e-02	11878	0.00635	41s
$p = 2$	CO2	12960	8.1650e-03	19796	0.00381	422s
$p = 2$	GEX4	12960	8.1650e-03	59384	0.00381	1240s
$p = 3$	CO4	25920	8.5690e-04	222072	0.00170	17606s
$p = 3$	GEX4	25920	8.5671e-04	83134	0.00272	6618s
$p = 3$	LEX4	25920	8.5690e-04	73898	0.00136	5906s

with periodic boundary conditions and $\hat{\boldsymbol{E}} = \mathbf{1}$. In (5.28), $\mathrm{i}^2 = -1$, ω denotes the angular frequency, $k = (k_x, k_y, k_z)^T$ is the wave number. Between these quantities the (exact) dispersion relation $\omega^2 = k^2/c^2$ holds with $k^2 = k_x^2 + k_y^2 + k_z^2$ and with $c = 1/\left(\varepsilon_\mathrm{r}\mu_\mathrm{r}\right)^{1/2}$, which is the speed of light.

As a first step, we project the exact initial conditions $\mathbf{E}(\mathbf{x}, 0)$ and $\partial_t \mathbf{E}(\mathbf{x}, 0)$ onto

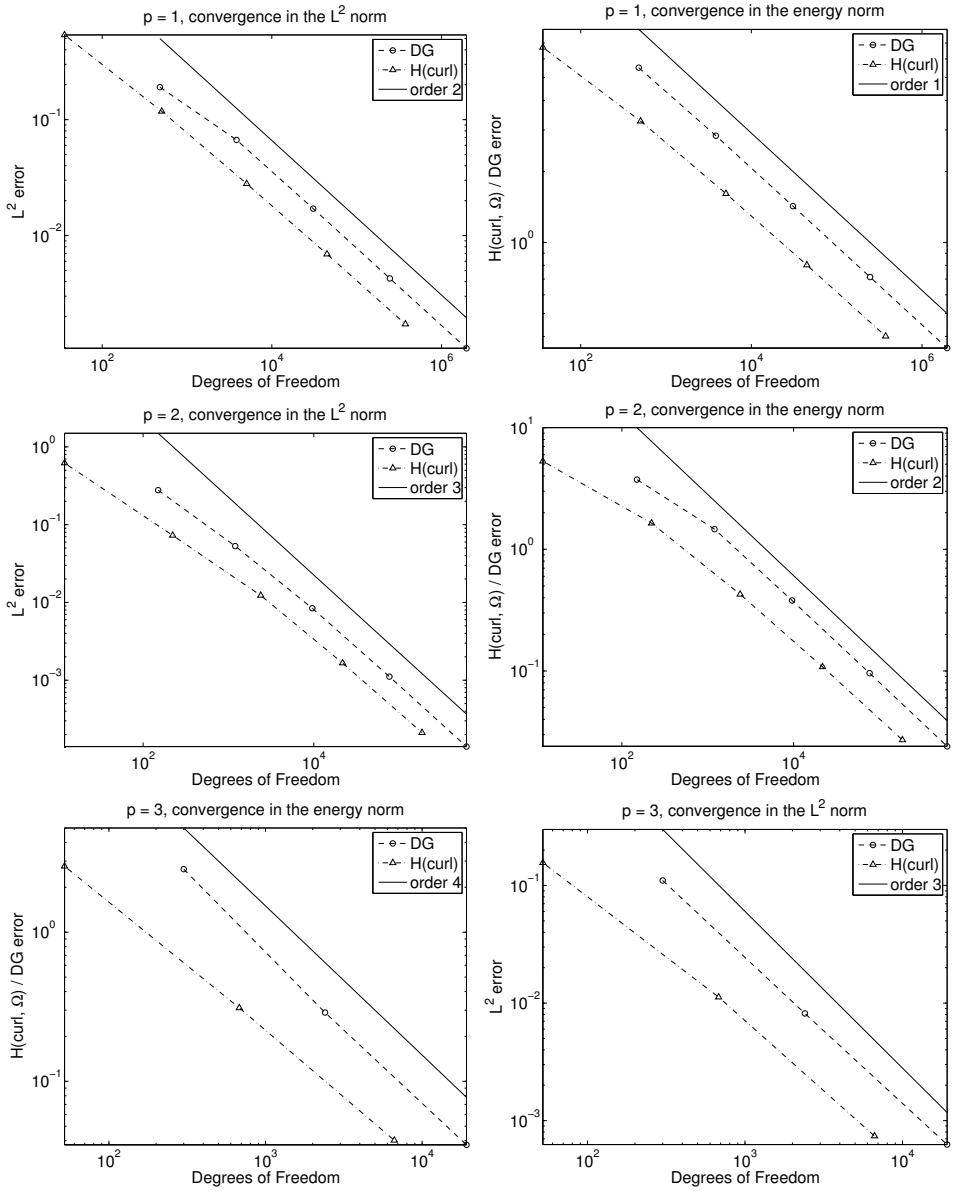


Figure 5.6: *Convergence plots in the L^2-norm (left column) and in the energy norm (right column) for test example (5.26) with $\sigma = 60\pi$. In each plot the convergence rates of the DG method and the $H(\mathrm{curl})$-conforming method are shown along with the expected order of convergence.*

Table IX: *Computational costs of the $H(\mathrm{curl})$-conforming method for example (5.26) with $\sigma = 450\pi$. An unstructured mesh of 432 elements is used with spatial polynomial orders $p = 1, 2, 3$.*

	method	# DoF	$L^2(\Omega)$ error	# matvecs	τ	CPU time
$p = 1$	CO2	744	1.2498e-01	32049	0.02380	17s
$p = 2$	CO2	3420	1.3102e-02	163451	0.01429	938s
$p = 2$	GEX4	3420	1.3102e-02	490287	0.01429	2677s
$p = 3$	CO4	9360	—	>1e+07	0.01530	>5e+05s
$p = 3$	GEX4	9360	—	>1e+07	0.01020	>5e+05s
$p = 3$	LEX4	9360	—	>1e+07	0.01530	>5e+05s

the finite-element space

$$\begin{aligned} E_h^j(0) &= \big(\mathbf{E}(\mathbf{x},0), \boldsymbol{\psi}_j\big)_\Omega, \qquad j = 1 \dots N, \\ \frac{\mathrm{d}}{\mathrm{d}t} E_h^j(0) &= \big(\partial_t \mathbf{E}(\mathbf{x},0), \boldsymbol{\psi}_j\big)_\Omega, \qquad j = 1 \dots N. \end{aligned} \tag{5.29}$$

We can now obtain the initial conditions for (5.27) through the relations $u_0 = u(0) = M_{\varepsilon_\mathrm{r}}^{-1} E_h(0)$ and $v_0 = v(0) = u'(0) = M_{\varepsilon_\mathrm{r}}^{-1} \frac{\mathrm{d}}{\mathrm{d}t} E_h(0)$. The time-exact discrete Fourier mode at time level $n\tau$ is then defined as

$$\begin{pmatrix} u_n \\ v_n \end{pmatrix} = \nu^n \begin{pmatrix} u_0 \\ v_0 \end{pmatrix} \quad \text{with} \quad \nu^n = \mathrm{e}^{-\mathrm{i}\omega_h n\tau}, \tag{5.30}$$

where ν^n is the exact amplification factor and ω_h is the semi-discrete numerical frequency.

To first see the impact of the space discretisation only, we consider the semi-discrete equation

$$M_\varepsilon u'' + S_\mu u = 0 \tag{5.31}$$

with periodic boundary conditions and a plane wave initial condition (5.28). In this case, (5.31) is equivalent to the discrete time-harmonic Maxwell eigenvalue problem

$$S_\mu u - \omega_h^2 M_\varepsilon u = 0 \tag{5.32}$$

with periodic boundary conditions. All semi-discrete eigenvalues ω_h^2 of (5.32) are real and non-negative, which entails that the space discretisation imposes no dissipation. In Table X, we show the numerical frequencies of the spatial DG discretisation for the Fourier mode with $k_x = 2\pi, k_y = -2\pi, k_z = 0$, i.e. with exact angular frequency $\omega_\mathrm{ex} = \sqrt{8}\pi$. The number of elements for each mesh is $N_\mathrm{el} = 5(\frac{1}{h})^3$ and in each element the local number of degrees of freedom is $\frac{1}{2}(p+1)(p+2)(p+3)$. To solve the eigenvalue discrete problem (5.32) of this size the `Matlab` implementation[7] of the Jacobi-Davidson iterative method [72, 73] is used. We note that for

[7] The software is available at `http://www.math.uu.nl/people/sleijpen` free of charge

Table X: *Semi-discrete frequencies ω_h of the DG method that approximate the exact frequency $\omega_{\text{ex}} = \sqrt{8}\pi$*

	$h = \frac{1}{2}$	$h = \frac{1}{4}$	$h = \frac{1}{8}$	$h = \frac{1}{16}$	ω_{ex}
$p = 1$	—	9.4286	9.0469	8.9271	8.8858
$p = 2$	9.4738	8.9276	8.8887	—	8.8858
$p = 3$	8.9146	8.8875	8.8858	—	8.8858

Table XI: *Frequency error $\omega_h - \omega_{\text{ex}}$ of the DG semi-discrete system with exact frequency $\omega_{\text{ex}} = \sqrt{8}\pi$*

	$h = \frac{1}{2}$	$h = \frac{1}{4}$	$h = \frac{1}{8}$	$h = \frac{1}{16}$
$p = 1$	—	5.4283e-01	1.6117e-01	4.1380e-02
$p = 2$	5.8800e-01	4.1831e-02	2.9628e-03	—
$p = 3$	2.8869e-02	1.7173e-03	3.0850e-05	—

other Fourier modes the same approximation properties apply as long as $\omega_h h$ is in the same region as shown in the tables. The frequency errors for the same meshes and polynomial orders are depicted in Table XI. Note that the frequency errors are signed, indicating phase advance.

To include the time integration in the dispersion analysis it suffices to apply a chosen time-integration method to the test model (5.16) with $\gamma = 0$. We are allowed to do that because the eigenvalues of $\tilde{S}_\mu$ are the same as the eigenvalues of $M_\varepsilon^{-1} S_\mu$, that is $\lambda = \omega_h^2$. Let $\mathcal{M}$ denote the amplification operator of any of the time-integration methods described in Section 5.4. So instead of (5.30) we now have the fully discrete Fourier mode at time level $n\tau$,

$$\nu_h^{n+1} \begin{pmatrix} u_0 \\ v_0 \end{pmatrix} = \mathcal{M} \nu_h^n \begin{pmatrix} u_0 \\ v_0 \end{pmatrix}, \tag{5.33}$$

which reduces to the eigenvalue problem

$$\nu_h \begin{pmatrix} u_0 \\ v_0 \end{pmatrix} = \mathcal{M} \begin{pmatrix} u_0 \\ v_0 \end{pmatrix}. \tag{5.34}$$

Solving this eigenvalue problem will produce two eigenpairs, representing two waves with the same wave number but travelling in opposite directions. Without loss of generality, we can discard the one with negative real part and establish the dispersive and dissipative properties of the fully discrete scheme through the relation

$$\nu_h = \mathrm{e}^{-\mathrm{i}\omega_h^\tau \tau},$$

where ω_h^τ represents the fully discrete numerical frequency. The real part of ω_h^τ defines the actual angular frequency in the discrete dispersion relation, while a

Table XII: *Frequency error imposed only by the time integration,* $\mathrm{Re}(\omega_h^\tau) - \omega_h$*, of the SSPRK(4, 3) method for semi-discrete numerical frequencies* ω_h *taken from Table X.*

	$h = \frac{1}{2}$	$h = \frac{1}{4}$	$h = \frac{1}{8}$	$h = \frac{1}{16}$
$p = 1$	—	7.1799e-05	3.6525e-06	2.1360e-07
$p = 2$	1.5242e-04	7.0867e-06	4.3347e-07	—
$p = 3$	2.9293e-05	1.8039e-06	1.1265e-07	—

Table XIII: *Frequency error imposed only by the time integration,* $\mathrm{Re}(\omega_h^\tau) - \omega_h$*, of the CO2 method for semi-discrete numerical frequencies* ω_h *taken from Table X.*

	$h = \frac{1}{2}$	$h = \frac{1}{4}$	$h = \frac{1}{8}$	$h = \frac{1}{16}$
$p = 1$	—	9.7283e-03	2.1439e-03	5.1472e-04
$p = 2$	1.4229e-02	2.9674e-03	7.3172e-04	—
$p = 3$	6.0353e-03	1.4930e-03	3.7292e-04	—

negative imaginary part indicates numerical dissipation. A non-negligible positive imaginary part would mean instability.

We show the frequency errors of the time-integration schemes SSPRK(4, 3), CO2 and LEX4 in Tables XII–XIV. They show that the frequency error of the time-integration method is at least an order smaller than the one of the DG method, as long as the order of the time-integration method is on a par with the order of the DG method. When this is not the case, such as when CO2 is used for $p = 2$ or $p = 3$, the frequency error of the time integration is commensurate with, or exceeds that of the DG discretisation.

Composition methods, such as CO2 and CO4, are known to be non-dissipative [37]. Thus combining them with a symmetric spatial discretisation results in an energy-conservative fully-discrete discretisation. Global Richardson extrapolation based on a composition method naturally inherits this property. However, local Richardson extrapolation may introduce a slight dissipation even when based on a non-dissipative scheme such as CO2. We show this in Table XV and note that the error is generally too small to have a real impact on simulations arising in practice. By comparison, the SSPRK(4, 3) scheme introduces a much more significant level of dissipation, shown in Table XVI.

Finally, we note that if a time-dependent boundary condition is used in (5.1) instead of a homogeneous one, order reduction may occur. See [11] for the possible effects of this.

5.6 Concluding remarks

We have investigated the time-dependent second-order Maxwell equation in three spatial dimensions. A direct comparison between the high-order DG-FEM and the

Table XIV: *Frequency error imposed only by the time integration,* $\mathrm{Re}(\omega_h^\tau) - \omega_h$*, of the LEX4 method for semi-discrete numerical frequencies* ω_h *taken from Table X. The negative values indicate that the frequency error caused by the time integration counteracts that of imposed by the space discretisation.*

	$h = \frac{1}{2}$	$h = \frac{1}{4}$	$h = \frac{1}{8}$	$h = \frac{1}{16}$
$p = 1$	—	-7.9554e-06	-4.0558e-07	-2.3730e-08
$p = 2$	-1.6866e-05	-7.8671e-07	-4.8152e-08	—
$p = 3$	-3.2488e-06	-2.0035e-07	-1.2516e-08	—

Table XV: *Imaginary part of the numerical frequency,* $\mathrm{Im}(\omega_h^\tau)$*, for the LEX4 time-integration method, where the semi-discrete numerical frequencies* ω_h *are taken from Table X. This term is responsible for numerical dissipation.*

	$h = \frac{1}{2}$	$h = \frac{1}{4}$	$h = \frac{1}{8}$	$h = \frac{1}{16}$
$p = 1$	—	-6.9642e-07	-1.6998e-08	-4.9043e-10
$p = 2$	-1.7825e-06	-3.9053e-08	-1.1891e-09	—
$p = 3$	-2.3027e-07	-7.0689e-09	-2.2071e-10	—

Table XVI: *Imaginary part of the numerical frequency,* $\mathrm{Im}(\omega_h^\tau)$*, for the SSPRK(4, 3) time-integration method, where the semi-discrete numerical frequencies* ω_h *are taken from Table X. This term is responsible for numerical dissipation.*

	$h = \frac{1}{2}$	$h = \frac{1}{4}$	$h = \frac{1}{8}$	$h = \frac{1}{16}$
$p = 1$	—	-7.5911e-04	-8.0688e-05	-9.5692e-06
$p = 2$	-1.3346e-03	-1.3217e-04	-1.6251e-05	—
$p = 3$	-3.8256e-04	-4.7337e-05	-5.9156e-06	—

high-order $H(\mathrm{curl})$-conforming FEM on both structured and unstructured meshes was provided when $H(\mathrm{curl})$-conforming hierarchic basis functions are used. It has revealed that, in case hierarchic basis functions are used, the computational cost is already lower for DG-FEM when $p = 3$, or even $p = 2$ on unstructured meshes. The computational tests have highlighted the fact that the inclusion of moderate conductivity renders many of the popular time-integration methods uncompetitive owing to a stringent time-step restriction. In these cases, global or local Richardson extrapolations based on the second-order composition method provide a viable alternative as they treat the conductivity term implicitly.

Through a numerical dispersion and dissipation analysis, we have also shown that the spatial discretisation dominates the frequency error as long as the order of the time integration is at least the same as the order of the spatial discretisation. Since the semi-discrete system is symmetric and therefore conserves (the discrete) energy, applying a composition method to integrate in time results in a fully-

discrete scheme that also conserves (the discrete) energy.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

In this thesis, we have derived, analysed and implemented high-order finite element methods for the Maxwell equations. The work carried out can be briefly summarised in the following points.

- We have shown that by applying the $(p+1)$st-order SSP-RK scheme to a nodal discontinuous Galerkin (DG) spatial discretisation with pth-order polynomials we can retain $(p+1)$st-order convergence in the pointwise l_2-norm. It has been also demonstrated, through numerical Fourier analysis, that for this method the fully discrete scheme is both dispersive and dissipative. However, the rate of convergence for the dispersion and dissipation error is higher than that of the pointwise l^2-error, and therefore often negligible.

- We have provided a framework to implement high-order finite element methods on tetrahedral meshes when hierarchic $H(\text{curl})$-conforming basis functions are used. We have highlighted, through a simple example, a possible discrepancy in the definition of one type of basis functions in the construction of the basis. A simple solution to this problem has also been proposed.

- We have derived optimal penalty parameters and error estimates for symmetric discontinuous Galerkin discretisations of the second-order time-harmonic Maxwell equations. Numerical examples on three-dimensional meshes have been presented to verify the sharpness of the theoretical estimates and also to show asymptotic convergence.

- For the time-dependent second-order Maxwell equation, we have provided a direct comparison between the high-order DG-FEM and the high-order $H(\text{curl})$-conforming FEM on both structured and unstructured tetrahedral meshes. It has revealed that, in case $H(\text{curl})$-conforming hierarchic basis functions are used, the computational costs are already lower for DG-FEM

when $p = 3$, or even $p = 2$ on unstructured meshes. We have also highlighted that conductivity renders many of the time-integration methods uncompetitive, in which cases global or local Richardson extrapolations based on the second-order composition method may be the alternative.

It is also clear, however, that many research questions have remained unanswered. Without aiming for completeness, we list those that could be the most direct continuations of the research presented here.

- Despite the improvements presented in this thesis, the CFL condition of high-order finite element methods is still too restrictive, since it is proportional to the inverse of the polynomial order. For the advection equation, it is possible to get rid of this dependence by the use of covolume filters. However, the generality of this approach is still to be investigated.

- The motivation behind using hierarchic $H(\mathrm{curl})$-conforming basis functions in several chapters of this thesis is that they are extremely useful in *hp*-adaptation. Therefore, a natural extension of this work would be to implement an *hp*-adaptive algorithm for the Maxwell equations. The key in such an algorithm is the error estimate on which the subsequent refinement strategy is based. A posteriori error estimates would, in principle, serve this purpose well, but they are not yet sufficiently accurate for high-order $H(\mathrm{curl})$-conforming FEM and DG-FEM.

- The relative poor performance of the $H(\mathrm{curl})$-conforming FEM compared with DG-FEM for the time-dependent second-order equation provides food for thought. One of the reasons is the bad conditioning of the mass matrix that result from hierarchic bases. This suggests that the use of more advanced preconditioning techniques may significantly improve the performance of the $H(\mathrm{curl})$-conforming FEM. Similarly, the DG-FEM discretisation of the time-harmonic second-order equations suffers from slow convergence of the iterative solver, mainly owing to the indefinite nature of the system. An appropriate preconditioner should greatly improve the performance of high-order DG-FEM on tetrahedral meshes.

BIBLIOGRAPHY

[1] M. Ainsworth. Dispersive and dissipative behaviour of high order discontinuous Galerkin finite element methods. *J. Comput. Phys.*, 198(1):106–130, 2004.

[2] M. Ainsworth and J. Coyle. Hierarchic finite element bases on unstructured tetrahedral meshes. *Internat. J. Numer. Methods Engrg.*, 58(14):2103–2130, 2003.

[3] M. Ainsworth, P. Monk, and W. Muniz. Dispersive and dissipative properties of discontinuous Galerkin finite element methods for the second-order wave equation. *J. Sci. Comput.*, 27:5–40, 2006.

[4] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini. Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.*, 39(5):1749–1779, 2001/02.

[5] S. Balay, K. Buschelman, W. D. Gropp, D. Kaushik, M. G. Knepley, L. C. McInnes, B. F. Smith, and H. Zhang. PETSc Web page, 2001. http://www.mcs.anl.gov/petsc.

[6] F. Bassi, S. Rebay, G. Mariotti, S. Pedinotti, and M. Savini. A high order accurate discontinuous finite element method for inviscid and viscous turbomachinery flows. In *Proceedings of the 1997 2nd European Conference on Turbomachinery - Fluid Dynamics and Thermodynamics, (Antwerpen, Belgium)*, pages 99–108, 1997.

[7] S. Benhassine, J. Carpes, and L. Pichon. Comparison of mass lumping techniques for solving the 3D Maxwell's equations in the time domain. *IEEE Trans. Magn.*, 36(4):1548–1552, 2000.

[8] A. Bossavit. A rationale for 'edge-elements' in 3-D fields computations. *IEEE Trans. Magn.*, 24(1):74–79, 1988.

[9] A. Bossavit. Solving Maxwell equations in a closed cavity, and the question of 'spurious modes'. *IEEE Trans. Magn.*, 26(2):702–705, 1990.

[10] A. Bossavit. On the representation of differential forms by potentials in dimension 3. In *Scientific computing in electrical engineering (Warnemünde, 2000)*, volume 18 of *Lect. Notes Comput. Sci. Eng.*, pages 97–104. Springer, Berlin, 2001.

[11] M. A. Botchev and J. G. Verwer. Numerical integration of damped Maxwell equations. *SIAM J. Sci. Comput.*, 31(2):1322–1346, 2009.

[12] F. Brezzi, G. Manzini, D. Marini, P. Pietra, and A. Russo. Discontinuous finite elements for diffusion problems. In *Atti Convegno in onore di F. Brioschi (Milan, 1997)*, pages 197–217. Istitutto Lombardo, Accademia di Scienza e Lettere, Milan, Italy, 1999.

[13] A. Buffa, P. Houston, and I. Perugia. Discontinuous Galerkin computation of the Maxwell eigenvalues on simplicial meshes. *J. Comput. Appl. Math.*, 204(2):317–333, 2007.

[14] A. Buffa and I. Perugia. Discontinuous Galerkin approximation of the Maxwell eigenproblem. *SIAM J. Numer. Anal.*, 44(5):2198–2226, 2006.

[15] A. Buffa, I. Perugia, and T. Warburton. The mortar-discontinuous Galerkin method for the 2D Maxwell eigenproblem. *J. Sci. Comput.*, 40(1-3):86–114, 2009. Available at `http://www.springerlink.com/content/85055g4h55622j19/`.

[16] M. H. Carpenter and C. A. Kennedy. Fourth order 2N-storage Runge-Kutta scheme, NASA-TM-109112 (NASA Langley Research Center, VA, 1994).

[17] M.-H. Chen, B. Cockburn, and F. Reitich. High-order RKDG methods for computational electromagnetics. *J. Sci. Comput.*, 22/23:205–226, 2005.

[18] Q. Chen and I. Babuška. Approximate optimal points for polynomial interpolation of real functions in an interval and in a triangle. *Comput. Methods Appl. Mech. Engrg.*, 128(3-4):405–417, 1995.

[19] Q. Chen and I. Babuška. The optimal symmetrical points for polynomial interpolation of real functions in the tetrahedron. *Comput. Methods Appl. Mech. Engrg.*, 137(1):89–94, 1996.

[20] B. Cockburn, G. E. Karniadakis, and C.-W. Shu. The development of discontinuous Galerkin methods. In *Discontinuous Galerkin methods (Newport, RI, 1999)*, volume 11 of *Lect. Notes Comput. Sci. Eng.*, pages 3–50. Springer, Berlin, 2000.

[21] B. Cockburn, F. Li, and C.-W. Shu. Locally divergence-free discontinuous Galerkin methods for the Maxwell equations. *J. Comput. Phys.*, 194(2):588–610, 2004.

[22] B. Cockburn and C.-W. Shu. Runge-Kutta discontinuous Galerkin methods for convection-dominated problems. *J. Sci. Comput.*, 16(3):173–261, 2001.

[23] L. Demkowicz, J. Kurtz, D. Pardo, M. Paszyński, W. Rachowicz, and A. Zdunek. *Computing with hp-adaptive finite elements. Vol. 2.* Chapman & Hall/CRC Applied Mathematics and Nonlinear Science Series. Chapman & Hall/CRC, Boca Raton, FL, 2008. Frontiers: three dimensional elliptic and Maxwell problems with applications.

[24] J. Diaz and M. J. Grote. Energy conserving explicit local time stepping for second-order wave equations. *SIAM Journal on Scientific Computing*, 31(3):1985–2014, 2009.

[25] Y. Epshteyn and B. Rivière. Estimation of penalty parameters for symmetric interior penalty Galerkin methods. *J. Comput. Appl. Math.*, 206(2):843–872, 2007.

[26] A. Ern and J.-L. Guermond. *Theory and practice of finite elements*, volume 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004.

[27] A. Fisher, R. N. Rieben, G. H. Rodrigue, and D. A. White. A generalized mass lumping technique for vector finite-element solutions of the time-dependent Maxwell equations. *IEEE Trans. Antennas and Propagation*, 53(9):2900–2910, 2005.

[28] 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.

[29] S. Gottlieb and S. J. Ruuth. Optimal strong-stability-preserving time-stepping schemes with fast downwind spatial discretizations. *J. Sci. Comput.*, 27(1-3):289–303, 2006.

[30] S. Gottlieb and C.-W. Shu. Total variation diminishing Runge-Kutta schemes. *Math. Comp.*, 67(221):73–85, 1998.

[31] S. Gottlieb, C.-W. Shu, and E. Tadmor. Strong stability-preserving high-order time discretization methods. *SIAM Rev.*, 43(1):89–112, 2001.

[32] D. J. Griffiths. *Introduction to Electrodynamics*. Benjamin Cummings; 3 edition, 1999.

[33] M. J. Grote, A. Schneebeli, and D. Schötzau. Interior penalty discontinuous Galerkin method for Maxwell's equations: energy norm error estimates. *J. Comput. Appl. Math.*, 204(2):375–386, 2007.

[34] M. J. Grote, A. Schneebeli, and D. Schötzau. Interior penalty discontinuous Galerkin method for Maxwell's equations: optimal L^2-norm error estimates. *IMA J. Numer. Anal.*, 28(3):440–468, 2008.

[35] M. J. Grote and D. Schötzau. Optimal error estimates for the fully discrete interior penalty DG method for the wave equation. *J. Sci. Comput.*, 40(1-3):257–272, 2009.

[36] B. Gustafsson, H.-O. Kreiss, and J. Oliger. *Time dependent problems and difference methods*. Pure and Applied Mathematics (New York). John Wiley & Sons Inc., New York, 1995. A Wiley-Interscience Publication.

[37] E. Hairer, C. Lubich, and G. Wanner. *Geometric numerical integration*, volume 31 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2002. Structure-preserving algorithms for ordinary differential equations.

[38] B. He and F. L. Teixeira. Differential forms, Galerkin duality, and sparse inverse approximations in finite element solutions of Maxwell equations. *IEEE Trans. Antennas and Propagation*, 55(5):1359–1368, 2007.

[39] J. S. Hesthaven. From electrostatics to almost optimal nodal sets for polynomial interpolation in a simplex. *SIAM J. Numer. Anal.*, 35(2):655–676, 1998.

[40] J. S. Hesthaven. High-order accurate methods in time-domain computational electromagnetics. A review. *Advances in Imaging and Electron Physics*, 127(1):59–123, 2003.

[41] J. S. Hesthaven and C. H. Teng. Stable spectral methods on tetrahedral elements. *SIAM J. Sci. Comput.*, 21(6):2352–2380, 2000.

[42] J. S. Hesthaven and T. Warburton. Nodal high-order methods on unstructured grids. I. Time-domain solution of Maxwell's equations. *J. Comput. Phys.*, 181(1):186–221, 2002.

[43] J. S. Hesthaven and T. Warburton. High-order nodal discontinuous Galerkin methods for the Maxwell eigenvalue problem. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 362(1816):493–524, 2004.

[44] J. S. Hesthaven and T. Warburton. *Nodal discontinuous Galerkin methods*, volume 54 of *Texts in Applied Mathematics*. Springer, New York, 2008. Algorithms, analysis, and applications.

[45] R. Hiptmair. Finite elements in computational electromagnetism. *Acta Numer.*, 11:237–339, 2002.

[46] P. Houston, I. Perugia, A. Schneebeli, and D. Schötzau. Interior penalty method for the indefinite time-harmonic Maxwell equations. *Numer. Math.*, 100(3):485–518, 2005.

[47] P. Houston, I. Perugia, and D. Schötzau. Mixed discontinuous Galerkin approximation of the Maxwell operator. *SIAM J. Numer. Anal.*, 42(1):434–459, 2004.

[48] P. Houston, I. Perugia, and D. Schötzau. Mixed discontinuous Galerkin approximation of the Maxwell operator: non-stabilized formulation. *J. Sci. Comput.*, 22/23:315–346, 2005.

[49] F. Q. Hu and H. L. Atkins. Eigensolution analysis of the discontinuous Galerkin method with nonuniform grids. I. One space dimension. *J. Comput. Phys.*, 182(2):516–545, 2002.

[50] F. Q. Hu, M. Y. Hussaini, and P. Rasetarinera. An analysis of the discontinuous Galerkin method for wave propagation problems. *J. Comput. Phys.*, 151(2):921–946, 1999.

[51] J. Jin. *The finite element method in electromagnetics*. Wiley-Interscience [John Wiley & Sons], New York, second edition, 2002.

[52] G. E. Karniadakis and S. J. Sherwin. *Spectral/hp element methods for computational fluid dynamics*. Numerical Mathematics and Scientific Computation. Oxford University Press, New York, second edition, 2005.

[53] R. J. LeVeque. *Finite volume methods for hyperbolic problems*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 2002.

[54] R. I. McLachlan. On the numerical integration of ordinary differential equations by symmetric composition methods. *SIAM J. Sci. Comput.*, 16(1):151–168, 1995.

[55] A. H. Mohammadian, V. Shankar, and W. F. Hall. Computation of electromagnetic scattering and radiation using a time-domain finite-volume discretization procedure. *Computer Physics Communications*, 68:175–196, Nov. 1991.

[56] P. Monk. *Finite element methods for Maxwell's equations*. Numerical Mathematics and Scientific Computation. Oxford University Press, New York, 2003.

[57] P. Monk and G. R. Richter. A discontinuous Galerkin method for linear symmetric hyperbolic systems in inhomogeneous media. *J. Sci. Comput.*, 22/23:443–477, 2005.

[58] J.-C. Nédélec. Mixed finite elements in $\mathbf{R}^3$. *Numer. Math.*, 35(3):315–341, 1980.

[59] J.-C. Nédélec. A new family of mixed finite elements in $\mathbf{R}^3$. *Numer. Math.*, 50(1):57–81, 1986.

[60] I. Perugia and D. Schötzau. The *hp*-local discontinuous Galerkin method for low-frequency time-harmonic Maxwell equations. *Math. Comp.*, 72(243):1179–1214, 2003.

[61] I. Perugia, D. Schötzau, and P. Monk. Stabilized interior penalty methods for the time-harmonic Maxwell equations. *Comput. Methods Appl. Mech. Engrg.*, 191(41-42):4675–4697, 2002.

[62] L. Pesch, A. Bell, H. Sollie, V. R. Ambati, O. Bokhove, and J. J. W. Van Der Vegt. hpGEM—a software framework for discontinuous Galerkin finite element methods. *ACM Trans. Math. Software*, 33(4):Art. 23, 25, 2007.

[63] W. Reed and T. Hill. Triangular mesh methods for the neutron transport equation. 1973.

[64] 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.

[65] G. Rodrigue and D. White. A vector finite element time-domain method for solving Maxwell's equations on unstructured hexahedral grids. *SIAM J. Sci. Comput.*, 23(3):683–706, 2001.

[66] S. J. Ruuth. Global optimization of explicit strong-stability-preserving Runge-Kutta methods. *Math. Comp.*, 75(253):183–207, 2006.

[67] J. M. Sanz-Serna and M. P. Calvo. *Numerical Hamiltonian problems*, volume 7 of *Applied Mathematics and Mathematical Computation*. Chapman & Hall, London, 1994.

[68] D. Sármány, F. Izsák, and J. J. W. van der Vegt. High-order accurate discontinuous Galerkin method for the indefinite time-harmonic Maxwell equations. Memorandum 1889, Department of Applied Mathematics, University of Twente, Enschede, the Netherlands, January 2009. Available at `http://eprints.eemcs.utwente.nl/14852/01/memo1889.pdf`.

[69] D. Schötzau, C. Schwab, and A. Toselli. Stabilized *hp*-DGFEM for incompressible flow. *Math. Models Methods Appl. Sci.*, 13(10):1413–1436, 2003.

[70] K. Shahbazi. An explicit expression for the penalty parameter of the interior penalty method. *J. Comput. Phys.*, 205(2):401–407, 2005.

[71] S. Sherwin. Dispersion analysis of the continuous and discontinuous Galerkin formulations. In *Discontinuous Galerkin methods (Newport, RI, 1999)*, volume 11 of *Lect. Notes Comput. Sci. Eng.*, pages 425–431. Springer, Berlin, 2000.

[72] G. L. G. Sleijpen and H. A. van der Vorst. A Jacobi-Davidson iteration method for linear eigenvalue problems. *SIAM J. Matrix Anal. Appl.*, 17(2):401–425, 1996.

[73] G. L. G. Sleijpen and H. A. van der Vorst. A Jacobi-Davidson iteration method for linear eigenvalue problems. *SIAM Rev.*, 42(2):267–293, 2000.

[74] P. Šolín, K. Segeth, and I. Doležel. *Higher-order finite element methods*. Studies in Advanced Mathematics. Chapman & Hall/CRC, Boca Raton, FL, 2004.

[75] R. J. Spiteri and S. J. Ruuth. A new class of optimal high-order strong-stability-preserving time discretization methods. *SIAM J. Numer. Anal.*, 40(2):469–491, 2002.

[76] G. Strang and G. J. Fix. *An analysis of the finite element method*. Prentice-Hall Inc., Englewood Cliffs, N. J., 1973. Prentice-Hall Series in Automatic Computation.

[77] A. Taflove and S. C. Hagness. *Computational electrodynamics: the finite-difference time-domain method*. Artech House Inc., Boston, MA, second edition, 2000.

[78] M. A. Taylor, B. A. Wingate, and R. E. Vincent. An algorithm for computing Fekete points in the triangle. *SIAM J. Numer. Anal.*, 38(5):1707–1720, 2000.

[79] E. F. Toro. *Riemann solvers and numerical methods for fluid dynamics*. Springer-Verlag, Berlin, second edition, 1999. A practical introduction.

[80] J. J. W. van der Vegt and S. K. Tomar. Discontinuous Galerkin method for linear free-surface gravity waves. *J. Sci. Comput.*, 22/23:531–567, 2005.

[81] T. Warburton and M. Embree. The role of the penalty in the local discontinuous Galerkin method for Maxwell's eigenvalue problem. *Comput. Methods Appl. Mech. Engrg.*, 195(25-28):3205–3223, 2006.

[82] T. Warburton and J. S. Hesthaven. On the constants in *hp*-finite element trace inverse inequalities. *Comput. Methods Appl. Mech. Engrg.*, 192(25):2765–2773, 2003.

[83] H. Whitney. *Geometric integration theory*. Princeton University Press, Princeton, N. J., 1957.

[84] J. H. Williamson. Low-storage Runge-Kutta schemes. *J. Comput. Phys.*, 35(1):48–56, 1980.

[85] Z. Ye, L. Du, Z. Fan, and R. Chen. Mass lumping techniques combined with 3D time-domain finite-element method for the vector wave equation. In *International Conference on Microwave and Millimeter Wave Technology, 2008*, volume 3, pages 1307–1310, 2008.

[86] K. S. Yee. Numerical solution of initial boundary value problems involving Maxwell's equation in isotropic media. *IEEE Trans. Antennas Prop.*, 14(3):302–307, 1966.

[87] O. C. Zienkiewicz and R. L. Taylor. *The finite element method. Vol. 1*. Butterworth-Heinemann, Oxford, fifth edition, 2000. The basis.

[88] O. C. Zienkiewicz and R. L. Taylor. *The finite element method. Vol. 2*. Butterworth-Heinemann, Oxford, fifth edition, 2000. Solid mechanics.

[89] O. C. Zienkiewicz and R. L. Taylor. *The finite element method. Vol. 3.* Butterworth-Heinemann, Oxford, fifth edition, 2000. Fluid dynamics.

SUMMARY

This thesis discusses numerical approximations of electromagnetic wave propagation, which is mathematically described by the Maxwell equations. These equations are typically either formulated as integral equations or as (partial) differential equations. Throughout this thesis, the numerical discretisation (i.e. approximation) of the partial differential equations is considered. More specifically, out of the numerous existing discretisation techniques this work focuses on $H(\mathrm{curl})$-conforming high-order finite element methods (FEM) and high-order discontinuous finite element methods (DG-FEM) for the Maxwell equations.

One of the first, and most obvious, questions in designing a high-order FEM and DG-FEM is the choice of basis functions. The Maxwell equations have a special geometric structure. If that is not well-represented by the basis functions, the numerical approximation may lead to spurious, non-physical solutions. Another important feature of a high-order basis is hierarchy. A hierarchic construction makes it easier to use different orders of approximation in different parts of the computational domain. The discussion of a set of basis functions that both preserve the geometric structure of the Maxwell equations – that is $H(\mathrm{curl})$-conformity for the formulations used in this thesis – and have a hierarchic structure forms part of the work presented here. In particular, Chapter 3 addresses the major difficulties with the implementation of the basis, especially with ensuring the $H(\mathrm{curl})$-conforming property of every basis function. In Chapter 5, the high-order hierarchic $H(\mathrm{curl})$-conforming FEM is then successfully applied to the time-dependent Maxwell wave equation in three spatial dimensions.

It is possible to introduce additional flexibility in the FEM by allowing the discrete representation of the solution to be discontinuous. In the resulting DG-FEM it is less important to mimic the geometric structure of the original equations in the definition of the basis functions. Rather, the job is then done by the definition of the numerical fluxes – the functions that are responsible for coupling the information between the discontinuous elements. As a consequence, the use of nodal basis functions, which are neither hierarchic nor $H(\mathrm{curl})$-conforming, have proved highly

popular, largely because it allows for a very efficient implementation. In Chapter 2, the nodal approach is applied to the first-order time-dependent Maxwell system in one and two dimensions. Having a hierarchic structure of the basis functions is, however, still advantageous in the DG-FEM when using varying-order approximations. In Chapter 4, the $H(\text{curl})$-conforming basis is implemented in the DG-FEM framework for the second-order time-harmonic Maxwell wave equation in three dimensions. An optimal estimate of the penalty parameter in the numerical flux is derived, which is especially useful in providing guidance on how and where to strike the balance between stability and computational efficiency.

High-order discretisation in space generally requires high-order discretisation in time. One family of the most widely-used high-order time-integration methods applied in combination with DG-FEM are the strong-stability-preserving Runge-Kutta (SSPRK) methods. In Chapter 2, the SSPRK method is implemented in a way that its order of accuracy matches that of the space discretisation. The resulting fully-discrete scheme is shown to be computationally more efficient than many of the alternatives using a fixed-order time-integration method. However, if the computational domain contains conductive materials, as is often the case, explicit RK methods, such as SSPRK, are no longer suitable. In that case, a time-integration method is needed that treats the conductivity term in an implicit manner. Chapter 5 describes such time-integration methods when they are applied to both $H(\text{curl})$-conforming FEM and DG-FEM semi-discrete schemes of the second-order time-dependent Maxwell wave equation in three-dimensions.

All the different forms of the Maxwell equations presented in this thesis describe wave propagation. It is therefore natural to ask how and whether a given numerical scheme affects the basic properties of the wave, such as dispersion and dissipation. In Chapter 2, the SSPRK-DG method is shown to be both dispersive and dissipative, albeit the dispersion and dissipation errors are often too small to matter much in many practical applications. By contrast, some of the methods described in Chapter 5 conserve the discrete energy, i.e. they are non-dissipative. They still induce numerical dispersion but, as in the case of SSPRK-DG, this is generally less of a worry if the order of the scheme is increased.