Springer Series in Computational Mathematics

43

For further volumes:
http://www.springer.com/series/797

Jichun Li • Yunqing Huang

Time-Domain Finite Element Methods for Maxwell's Equations in Metamaterials

Springer

Jichun Li
Mathematical Sciences
University of Nevada Las Vegas
Las Vegas
Nevada
USA

Yunqing Huang
Xiangtan University
Xiangtan
Hunan
China, People's Republic

ISSN 0179-3632
ISBN 978-3-642-33788-8 ISBN 978-3-642-33789-5 (eBook)
DOI 10.1007/978-3-642-33789-5
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012953285

Mathematics Subject Classification (2010): 65N30, 65L05, 65N15, 65F10, 35L15, 78M10, 78M05

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Electromagnetic metamaterials are artificially structured composite materials that exhibit a frequency band where the effective index of refraction becomes negative. Since the successful construction of such metamaterials in 2000, the study of metamaterials has attracted great attention of researchers across many disciplines. There is currently an enormous effort in the electrical engineering, material science, physics, and optics communities to come up with various ways of constructing efficient metamaterials and using them for potentially revolutionary applications in antenna and radar design, subwavelength imaging, and invisibility cloak design. Hence, simulation of electromagnetic phenomena in metamaterials becomes a very important issue, which is the subject of this book. In the mathematics community, there is an increasing interest in the study of metamaterials as evidenced by the Hot Topics Workshops on Negative Index Materials held at IMA (Institute for Mathematics and its Applications) of the University of Minnesota during October 2–4, 2006, which was the first public exposure of this subject to the mathematics community. During January 25–29, 2010, the leading author (Jichun Li) cochaired a workshop on "Metamaterials: Applications, Analysis and Modeling" at IPAM (Institute for Pure and Applied Mathematics) of the University of California at Los Angeles to expose this subject once more to the general mathematics community.

The purpose of this book is to provide a detailed introduction to the basic mathematical analysis of those model equations resulting from metamaterial simulations. We focus on developing and analyzing time-domain finite element methods for solving those metamaterial model equations. The book is intended to be self-contained in terms of finite element methods. Though there are many other types of numerical methods developed for metamaterial simulations, we restrict the contents to finite element methods because of our own research interests and experiences. The book starts with a brief introduction to metamaterials in Chap. 1. Here we discuss the origins of metamaterials, their basic electromagnetic and optical properties, some metamaterial structures and potential applications in subwavelength imaging, antenna design, invisibility cloak, and biosensing. At the end of this chapter, we introduce the governing equations for modeling wave propagation in metamaterials.

In Chap. 2, we provide a self-contained introduction to finite element methods. We start with the basic Lagrange finite elements and the corresponding interpolation error estimates. Then we present the basic finite element error analysis techniques for the second-order elliptic problems and teach readers how to code a simple Q1 element for solving elliptic problems.

After the preparatory work of Chap. 2, we move on to introduce the divergence-conforming and curl-conforming finite elements in Chap. 3. Since these elements play very important roles in metamaterial simulations, detailed constructions of these elements and their interpolation error estimates are discussed. After these, we present both explicit and implicit schemes for solving the Drude metamaterial model. The stability and error estimate analysis are carried out for those schemes. Finally, we extend similar schemes and analysis developed for the Drude model to the Lorentz model, and the Drude-Lorentz model, which are popular metamaterial models used by physicists and engineers.

In Chap. 4, we introduce the discontinuous Galerkin method and present its application to metamaterial simulations. Here, three types of discontinuous Galerkin methods are presented: one for integro-differential vector wave equations; and the other two for metamaterial Maxwell's equations written in conservation laws. MATLAB codes are provided for the practical implementation.

From our computational experiments with the lowest-order rectangular and cubic edge elements, we found that at element centers, these edge elements achieve one order higher convergence rate than the theoretical analysis suggested. This is a new superconvergence phenomenon; hence we devote Chap. 5 to the analysis of this phenomenon. The results and proofs are original, since no other books cover such superconvergence results in the infinity norm.

To develop an efficient adaptive finite element method, a posterior error estimator plays a very important role. There are several books covering this topic, but they mainly focus on classic elliptic and parabolic equations. To fill the gap, in Chap. 6 we venture to introduce some basic techniques recently developed for a posterior error analysis of Maxwell's equations. Here we first present detailed derivations of a posterior error estimator for the standard time-harmonic Maxwell's equations, then extend the analysis to the time-dependent integro-differential Maxwell's equations in cold plasma.

In Chap. 7, we present a detailed discussion on how to code the two-dimensional edge element for solving metamaterial Maxwell's equations. Considering that programming edge element is difficult and no other book has a detailed discussion on this task, we cover the whole programming process including mesh generation, calculation of the element matrices, assembly process, and postprocessing of numerical solutions. The complete MATLAB source codes are provided in the hope that the readers can easily modify our codes to solve other similar models interesting to them.

In order to model practical wave propagation problems in unbounded domains, we feel that readers have to understand how to construct the Perfectly Matched Layers (PMLs). In Chap. 8, we provide a succinct discussion of PMLs developed for free space, lossy media, dispersive media, and metamaterials.

In the last chapter (Chap. 9) of this book, we present several interesting simulations of wave propagation in metamaterials. Here we demonstrate the negative refraction index phenomenon (i.e., backward wave propagation inside metamaterials), invisibility cloak in both frequency domain and time domain, and solar cell designs with metamaterials. Finally, we mention some open issues which need more attention or have not been well studied.

Overall, this book is intended to bring readers to the front field of metamaterial simulations by finite element methods. Inevitably, there are some interesting topics left out of this book, since there is a tremendous effort going on in this area and it is hard for us to keep abreast of the vast amount of literature across many disciplines. The contents are a reflection of our own interests and related subjects. Part of the material has been given as a series of lectures by Jichun Li at Xiangtan University of China in December 2010, in the 2011 Winter Enrichment Program at King Abdullah University of Science and Technology (KAUST) of Saudi Arabia in January 2011, and at Peking University of China in August 2012. Hence, the book can also be used as a one-semester course for graduate students in physics, engineering, material sciences, optics, and mathematics interested in wave propagation simulations. We assume that all potential readers should have some basic knowledge about electromagnetic theory, partial differential equations, functional analysis, and have some training in numerical methods for solving differential equations.

Thanks are due to our family's kind love and support, without which we would not have finished this book. Special thanks go to Wei Yang, one of our talented students, who helped us create many figures for the book. We are grateful to Global Science Press for giving permission to reproduce some material and figures from our published papers in *Advances in Applied Mathematics and Mechanics*. We also benefited from David, Jichun Li's high school son, who spent a great amount of time polishing our English.

In closing, Jichun Li is especially grateful for Bairen Professorship support from Xiangtan University, which provided a very pleasant environment for writing this book. He also wants to thank the support from Mathworks Book Program provided by *mathworks.com*. Last, but by no means least, we like to thank National Science Foundation of both China and USA for grant support which has made our research in this area possible.

Las Vegas, NV, USA — Jichun Li
Xiangtan, Hunan, China — Yunqing Huang

Contents

Chapter 1
Introduction to Metamaterials

In this chapter, we start with a brief discussion on the origins of metamaterials, and their basic electromagnetic and optical properties. We then present some metamaterial structures and potential applications in areas such as sub-wavelength imaging, antenna design, invisibility cloak, and biosensing. After all these, we then move to the related mathematical problems by introducing the governing equations used to model the wave propagation in metamaterials. Finally, a brief overview of some popular computational methods for solving Maxwell's equations is provided.

1.1 The Concept of Metamaterials

The prefix "meta" means "beyond," and in this sense the terminology "metamaterials" implies artificially structured composite materials consisting of unit cells much smaller than the wavelength of the incident radiation and displaying properties not usually found in natural materials. More specifically, we are interested in a metamaterial with simultaneously negative electric **permittivity** ϵ and magnetic **permeability** μ. In general, both permittivity ϵ and permeability μ depend on the molecular and perhaps crystalline structure of the material, as well as bulk properties such as density and temperature.

Back in 1968, Russian physicist Victor Veselago wrote a seminar paper [288] on metamaterials (he then called left-handed materials). In that paper, he speculated that the strikingly unusual phenomena could be expected in a hypothetical left-handed material in which the electric field **E**, the magnetic field **H** and the wave vector **k** form a left-handed system. The paper explicitly presented that to achieve such a left-handed material, the required material parameters should be simultaneously negative for both permittivity and permeability. However, due to the non-existence of such materials in nature, Veselago's paper did not make a big impact until the first successful construction of such a medium by Smith et al. in 2000 [271], and the first experimental demonstration of the negative refractive

J. Li and Y. Huang, *Time-Domain Finite Element Methods for Maxwell's Equations in Metamaterials*, Springer Series in Computational Mathematics 43,
DOI 10.1007/978-3-642-33789-5_1, © Springer-Verlag Berlin Heidelberg 2013

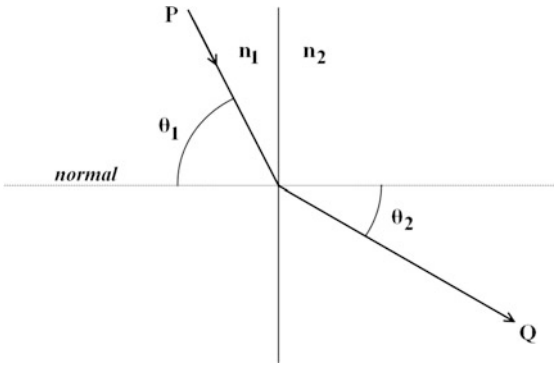

Fig. 1.1 Demonstration of Snell's law

index in 2001 [260]. Another catalyst was caused by Pendry's landmark work on perfect lens [234], which sparked the attempt to consider metamaterials for many potentially exciting applications. According to [274, p. 317], these four seminar papers together made the birth of the subject of metamaterials. Since 2000, there has been a tremendous growing interest in the study of metamaterials and their potential applications in areas ranging from electronics, telecommunications to sensing, radar technology, sub-wavelength imaging, data storage, and design of invisiblity cloak.

1.1.1 Basic Electromagnetic and Optical Properties

The optical properties of many materials can be characterized by the so-called **refractive index** (or index of refraction) n, which is defined as

$$n = \frac{c}{v}, \tag{1.1}$$

where c and v denote the speeds of light in vacuum and in the underlying material, respectively. This definition represents the optical density of the underlying medium. Hence for a normal medium, the number n is typically greater than one.

The refractive index n is often seen in the Snell's law (see Fig. 1.1):

$$n_1 \sin\theta_1 = n_2 \sin\theta_2, \tag{1.2}$$

which states that the ratio of the sines of the angles of incidence and refraction is equivalent to the reciprocal ratio of the refraction indices in two different isotropic media. Here θ_1 and θ_2 denote the incidence angle and refraction angle, respectively.

The refractive index n can also be defined using the well-known Maxwell relation

$$n = \sqrt{\epsilon_r \mu_r}. \tag{1.3}$$

This relation connects the refractive index n, an optical quantity, with two electromagnetic quantities: the permittivity ϵ_r and permeability μ_r of a medium relative to the permittivity ϵ_0 and permeability μ_0 in vacuum. Note that $\epsilon_0 = \epsilon/\epsilon_r = 8.854{\cdot}10^{-12}\,\mathrm{N/A^2}$ and $\mu_0 = \mu/\mu_r = 4\pi{\cdot}10^{-7}$ force/m. It is know that vacuum has a refractive index of 1, and the speed of light in vacuum $c = 1/\sqrt{\epsilon_0\mu_0} \approx 3\times10^8$ m/s.

One important concept in study of wave propagation problems is **phase velocity**, which is the rate at which the phase of the wave propagates in space. This is the speed at which the phase of any one frequency component of the wave travels. Mathematically, the phase velocity v_p is defined as the ratio of the wavelength λ (the distance between any two points with the same phase, such as between crests, or troughs) to period T (measured in seconds), i.e.,

$$v_p = \frac{\lambda}{T},$$

which can also be represented as

$$v_p = \frac{\omega}{k}, \tag{1.4}$$

where $\omega \equiv \frac{2\pi}{T}$ is the wave's angular frequency (measured in radians per second), and $k \equiv \frac{2\pi}{\lambda}$ is the angular wavenumber.

Another important concept in wave propagation is **group velocity**, which is used to describe the velocity with which the overall shape of the wave's amplitudes (known as the modulation or envelope of the wave) propagates through space. Mathematically the group velocity v_g is defined as

$$v_g = \frac{\partial\omega}{\partial k}. \tag{1.5}$$

The function $\omega = \omega(k)$ is known as the dispersion relation. If ω is directly proportional to k, then the group velocity is exactly equal to the phase velocity. Otherwise, the group velocity will behave very differently from the phase velocity. For example, in a dispersive medium (in which the phase velocity of a wave depends on frequency), the envelope of the wave packet become distorted as the wave propagates, since waves with different frequencies move at different speeds.

From (1.4) and (1.5), we can relate the group velocity to the phase velocity as follows:

$$\frac{1}{v_g} = \frac{1}{v_p} + \omega\frac{\partial}{\partial\omega}\left(\frac{1}{v_p}\right). \tag{1.6}$$

In the normal dispersion case, $\frac{\partial}{\partial\omega}(\frac{1}{v_p}) > 0$ implies that $v_g < v_p$. Under this situation, the group velocity is often thought of as the velocity at which energy or information is conveyed along a wave. However, if the wave travels through

an absorptive medium, this does not always hold. For example, in the anomalous dispersion, $\frac{\partial}{\partial\omega}(\frac{1}{v_p}) < 0$ implies that $v_g > v_p$. A real application of this fact is that laser light pulses are sent through specially prepared materials in order to have the group velocity significantly exceed the speed of light in vacuum. It is also possible to reduce the group velocity to zero, which makes the pulse immobile; or to have a negative group velocity, which makes the pulse appear to propagate backwards. In these cases, the group velocity loses its usual meaning as the transfer velocity of energy or information.

For isotropic double negative metamaterials, Veselago [288] showed that the phase velocity would be antiparallel to the direction of the energy flow, which is contrary to wave propagation in natural materials. This fact can be justified as follows. Let us denote the Poynting vector

$$\mathbf{S} = \frac{1}{2}\mathrm{Re}(\mathbf{E} \times \mathbf{H}^*),$$

where the star denotes complex conjugate. The Poynting vector $\mathbf{S}$ gives the magnitude and direction of power flow. Assume a plane wave propagating in a medium as

$$\mathbf{E} = \tilde{\mathbf{E}} e^{j(\omega t - \mathbf{k}\cdot\mathbf{r})}, \quad \mathbf{H} = \tilde{\mathbf{H}} e^{j(\omega t - \mathbf{k}\cdot\mathbf{r})}, \tag{1.7}$$

where $j = \sqrt{-1}$ is the imaginary unit, then substituting (1.7) into Maxwell's equations (1.9) and (1.10) given below in Sect. 1.2 with the constitutive relations (1.11), we have

$$\epsilon\omega\tilde{\mathbf{E}} = -\mathbf{k} \times \tilde{\mathbf{H}}, \quad \mu\omega\tilde{\mathbf{H}} = \mathbf{k} \times \tilde{\mathbf{E}}, \tag{1.8}$$

which shows that: If ϵ and $\mu > 0$, then vectors $\tilde{\mathbf{E}}$, $\tilde{\mathbf{H}}$ and $\mathbf{k}$ obey the right-hand rule; If ϵ and $\mu < 0$, then vectors $\tilde{\mathbf{E}}$, $\tilde{\mathbf{H}}$ and $\mathbf{k}$ obey the left-hand rule, i.e., $\mathbf{S}$ and $\mathbf{k}$ have opposite directions.

Hence, if we assume that the energy flux moving away from the source is the positive direction as usual, then the phase velocity of a propagating wave in a metamaterial points towards the source. For this reason and the definition of $n = c/v$, metamaterials could be considered as having a negative refractive index, i.e.,

$$n = -\sqrt{\epsilon_r \mu_r}, \quad \text{when } \epsilon_r < 0,\ \mu_r < 0.$$

One striking property for metamaterials is the so-called re-focusing property. Let us assume that a line source is placed $\frac{d}{2}$ before a metamaterial slab with width d and refractive index $n_r = -1$, the medium outside the slab is free space (i.e., $n_i = 1$). By Snell's law (1.2), the refraction angle θ_r is equal to the negative incidence angle θ_i. Hence all rays emanating from the line source will be refocused inside the metamaterial slab and have another focus at the back of the slab (see Fig. 1.2).

Another interesting property for metamaterials is that the Doppler effect (or Doppler shift) in metamaterials is reversed. Recall that the well-known Doppler effect tells us that: For wave propagating in a standard medium (such as sound wave in air), the wave frequency increases for an observer as the source of the wave moves

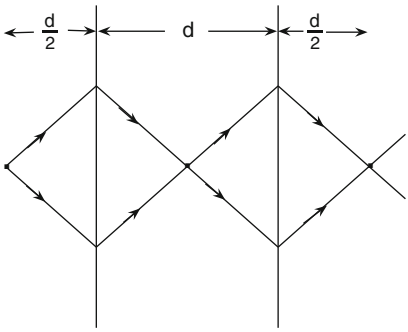

Fig. 1.2 Demonstration of the refocusing property

closer; while the wave frequency decreases for the observer as the source of the wave moves away. A simple example of Doppler effect is that when an ambulance approaches, the sound wave generated from its siren is compressed, which increases the wave frequency or pitch; when the ambulance moves away, the sound wave is stretched, which causes the siren's pitch to decrease. On the other hand, for the electromagnetic wave propagating in a metamaterial, the wave frequency decreases for an observer as the wave source moves closer. This can be very scary. Just imagine that if the air were filled with metamaterials, then a missile could reach the target without any awareness.

1.1.2 Basic Structures

The first double negative metamaterial was constructed by a group of physicists at the University of California at San Diego led by David Smith et al. [271]. The material consists of a two-dimensional array of repeated unit cells of square copper split ring resonators (SRRs) and copper wire strips on fiber glass circuit board. The SRR is made of two concentric rings separated by a gap, and both rings have splits at opposite sides, see Fig. 1.3. By careful design of the split width, gap distance, metal width and radius, the SRR can hopefully create a strong magnetic resonance which leads to negative permeability μ. While the metal wire is used to provide the negative permittivity ϵ by carefully choosing the distance between the wires and the size of their cross section. Experiments carried out by Shelby et al. [260] demonstrate that this structure shows negative refraction index and left-handed behavior for incident plane waves with electric field polarized parallel to the continuous wire and magnetic field perpendicular to the SRR.

Various modifications of SRRs have been proposed in the literature, aiming mainly to make the structure easy to fabricate, reduce the overall size of the cell element, and reduce the loss of the structure. For example, a set of split ring resonators was investigated by Aydin et al. [15]. Their constructions are shown in

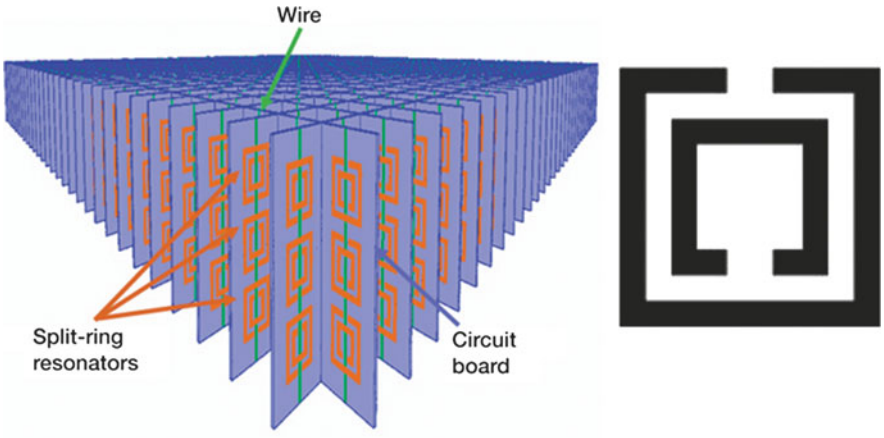

Fig. 1.3 (*Left*): An exemplary metamaterial formed by square split ring resonators (SRRs) and metal wires (Source: http://en.wikipedia.org/wiki/File:Left-handed_metamaterial_array_configuration.jpg) (Author: Cynthia.L.Dreibelbis@nasa.gov). (*Right*): A unit cell of square split ring resonators (SRRs)

Fig. 1.4 Some split ring resonators designed by Aydin et al. [15] (Reproduced with permission from Fig. 11 of [15])

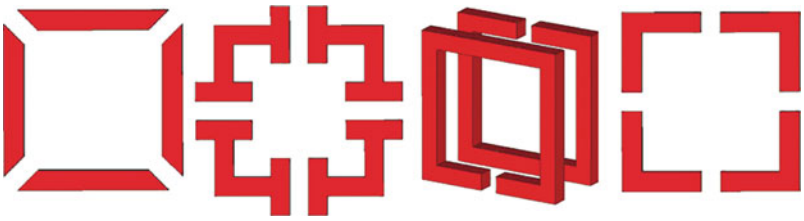

Fig. 1.5 Some split ring resonators studied by Kafesaki et al. [164] (Reproduced with permission from Fig. 16 of [164])

Fig. 1.4. The first three are single rings split one, two and four times, respectively. The fourth and fifth are double rings split four and eight times, respectively.

In 2005, Kafesaki et al. [164] carried out a comprehensive numerical study of many SRRs (see Fig. 1.5). They studied the magnetic and the electric response of single-ring and double-ring SRRs, and how the responses of SRRs depend on the length, width and depth of the metallic sides for different kinds of SRRs.

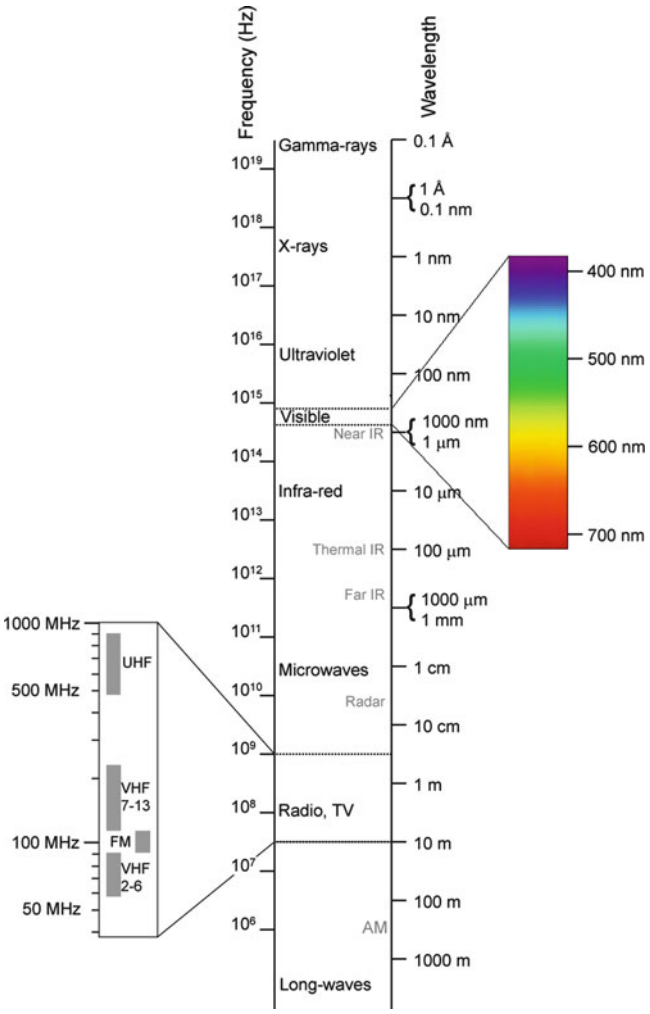

Fig. 1.6 The whole electromagnetic spectrum (Source: http://en.wikipedia.org/wiki/Electromagnetic_spectrum)

Recently, in search of higher-frequency resonators, researchers found that the resonant frequency saturates as the SRR size becomes smaller and smaller. Extension of metamaterials based on split ring resonators to near-infrared and visible wavelengths (the whole electromagnetic spectrum is shown in Fig. 1.6) becomes quite challenging and often involves difficult fabrication problem. A popular structure in optical wavelengths is a fishnet design, which consists of a metal-dielectric-metal sandwich. A square array of holes riddles the sandwich, which makes the structure similar to a real fishnet. The holes may be circular, elliptical or rectangular.

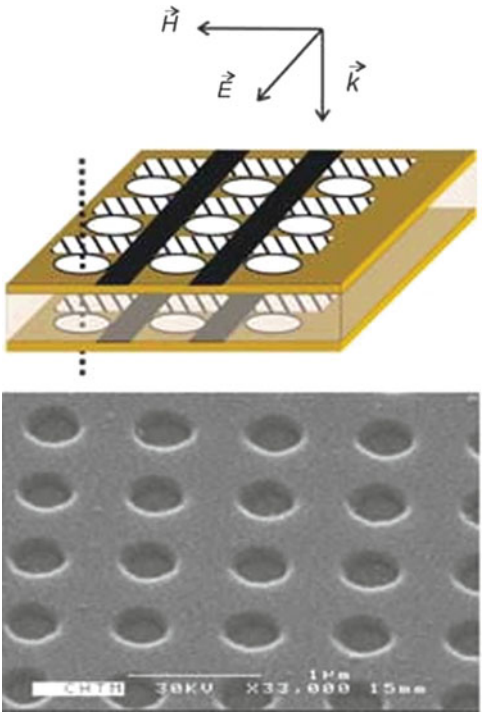

Fig. 1.7 (*Top*): A multilayer fishnet structure designed by Zhang et al. [301, Fig. 1]. (*Bottom*): The scanning electron microscopy (SEM) picture of the fabricated structure (Reprinted with permission from Zhang et al. [301]. Copyright (2005) by the American Physical Society)

Figure 1.7 shows a multilayer fishnet structure designed by Zhang et al. [301]. It consists of an Al_2O_3 dielectric layer between two Au films perforated with a square periodic array of circular holes (period 838 nm; hole diameter is about 360 nm) atop a glass substrate.

In [285], Valentine et al. experimentally demonstrated the first 3-D fishnet metamaterial (see Fig. 1.8), which is fabricated on a multilayer metal-dielectric stack. This structure consists of alternating layers of 30 nm silver (Ag) and 50 nm magnesium fluoride (MgF_2).

All the structures mentioned so far have anisotropic properties. To construct an isotropic metamaterial, the unit cell should have some symmetries. Some 3-D isotropic resonators have been proposed [124, 232, 236]. One example is shown in Fig. 1.9.

1.1.3 Potential Applications

1.1.3.1 Subwavelength Imaging

It is known that conventional lens-based imaging devices cannot provide resolution better than $\lambda/2$, where λ is the radiation wavelength. Such restriction is the

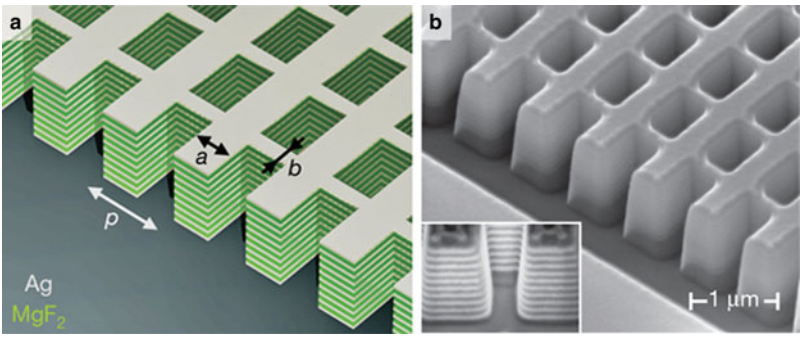

Fig. 1.8 (*Left*): The 21-layer fishnet structure designed by Valentine et al. [285, Fig. 1]. The dimensions of the unit cell are $p = 860$ nm, $a = 565$ nm (width of wide slabs) and $b = 265$ nm (width of thin slabs). The structure consists of alternating layers of 30 nm silver (Ag) and 50 nm magnesium fluoride (MgF2), (*Right*): The SEM image of the 21-layer fishnet structure with the side etched, showing the cross-section (Reprinted by permission from Macmillan Publishers Ltd: Nature [285], copyright (2008))

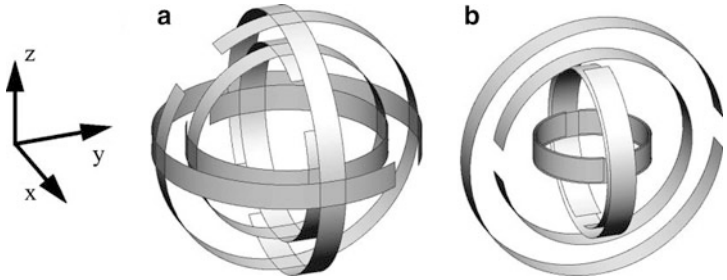

Fig. 1.9 3D isotropic resonators: Gay-Balmaz et al.'s design [124, Fig. 6]. (**a**) The structure is built from three identical SRRs normal to each other. (**b**) The structure is composed of three SRRs of increasing size (Reprinted with permission from Gay-Balmaz and Martin [124]. Copyright (2002), American Institute of Physics)

so-called diffraction limit. In recent years, several techniques based on the use of metamaterials have been proposed for subwavelength imaging in different ranges of electromagnetic spectrum. Proposed techniques include perfect lens [234], silver superlenses [116], hyperlenses [205, 221, 272], and wire medium lenses [265, 266].

For example, Silveirinha et al. [265] showed that a wire medium lens made of silver nanorods could achieve subwavelength resolution of $\lambda/10$ at 33 THz. Figure 1.10 presents the results of [265].

Another interesting example was proposed by Fang et al. [116]. An object "NANO" with 40 nm linewidth was imaged by silver superlens. The object was clearly imaged even when the incoming wave had 365 nm wavelength, which means that $\lambda/9$ image resolution was obtained.

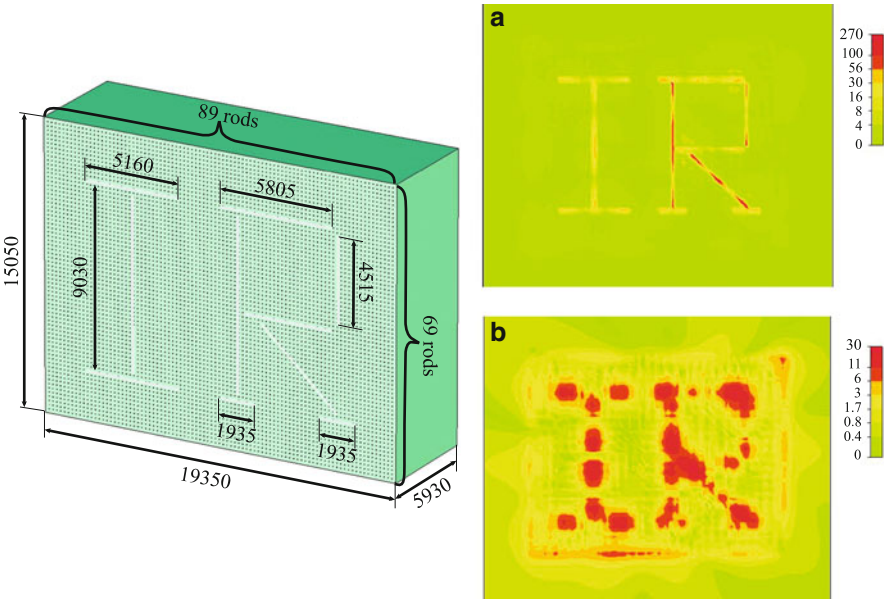

Fig. 1.10 (*Left*): The setup of the imaging simulation, where the numbers are in nm. (*Right*) Distributions of $|E_z|$ at 33 THz at the source plane (*Top*) and image plane (*Bottom*) (Source: Reprinted with permission from Silveirinha et al. [265]. Copyright (2007) by the American Physical Society)

1.1.3.2 Circuit Applications

Due to the small dimensions of SRRs and complementary split ring resonators (CSRRs, a dual of SRR by switching metal and air) relative to the signal wavelength at their resonance frequency, SRR and CSRR-based transmission lines are useful for device miniaturization. Applications in microwave passive components such as impedance inverters, power dividers [257], couplers, and filters have been discussed. SRRs are also useful particles in many other applications such as magnetoinductive and electroinductive wave components [37, 275], frequency selective surfaces [18].

1.1.3.3 Antenna Applications

Recently, researchers have proposed some methods to obtain miniaturized antennas made of ideal homogenized metamaterials. The first design of a subwavelength antenna with metamaterials for the case of dipole and monopole radiators was proposed by Ziokowski's group [312]. The basic design consists of an electrically

short electric dipole (or monopole) surrounded by a double-negative (DNG) or an epsilon-negative (ENG) spherical shell with an electrically short radius. The compact resonance arises at the interface between the DNG (or ENG) shell and the free space. Similar ideas have been used to design subwavelength patch antennas [7, 38] and leaky-wave antennas [8].

1.1.3.4 Cloaking

Invisibility has long been a dream of human beings. Cloaking devices are advanced stealth technologies still in development that can make objects partially or wholly invisible to some portions of the electromagnetic spectrum.

Generally speaking, there are several major approaches to render objects invisible. For example, Alu and Engheta [6] proposed to use plasmonic coatings to cancel the dipolar scattering. But this technique is limited to the sub-wavelength scale of the object, and the coating depends on the geometry and material parameters of the object. Milton and Nicorovici [212] discovered that using a metamaterial coating would cloak polarizable line dipoles. But the coating is affected by the objects placed inside. Leonhardt [180] and independently Pendry, Schurig and Smith [237] discovered a coordinate transformation mechanism for electromagnetic cloaking. Their mechanism was quite similar to that of Greenleaf et al. [130, 131] introduced for conductivity. Their main idea is to guide electromagnetic wave around the cloaked region, and many later work has adopted this technique.

In May 2006, the first full wave numerical simulations on cylindrical cloaking was carried out by Cummer et al. [96]. A few months later, the first experiment of such a cloak at microwave frequencies was successfully demonstrated by Schurig et al. [254], where the cloak surrounding a 25-mm-radius Cu cylinder was measured.

After 2006, numerous studies have been devoted to cloaking, mainly inspired by [96, 254]. For example, in 2008, Liang et al. [197] performed a time-dependent simulation for the cylindrical cloak using finite-difference time-domain method. Their simulation (cf. Fig. 1.11) clearly shows the dynamical process of the electromagnetic wave in the cloaking structure. Figure 1.11 is obtained by considering only the E-polarized modes with permittivity and permeability components ϵ_z, μ_r and μ_θ satisfying the Lorentzian dispersive function $f_j(\omega) = \omega_p^2/(w_{aj}^2 - \omega^2 - j\omega\gamma)$, where $j = z, r, \theta$. The setup of the cloaking system is shown in Fig. 1.11a with R_1 and $R_2 = 2R_1$ as the inner and the outer cylindrical radii of the cloaking structure. A perfect electric conductor shell is put against the inner surface of the structure. An incident plane wave with frequency ω_0 moves form left side towards the cloaking structure, which is surrounded by the free space. As we can see, the cloaking effect is built up step by step. Finally, the field gets to the stable state shown in Fig. 1.11f, which clearly shows that the plane wave pattern gets recovered after the wave passes through the cloaking structure.

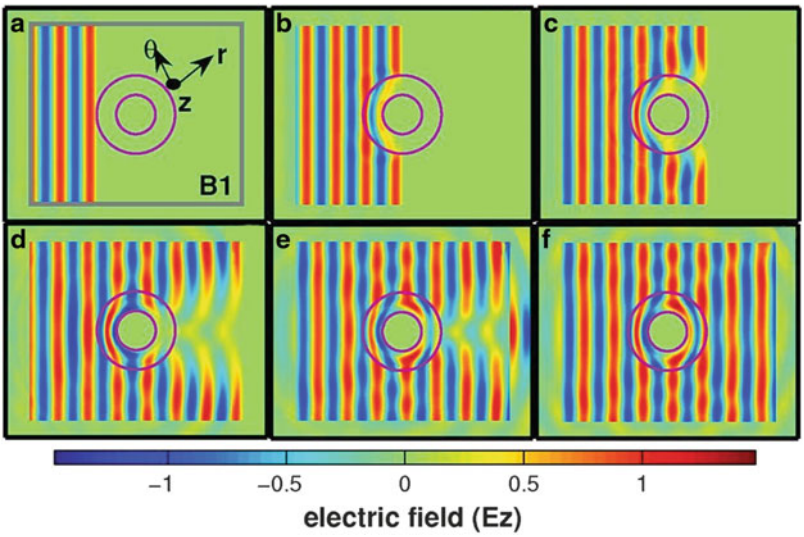

Fig. 1.11 The distribution of the electric field at different times: (**a**) $t = 2.28T$; (**b**) $t = 3.60T$; (**c**) $t = 4.92T$; (**d**) $t = 7.20T$; (**e**) $t = 9.00T$; (**f**) Stable state. T is the period of the incident wave (Reprinted with permission from Liang et al. [197]. Copyright (2008), American Institute of Physics)

1.1.3.5 Biosensing

Another potential application field of metamaterials is on biosensing. Conventional biosensors (such as those based on electro-mechanical transduction, fluorescence, nanomaterials, and surface plasmon resonance) often involve labor-intensive sample preparation and very sophisticated equipment.

In recent years, researchers have proposed to use metamaterials as candidates for detection of highly sensitive chemical, biochemical and biological analytes. For example, Lee et al. [177] studied the possibility of using split-ring resonators (SRRs) for biosensors. The basic principle is based on the fact that SRR can be considered to be a simple LC circuit with a response frequency of $f = 1/2\pi\sqrt{LC}$, which shows that the resonant frequency varies in terms of the changes in the inductance L and/or capacitance C. Hence the resonant frequency of SRR shall be shifted before and after the introduction of biomaterials.

Planar metamaterials were proposed to serve as thin-film sensors recently by O'Hara et al. [230]. They found that a resonant frequency response can be tuned through metamaterial designs. Though their metamaterial design can only detect thin films having a thickness less than 100 nm, their work presents a promising outlook for THz sensing technology.

1.1.3.6 Particle Detection

It is known that when charged particles move in a medium with velocity larger than $\frac{c}{n}$ (the phase velocity of light in the medium), Cherenkov radiation (CR) is emitted. Recall that c is the speed of light in vacuum, and n is the index of refraction of the medium. An example of CR is the blue glow seen in a nuclear reactor. Devices sensitive to Cherenkov radiation, called Cherenkov detectors, have been used extensively for detecting fast moving charged particles, and measuring the intensity of reactions etc.

Since the recently constructed metamaterials have negative refractive index, which results in the so-called reversed CR [288], a phenomenon can be used to improve the Cherenkov detectors. The reason is that in a conventional dielectric medium, the emitted radiation travels in the same direction as the particles, which will interfere with the detection of those photons. However, in metamaterials, photons and charged particles move in opposite directions so that their physical interference is reduced. Though great progress has been made in the past decade on theoretical, numerical and experimental study of reversed CR [66, 104, 123], many challenging issues need to be resolved before the reversed CR can be put in practical applications. Since the intensity of CR increases with frequency, the optical or ultraviolent spectrum is more useful for detection. However, fabrication techniques for creating low loss metamaterials at optical or ultraviolent frequencies [57] is far less mature.

1.2 Governing Equations for Metamaterials

The Maxwell's equations are the fundamental equations for understanding most electromagnetic and optical phenomena. In time domain, the general Maxwell's equations can be written as

$$\text{Faraday's law (1831):} \quad \nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \tag{1.9}$$

$$\text{Ampere's law (1820):} \quad \nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t}, \tag{1.10}$$

which are used to describe the relationship between electric field $\mathbf{E}(\mathbf{x}, t)$ and magnetic field $\mathbf{H}(\mathbf{x}, t)$, and the underlying electromagnetic materials can be described by two material parameters: the permittivity ϵ and the permeability μ. In (1.9) and (1.10), we use the electric flux density $\mathbf{D}(\mathbf{x}, t)$ and magnetic flux density $\mathbf{B}(\mathbf{x}, t)$, which are related to the fields $\mathbf{E}$ and $\mathbf{H}$ through the constitutive relations given by

$$\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P} \equiv \epsilon \mathbf{E}, \quad \mathbf{B} = \mu_0 \mathbf{H} + \mathbf{M} \equiv \mu \mathbf{H}, \tag{1.11}$$

where ϵ_0 is the vacuum permittivity, μ_0 is the vacuum permeability, and $\mathbf{P}$ and $\mathbf{M}$ are the induced polarization and magnetization, respectively. Note that $\mathbf{P}$ and $\mathbf{M}$ are

caused by the impinging fields, which can influence the organization of electrical charges and magnetic dipoles in a medium. How big the induced polarization $\mathbf{P}$ and magnetization $\mathbf{M}$ are depends on the particular material involved. For example, in vacuum, $\epsilon = \epsilon_0, \mu = \mu_0$, hence $\mathbf{P} = \mathbf{M} = 0$; while in pure water, $\epsilon = 80\epsilon_0$ and $\mu = \mu_0$, which lead to $\mathbf{P} = 79\epsilon_0\mathbf{E}$ and $\mathbf{M} = 0$.

For metamaterials, the permittivity ϵ and the permeability μ are not just simple constants due to the complicated interaction between electromagnetic fields and meta-atoms (i.e., the unit cell structure). Since the scale of inhomogeneities in a metamaterial is much smaller than the wavelength of interest, the responses of the metamaterial to external fields can be homogenized and are described using effective permittivity and effective permeability. A popular model for metamaterial is the lossy Drude model [311, 313], which in frequency domain is described by:

$$\epsilon(\omega) = \epsilon_0(1 - \frac{\omega_{pe}^2}{\omega(\omega - j\Gamma_e)}) = \epsilon_0\epsilon_r, \tag{1.12}$$

$$\mu(\omega) = \mu_0(1 - \frac{\omega_{pm}^2}{\omega(\omega - j\Gamma_m)}) = \mu_0\mu_r, \tag{1.13}$$

where ω_{pe} and ω_{pm} are the electric and magnetic plasma frequencies, Γ_e and Γ_m are the electric and magnetic damping frequencies, and ω is a general frequency. A simple case for achieving negative refraction index $n = -\sqrt{\epsilon_r\mu_r} = -1$ is to choose $\Gamma_e = \Gamma_m = 0$ and $\omega_{pe} = \omega_{pm} = \sqrt{2}\omega$.

A derivation of (1.12) is given in [235] for very thin metallic wires assembled into a periodic lattice. Assuming that the wires have radius r, and are arranged in a simple cubic lattice with distance a between wires, and σ is the conductivity of the metal, Pendry et al. [235] showed that

$$\omega_{pe}^2 = \frac{2\pi c^2}{a^2 \ln(a/r)}, \quad \Gamma_e = \frac{\epsilon_0 a^2 \omega_{pe}^2}{\pi r^2 \sigma}, \tag{1.14}$$

where c denotes the speed of light in vacuum.

Using a time-harmonic variation of $exp(j\omega t)$, from (1.11) to (1.13) we can obtain the corresponding time domain equations for the polarization $\mathbf{P}$ and the magnetization $\mathbf{M}$ as follows:

$$\frac{\partial^2\mathbf{P}}{\partial t^2} + \Gamma_e\frac{\partial\mathbf{P}}{\partial t} = \epsilon_0\omega_{pe}^2\mathbf{E}, \tag{1.15}$$

$$\frac{\partial^2\mathbf{M}}{\partial t^2} + \Gamma_m\frac{\partial\mathbf{M}}{\partial t} = \mu_0\omega_{pm}^2\mathbf{H}. \tag{1.16}$$

Furthermore, if we denote the induced electric and magnetic currents

$$\mathbf{J} = \frac{\partial\mathbf{P}}{\partial t}, \quad \mathbf{K} = \frac{\partial\mathbf{M}}{\partial t}, \tag{1.17}$$

then we can obtain the governing equations for modeling the wave propagation in a DNG medium described by the Drude model [189]:

$$\epsilon_0 \frac{\partial \mathbf{E}}{\partial t} = \nabla \times \mathbf{H} - \mathbf{J}, \tag{1.18}$$

$$\mu_0 \frac{\partial \mathbf{H}}{\partial t} = -\nabla \times \mathbf{E} - \mathbf{K}, \tag{1.19}$$

$$\frac{1}{\epsilon_0 \omega_{pe}^2} \frac{\partial \mathbf{J}}{\partial t} + \frac{\Gamma_e}{\epsilon_0 \omega_{pe}^2} \mathbf{J} = \mathbf{E}, \tag{1.20}$$

$$\frac{1}{\mu_0 \omega_{pm}^2} \frac{\partial \mathbf{K}}{\partial t} + \frac{\Gamma_m}{\mu_0 \omega_{pm}^2} \mathbf{K} = \mathbf{H}. \tag{1.21}$$

Note that the two-dimensional transverse magnetic model of [311, Eq. (10)] can be obtained directly from (1.18) to (1.21) by assuming that components $E_y, H_x, H_z \neq 0$, while the rest components are 0.

Another popular model used for modeling wave propagation in metamaterials is described by the so-called Lorentz model [259, 269, 313], which in frequency domain is given by

$$\epsilon(\omega) = \epsilon_0 \left(1 - \frac{\omega_{pe}^2}{\omega^2 - \omega_{e0}^2 - j\Gamma_e \omega}\right), \quad \mu(\omega) = \mu_0 \left(1 - \frac{\omega_{pm}^2}{\omega^2 - \omega_{m0}^2 - j\Gamma_m \omega}\right), \tag{1.22}$$

where $\omega_{pe}, \omega_{pm}, \Gamma_e$ and Γ_m have the same meaning as the Drude model. Furthermore, ω_{e0} and ω_{m0} are the electric and magnetic resonance frequencies, respectively.

A derivation of $\mu_r(\omega) = 1 - \frac{F\omega^2}{\omega^2 - \omega_{m0}^2 - j\Gamma_m \omega}$ is shown by Pendry et al. for a composite medium consisting of a square array of cylinders with split ring structure (cf. [236, Fig. 3]) formed by two sheets separated by a distance d. More specifically, they derived

$$\omega_{m0}^2 = \frac{3dc^2}{\pi^2 r^3}, \quad \Gamma_m = \frac{2\sigma}{\mu_0 r}, \quad F = \frac{\pi r^2}{a^2},$$

where the parameters a, c, r and σ have the same meaning as in (1.14). Later, Smith and Kroll [269] changed $F\omega^2$ to $F\omega_0^2$ to ensure that $\mu_r(\omega) \to 1$ as $\omega \to \infty$. This new choice results the Lorentz model (1.22) with $\omega_{pm}^2 = F\omega_0^2$.

Transforming (1.22) into time domain, we obtain the Lorentz model equations for metamaterials:

$$\epsilon_0 \frac{\partial \mathbf{E}}{\partial t} + \frac{\partial \mathbf{P}}{\partial t} - \nabla \times \mathbf{H} = 0, \tag{1.23}$$

$$\mu_0 \frac{\partial \mathbf{H}}{\partial t} + \frac{\partial \mathbf{M}}{\partial t} + \nabla \times \mathbf{E} = 0, \tag{1.24}$$

$$\frac{1}{\epsilon_0\omega_{pe}^2}\frac{\partial^2\mathbf{P}}{\partial t^2}+\frac{\Gamma_e}{\epsilon_0\omega_{pe}^2}\frac{\partial\mathbf{P}}{\partial t}+\frac{\omega_{e0}^2}{\epsilon_0\omega_{pe}^2}\mathbf{P}-\mathbf{E}=0, \tag{1.25}$$

$$\frac{1}{\mu_0\omega_{pm}^2}\frac{\partial^2\mathbf{M}}{\partial t^2}+\frac{\Gamma_m}{\mu_0\omega_{pm}^2}\frac{\partial\mathbf{M}}{\partial t}+\frac{\omega_{m0}^2}{\mu_0\omega_{pm}^2}\mathbf{M}-\mathbf{H}=0. \tag{1.26}$$

The last popular model we want to mention is a mixed model used by engineers and physicists [123, 234, 236, 259, 269], in which the permittivity is described by the Drude model, while the permeability is described by the Lorentz model. More precisely, the permittivity is described by the Drude model [234, 236]:

$$\epsilon(\omega)=\epsilon_0(1-\frac{\omega_p^2}{\omega(\omega+j\nu)}), \tag{1.27}$$

where ω is the excitation angular frequency, $\omega_p > 0$ is the effective plasma frequency, and $\nu \geq 0$ is the loss parameter. On the other hand, the permeability can be described by the Lorentz model [259, 269]:

$$\mu(\omega)=\mu_0(1-\frac{F\omega_0^2}{\omega^2+j\gamma\omega-\omega_0^2}), \tag{1.28}$$

where $\omega_0 > 0$ is the resonant frequency, $\gamma \geq 0$ is the loss parameter, and $F \in (0, 1)$ is a parameter depending on the geometry of the unit cell of the metamaterial.

Using a time-harmonic variation of $exp(j\omega t)$, and substituting (1.27) and (1.28) into (1.11), respectively, we obtain the time-domain equation for the polarization:

$$\frac{\partial^2\mathbf{P}}{\partial t^2}+\nu\frac{\partial\mathbf{P}}{\partial t}=\epsilon_0\omega_p^2\mathbf{E}, \tag{1.29}$$

and the equation for the magnetization:

$$\frac{\partial^2\mathbf{M}}{\partial t^2}+\gamma\frac{\partial\mathbf{M}}{\partial t}+\omega_0^2\mathbf{M}=\mu_0F\omega_0^2\mathbf{H}. \tag{1.30}$$

To facility the mathematical study of the model, by introducing the induced electric current $\mathbf{J}=\frac{\partial\mathbf{P}}{\partial t}$ and magnetic current $\mathbf{K}=\frac{\partial\mathbf{M}}{\partial t}$, we can write the time domain governing equations for the Drude-Lorentz model as following:

$$\epsilon_0\frac{\partial\mathbf{E}}{\partial t}=\nabla\times\mathbf{H}-\mathbf{J}, \tag{1.31}$$

$$\mu_0\frac{\partial\mathbf{H}}{\partial t}=-\nabla\times\mathbf{E}-\mathbf{K}, \tag{1.32}$$

$$\frac{1}{\mu_0\omega_0^2F}\frac{\partial\mathbf{K}}{\partial t}+\frac{\gamma}{\mu_0\omega_0^2F}\mathbf{K}+\frac{1}{\mu_0F}\mathbf{M}=\mathbf{H}, \tag{1.33}$$

$$\frac{1}{\mu_0 F}\frac{\partial \mathbf{M}}{\partial t} = \frac{1}{\mu_0 F}\mathbf{K}, \tag{1.34}$$

$$\frac{1}{\epsilon_0 \omega_p^2}\frac{\partial \mathbf{J}}{\partial t} + \frac{\nu}{\epsilon_0 \omega_p^2}\mathbf{J} = \mathbf{E}. \tag{1.35}$$

In later chapters, we will develop various numerical methods for solving the Drude model (1.18)–(1.21), the Lorentz model (1.23)–(1.26), and the Drude-Lorentz model (1.31)–(1.35).

1.3 A Brief Overview of Computational Electromagnetics

Generally speaking, computational electromagnetics [45] can be classified into either frequency-domain simulation or time-domain simulation. Each category can be further classified into surface-based or volume-based methods. The method of moments (MoM) or boundary element method (BEM) is formulated as integral equations given on the surface of the physical domain. Note that MoM [138] or BEM [55, 56] is applicable to problems for which Green's functions of the underlying partial differential equations are available, which limits its applicability. Hence the volume-based methods such as the finite element method, the finite difference method, the finite volume method (e.g. [76,77,226,239]), and the spectral method (direct applications in computational electromagnetics see [168, 179]; applications in broader areas see [59, 142, 261, 282]) are quite popular.

One of the most favorite methods is the so-called finite-difference time-domain (FDTD) method proposed by Yee in 1966 [299]. Due to its simplicity, the FDTD method is very popular in electrical engineering community, and it is especially useful for broadband simulations, since one single simulation can cover a wide range of frequencies. For more details on the FDTD method, readers can consult Taflove and Hagness' book [276] and references cited therein. This book also provides complete 1-D to 3-D MATLAB source codes so that readers can learn the FDTD method quickly. A recent FDTD book by Hao and Mittra [137] focuses on the simulation of metamaterial models.

But the FDTD method has a major disadvantage when it is used for complex geometry simulation. In this case, the finite element method (FEM) is a better choice as evidenced by several published books in this area. For example, books [267] and [162] focus on how to develop and implement FEMs for solving Maxwell's equations. [267] even provides the Fortran source codes, but it only discusses the standard Lagrange finite elements, which are used to solve the Maxwell's equations written in scalar or vector potentials. Though edge elements are mentioned in this book, no implementation is provided. During 2006 and 2007, Demkowicz et al. published two books [97,98] on hp-adaptive finite element methods for solving both elliptic and time-harmonic Maxwell's equations. Demkowicz also publicized his 2-D Fortran 95 code in [97]. The code implements both rectangular and triangular

edge elements of different orders. In 2008, Hesthaven and Warburton published a very nice package *nudg* in their book [141]. *nudg* has both MATLAB and C++ versions, and can be used to solve the time-dependent Maxwell's equations written in conservation laws. If readers are interested in the finite element theory for Maxwell's equations, the best reference is Monk's book [217]. For a broad coverage on various methods (including FDTD method and FEM) and applications to Maxwell's equations, readers may consult the book by Cohen [85] and the book by Bondeson et al. [42]. However, all those books mentioned above mainly focus on Maxwell's equations in free space, except that [137] is devoted to Maxwell's equations in metamaterials.

In the rest of the book, we will focus on the finite element method due to our experience and interest.

1.4 Bibliographical Remarks

Though the field of metamaterials was born in 2000 [274], it has grown so rapidly that about 20 books (many are edited books) have been published since 2005. For more backgrounds on metamaterials, readers are encouraged to consult them [19, 57, 58, 61, 93, 94, 106, 109, 137, 171, 181, 208–210, 220, 228, 245, 256, 263, 274, 314]. However, they are almost exclusively focused on physics and applications of metamaterials. The only book focused on modeling of metamaterials is [137], which unfortunately covers only finite difference methods.

Chapter 2
Introduction to Finite Element Methods

The finite element method (FEM) is arguably one of the most robust and popular numerical methods used for solving various partial differential equations (PDEs). Due to the diligent work of many researchers over the past several decades, the fundamental theory and implementation of FEM have been well established as evidenced by many excellent books published in this area (e.g., [4, 20, 21, 39, 51, 54, 65, 78, 158, 163, 243]).

In this chapter, we provide a brief introduction to the basic FEM theory and programming techniques in order to prepare readers for extending these skills to solve metamaterial Maxwell's equations in later chapters.

The outline of this chapter is as follows: In Sect. 2.1, we introduce some basic concepts about constructing two-dimensional (2-D) and three-dimensional (3-D) Lagrange finite elements. Then in Sect. 2.2, we provide a succinct introduction to Sobolev spaces. After that, we present some classic finite element results such as the interpolation error estimates for Lagrange finite elements in Sect. 2.3. To prepare readers for more complicated analysis and algorithmic implementation in later chapters, we then provide a brief introduction to some basic finite element error analysis tools for elliptic type problems in Sect. 2.4. Finally, in Sect. 2.5, we introduce some standard coding techniques for implementing Lagrange finite elements for solving the second order elliptic problems.

2.1 Introduction to Finite Elements

Suppose that we want to numerically solve a given PDE on a fixed domain Ω. To use the finite element method, basically we need to proceed the following steps:

1. Rewrite a given PDE into an equivalent weak formulation.
2. Subdivide the physical domain Ω into smaller simple geometrical subdomains (or elements). Often we use tetrahedra, hexahedra or prisms for a 3-D domain, and triangles or quadrilaterals in a 2-D domain.

J. Li and Y. Huang, *Time-Domain Finite Element Methods for Maxwell's Equations in Metamaterials*, Springer Series in Computational Mathematics 43,
DOI 10.1007/978-3-642-33789-5_2, © Springer-Verlag Berlin Heidelberg 2013

3. Design a proper finite element, which is often denoted as a triple (K, P_K, Σ_K) according to [78]. Here K is a geometric element, P_K is a space of functions on K, and Σ_K is the so-called *degrees of freedom* of the finite element. For efficiency and simplicity reasons, P_K is often formed by polynomials. The degrees of freedom are often formed by values (or derivatives) of a function at the element vertices, or some integral forms of a function on the element edges and/or on the element.
4. Construct a finite element solution formed by basis functions of P_K to approximate the infinite dimensional solution in the weak formulation. Doing this leads to a system of discretized linear (or nonlinear) equations.
5. Solve the system of discretized equations and postprocess the obtained solution to get the numerical solution for the original given PDE.

In this section, we focus on the third step. The rest steps will be elaborated in later sections.

First, let us introduce some common notation for polynomial spaces used throughout the book. Let P_k be the space of polynomials of maximum total degree k in d variables $x_1, \cdots, x_d$, and $\tilde{P}_k$ be the space of polynomials of total degree exactly k in d variables $x_1, \cdots, x_d$. Hence a polynomial $p \in P_k$ if and only if it can be written as

$$p(\mathbf{x}) = \sum_{\alpha_1+\cdots+\alpha_d \le k} c_{\alpha_1,\cdots,\alpha_d} x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_d^{\alpha_d}$$

at any point $\mathbf{x} = (x_1, \cdots, x_d)$, and a polynomial $\tilde{p} \in \tilde{P}_k$ if and only if it can be written as

$$\tilde{p}(\mathbf{x}) = \sum_{\alpha_1+\cdots+\alpha_d = k} c_{\alpha_1,\cdots,\alpha_d} x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_d^{\alpha_d}$$

for proper coefficients $c_{\alpha_1,\cdots,\alpha_d}$. Here all α_i are assumed to be non-negative integers.

It is easy to see that in $\mathscr{R}^d$, the dimensions of the spaces P_k and $\tilde{P}_k$ are

$$dim(P_k) = \binom{k+d}{k} = \frac{(k+d)\cdots(k+1)}{d!} \tag{2.1}$$

and

$$dim(\tilde{P}_k) = dim(P_k) - dim(P_{k-1}), \tag{2.2}$$

respectively.

On a d-dimensional rectangle, we need a tensor-product polynomial space $Q_{l_1,\cdots,l_d}$, which is formed by polynomials of maximum degree l_k in x_k, where $1 \le k \le d$, i.e., a polynomial $q \in Q_{l_1,\cdots,l_d}$ if and only if it can be written as

$$q(\mathbf{x}) = \sum_{0\le\alpha_1\le l_1,\cdots,0\le\alpha_d\le l_d} c_{\alpha_1,\cdots,\alpha_d} x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_d^{\alpha_d}$$

for some coefficients $c_{\alpha_1,\cdots,\alpha_d}$.

The dimension of $Q_{l_1,\cdots,l_d}$ is easy to calculate as

$$dim(Q_{l_1,\cdots,l_d}) = (l_1+1)(l_2+1)\cdots(l_d+1).$$

Before we move forward, let us introduce the *unisolvent* concept used in finite element.

Definition 2.1. A finite element (K, P_K, Σ_K) is *unisolvent* if the set of degrees of freedom Σ_K uniquely defines a function in P_K.

Below are some examples of unisolvent finite elements.

Example 2.1. Consider a triangle K with vertices $(x_i, y_i), i = 1, 2, 3,$ ordered counterclockwisely. It is known that the area A of this triangle can be calculated as

$$A = \frac{1}{2}\begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix}.$$

Now we can define the so-called *barycentric coordinates* $\lambda_i(x, y)$ of the triangle, which is often denoted as

$$\lambda_i(x, y) = \frac{1}{2A}(\alpha_i + \beta_i x + \gamma_i y), \quad i = 1, 2, 3, \tag{2.3}$$

where constants α_i, β_i and γ_i are

$$\alpha_i = x_j y_k - x_k y_j, \ \ \beta_i = y_j - y_k, \ \ \gamma_i = -(x_j - x_k),$$

where $i \neq j \neq k$, and i, j and k permute naturally.

It is not difficult to see that: for any $1 \le i, j \le 3$, we have

$$\lambda_i(x_j, y_j) = \delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases}$$

from which we see that any function $u \in P_1$ on K can be uniquely represented by

$$u(x, y) = \sum_{i=1}^{3} u(x_i, y_i)\lambda_i(x, y) \quad \forall\ (x, y) \in K.$$

Using the triple notation, we can denote this unisolvent element (often called as P_1 element) as:

$$K = \{\text{The triangle with vertices } (x_i, y_i), i = 1, 2, 3, \}$$
$$P_K = \text{Polynomials of degree 1 in variables } x \text{ and } y,$$
$$\Sigma_K = \{\text{Function values at the vertices: } u(x_i, y_i), i = 1, 2, 3.\}$$

Using the barycentric coordinates λ_i, we can define other finite elements. Below is the so-called P_2 element:

Example 2.2.

$$K = \{\text{A triangle with vertices } a_i(x_i, y_i), i = 1, 2, 3,$$
$$\text{and edge midpoints } a_{ij},\ 1 \le i < j \le 3\},$$
$$P_K = \text{Polynomials of degree 2 in variables } x \text{ and } y.$$
$$\Sigma_K = \{\text{Function values at vertices and midpoints: } u(a_i), u(a_{ij}), i, j = 1, 2, 3.\}$$

We can prove that P_2 element is unisolvent.

Lemma 2.1. *Any function $u \in P_2$ on K is uniquely determined by its values at all vertices and edge midpoints, i.e., by $u(a_i)$, $1 \le i \le 3$, and $u(a_{ij})$, $1 \le i < j \le 3$.*

Proof. Note that the total number of degrees of freedom in Σ_K is equal to 6, which is the same as $dim(P_2)$. Hence we only need to show that if $u(a_i) = u(a_{ij}) = 0$, then $u \equiv 0$. Note that the restriction of u to edge a_2a_3 is a quadratic function in one variable and vanishes at three distinct points (i.e., at a_2, a_3 and a_{23}), hence $u(x, y)$ must be zero on this edge. This implies that u should contain a factor $\lambda_1(x, y)$.

By the same argument, u must be zero on edge a_1a_2, which implies that u should contain a factor $\lambda_3(x, y)$. Similarly, because u is zero on edge a_1a_3, u should also contain a factor $\lambda_2(x, y)$. Therefore, we can write

$$u(x, y) = c\lambda_1(x, y)\lambda_2(x, y)\lambda_3(x, y),$$

which becomes a third-order polynomial unless $c = 0$. Hence, $u \equiv 0$, which concludes the proof. □

Actually, any function u in P_2 element can be explicitly represented as [78, p. 47]:

$$u(x, y) = \sum_{i=1}^{3} u(a_i)\lambda_i(x, y)(2\lambda_i(x, y) - 1) + \sum_{1 \le i < j \le 3} 4u(a_{ij})\lambda_i(x, y)\lambda_j(x, y).$$

Similarly, we can construct a P_1 element on a tetrahedron.

Example 2.3.

$$\begin{aligned} Denote \quad K &= \{\text{A tetrahedron with four vertices } V_i(x_i, y_i, z_i), i = 1, 2, 3, 4\}, \\ P_K &= \text{Polynomials of degree 1 in three variables}, \\ \Sigma_K &= \{\text{Function values at the vertices: } u(V_i), i = 1, 2, 3, 4\}. \end{aligned}$$

On element K, any function u of P_K can be written as

$$u(x, y, z) = a + bx + cy + dz, \tag{2.4}$$

where the coefficients a, b, c, and d can be determined by enforcing (2.4) equal to the given values $u(V_i)$ at the four vertices of the tetrahedron. Introducing the short notation $u_i = u(V_i)$, we have

$$\begin{aligned} a + bx_1 + cy_1 + dz_1 &= u_1, \\ a + bx_2 + cy_2 + dz_2 &= u_2, \\ a + bx_3 + cy_3 + dz_3 &= u_3, \\ a + bx_4 + cy_4 + dz_4 &= u_4, \end{aligned}$$

solving which we obtain

$$a = \frac{1}{6V}\begin{vmatrix} u_1 & u_2 & u_3 & u_4 \\ x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \\ z_1 & z_2 & z_3 & z_4 \end{vmatrix} = \frac{1}{6V}(a_1u_1 + a_2u_2 + a_3u_3 + a_4u_4),$$

$$b = \frac{1}{6V}\begin{vmatrix} 1 & 1 & 1 & 1 \\ u_1 & u_2 & u_3 & u_4 \\ y_1 & y_2 & y_3 & y_4 \\ z_1 & z_2 & z_3 & z_4 \end{vmatrix} = \frac{1}{6V}(b_1u_1 + b_2u_2 + b_3u_3 + b_4u_4),$$

$$c = \frac{1}{6V}\begin{vmatrix} 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 \\ u_1 & u_2 & u_3 & u_4 \\ z_1 & z_2 & z_3 & z_4 \end{vmatrix} = \frac{1}{6V}(c_1u_1 + c_2u_2 + c_3u_3 + c_4u_4),$$

$$d = \frac{1}{6V}\begin{vmatrix} 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \\ u_1 & u_2 & u_3 & u_4 \end{vmatrix} = \frac{1}{6V}(d_1u_1 + d_2u_2 + d_3u_3 + d_4u_4),$$

where the coefficients a, b, c and d can be determined from the expansion of determinants, and V denotes the volume of the element, which can be expressed as

$$V = \frac{1}{6}\begin{vmatrix} 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \\ z_1 & z_2 & z_3 & z_4 \end{vmatrix}.$$

Substituting a, b, c and d back into (2.4) and collecting like terms u_j, we have

$$u(x, y, z) = \sum_{j=1}^{4} u(x_j, y_j, z_j) N_j(x, y, z), \tag{2.5}$$

where functions N_j are given by

$$N_j(x, y, z) = \frac{1}{6V}(a_j + b_j x + c_j y + d_j z). \tag{2.6}$$

We like to remark that $N_j(x, y, z)$ are often called the *shape functions* of the finite element, and they have the property

$$N_j(x_i, y_i, z_i) = \delta_{ij} = \begin{cases} 1 & i = j, \\ 0 & i \neq j. \end{cases}$$

By a similar technique, we can construct finite elements on d-rectangles. First, we give an example on a rectangular element.

Example 2.4. Consider a rectangle $K = [x_c - h_x, x_c + h_x] \times [y_c - h_y, y_c + h_y]$, whose four vertices are oriented counterclockwisely, starting with the bottom-left vertex $V_1 = (x_c - h_x, y_c - h_y)$. Then we can denote the Q_1 rectangular element by the triple:

$$\begin{aligned} K &= \{\text{The rectangle with four vertices } V_i, i = 1, 2, 3, 4\}, \\ P_K &= \text{Polynomial } Q_{1,1} \text{ on } K, \\ \Sigma_K &= \{\text{Function values at the vertices: } u(V_i), i = 1, 2, 3, 4\}. \end{aligned}$$

It is easy to check that any function u of $Q_{1,1}$ on K can be uniquely represented as follows:

$$u(x, y) = \sum_{j=1}^{4} u(x_j, y_j) N_j(x, y),$$

where the shape functions N_j are given by

$$N_1(x, y) = \frac{1}{A}(x_c + h_x - x)(y_c + h_y - y),$$

$$N_2(x,y) = \frac{1}{A}(x - x_c + h_x)(y_c + h_y - y),$$
$$N_3(x,y) = \frac{1}{A}(x - x_c + h_x)(y - y_c + h_y),$$
$$N_4(x,y) = \frac{1}{A}(x - x_c + h_x)(y - y_c + h_y),$$

here $A = 4h_x h_y$ denotes the area of the rectangle.

Now we give an example on a cubic element.

Example 2.5. Consider a cube

$$K = [x_c - h_x, x_c + h_x] \times [y_c - h_y, y_c + h_y] \times [z_c - h_z, z_c + h_z],$$

whose eight vertices are oriented counterclockwisely (four on the bottom face, and four on the top face), starting with the front-bottom-left vertex $V_1 = (x_c - h_x, y_c - h_y, z_c - h_z)$. We can define a unisolvent Q_1 cubic element by the triple:

$$K = \{\text{The cube with 8 vertices } V_i, i = 1, 2, \cdots, 8\},$$
$$P_K = \text{Polynomial } Q_{1,1,1} \text{ on } K,$$
$$\Sigma_K = \{\text{Function values at the vertices: } u(V_i), i = 1, 2, \cdots, 8\}.$$

It is not difficult to check that any function u of $Q_{1,1,1}$ on K can be uniquely represented as follows:

$$u(x,y,z) = \sum_{j=1}^{8} u(x_j, y_j, z_j) N_j(x,y,z),$$

where the shape functions N_j are given by

$$N_1(x,y,z) = \frac{1}{V}(x_c + h_x - x)(y_c + h_y - y)(z_c + h_z - z),$$
$$N_2(x,y,z) = \frac{1}{V}(x - x_c + h_x)(y_c + h_y - y)(z_c + h_z - z),$$
$$N_3(x,y,z) = \frac{1}{V}(x - x_c + h_x)(y - y_c + h_y)(z_c + h_z - z),$$
$$N_4(x,y,z) = \frac{1}{V}(x_c + h_x - x)(y - y_c + h_y)(z_c + h_z - z),$$

and the other four N_j have the same form as above except that the last terms are changed to $z - (z_c - h_z)$. Here $V = 8h_x h_y h_z$ denotes the volume of the cube.

2.2 Functional Analysis and Sobolev Spaces

2.2.1 Basic Functional Analysis

In our later analysis of Maxwell's equations, we shall appeal to many basic theorems from functional analysis. In this section, we simply summarize some definitions and theorems to be used in later sections. Readers interested in details can consult specialized books such as [53].

Let X be a normed linear space with norm $||\cdot||_X$.

Definition 2.2. A sequence $\{u_k\}_{k=1}^{\infty} \subset X$ *converges* to $u \in X$, denoted as $u_k \to u$, if $\lim_{k\to\infty} ||u_k - u||_X = 0$. If for any $\epsilon > 0$, there exists $N > 0$ such that

$$||u_k - u_m||_X < \epsilon \quad \text{for any } k, m \geq N,$$

then the sequence $\{u_k\}_{k=1}^{\infty} \subset X$ is called a *Cauchy sequence*.

Definition 2.3. The space X is called *complete* if each Cauchy sequence in X converges; namely, whenever $\{u_k\}_{k=1}^{\infty}$ is a Cauchy sequence, there exists $u \in X$ such that $u_k \to u$.

Definition 2.4. A complete normed linear space is called a *Banach space*.

Definition 2.5. A collection $\mathscr{M}$ of subsets of $\mathscr{R}^d$ is called σ*-algebra* if

(i) $\emptyset, \mathscr{R}^d \in \mathscr{M}$;
(ii) $A \in \mathscr{M}$ implies that $\mathscr{R}^d \setminus A \in \mathscr{M}$;
(iii) $\{A_k\}_{k=1}^{\infty} \subset \mathscr{M}$ implies that $\cup_{k=1}^{\infty} A_k, \cap_{k=1}^{\infty} A_k \in \mathscr{M}$.

It can be proved that there exists a σ-algebra $\mathscr{M}$ of subsets of $\mathscr{R}^d$ and a mapping $|\cdot| : \mathscr{M} \to [0, +\infty]$ with the following properties:

(i) Every open subset of $\mathscr{R}^d$ and every closed subset of $\mathscr{R}^d$ belong to $\mathscr{M}$.
(ii) If A is a ball of $\mathscr{R}^d$, then $|A|$ equals the d-dimensional volume of A.

The sets in $\mathscr{M}$ are often called *Lebesgue measurable* sets and $|\cdot|$ is *Lebesgue measure*. Hence Lebesgue measure provides a way of describing the volume of subsets of $\mathscr{R}^d$.

Definition 2.6. A function $f : \mathscr{R}^d \to \mathscr{R}$ is called a *measurable function* if $f^{-1}(S) \in \mathscr{M}$ for every open subset $S \subset \mathscr{R}$.

Definition 2.7. Let Ω be an open subset of $\mathscr{R}^d$, and $1 \leq p \leq \infty$. For a measurable function $f : \Omega \to \mathscr{R}$, we define the norm

$$||f||_{L^p(\Omega)} = (\int_{\Omega} |f|^p dx)^{1/p} \quad \text{if } 1 \leq p < \infty$$

and

$$||f||_{L^\infty(\Omega)} = \inf_{S\subset\Omega,|S|=0} \sup_{\Omega\setminus S} |f(x)| = \text{ess sup}_\Omega |f(x)| \quad \text{if } p=\infty.$$

Furthermore, $L^p(\Omega)$ is defined to be the linear space of all measurable functions $f : \Omega \to \mathscr{R}$ for which $||f||_{L^p(\Omega)} < \infty$.

It is known that $L^p(\Omega), 1 \le p \le \infty$, is a Banach space.

Definition 2.8. Let X be a real linear space. A mapping $(\cdot,\cdot) : X \times X \to \mathscr{R}$ is called an *inner product* if

(i) $(u, v) = (v, u)$ for any $u, v \in X$;
(ii) Each of the maps $u \to (u, v)$ and $v \to (u, v)$ is linear on X;
(iii) $(u, u) \ge 0$ for any $u \in X$;
(iv) $(u, u) = 0$ if any only if $u = 0$.

Furthermore, if $(\cdot,\cdot)$ is an inner product, the associated norm is

$$||u||_X = (u, u)^{1/2} \quad \text{for any} \quad u \in X.$$

Moreover, if X is complete with respect to the norm $||\cdot||_X$, then X is called a *Hilbert space*.

A simple example of a Hilbert space is $L^2(\Omega)$, which has the scalar inner product

$$(u, v) = \int_\Omega u(x)v(x)dx.$$

Two elementary estimates for Hilbert spaces are often used in numerical analysis. One is the Cauchy-Schwarz inequality

$$|(u, v)| \le ||u||_X||v||_X \quad \forall\, u, v \in X. \tag{2.7}$$

The other one is the arithmetic-geometric mean inequality: for any $u, v \in X$ and $\delta > 0$, we have

$$|(u, v)| \le \frac{\delta}{2}||u||_X^2 + \frac{1}{2\delta}||v||_X^2. \tag{2.8}$$

2.2.2 Sobolev Spaces

Sobolev spaces are named after the Russian mathematician Sergei Sobolev, who introduced this concept around 1950. A Sobolev space is a space of functions equipped with a norm which can be used to measure both the size and regularity of a function. Hence, Sobolev spaces play a very important role in analyzing partial

differential equations. In this section, we just present some important properties of Sobolev spaces related to our later usage. For more comprehensive discussions, readers can consult specialized books on Sobolev spaces (e.g. [2]).

Let $C_0^\infty(\Omega)$ (or $D(\Omega)$) denote the space of infinitely differentiable functions with compact support in Ω. Furthermore, let $\alpha = (\alpha_1, \cdots, \alpha_d)$ be an d-tuple of nonnegative integers and denote its length by $|\alpha| = \sum_{i=1}^{d} \alpha_i$.

Before we introduce the weak derivative concept, let us provide some motivation. Suppose we are given a function $u \in C^1(\Omega)$. If $\phi \in C_0^\infty(\Omega)$, using integration by parts we have

$$\int_\Omega u\phi_{x_i} dx = -\int_\Omega u_{x_i}\phi dx \quad i = 1, 2, \cdots, d. \tag{2.9}$$

Here no boundary integral term exists, since ϕ has compact support in Ω and vanishes on $\partial\Omega$. Similarly, if $u \in C^k(\Omega)$ for some integer $k > 1$, we can obtain

$$\int_\Omega uD^\alpha\phi dx = (-1)^{|\alpha|}\int_\Omega D^\alpha u\phi dx \quad \forall\, \phi \in C_0^\infty(\Omega). \tag{2.10}$$

Now (2.10) has a problem if u is not C^k, since $D^\alpha u$ on the right hand side has no obvious meaning. We resolve this difficulty by introducing the concept of weak derivative. Let us denote the set of locally integrable functions

$$L^1_{loc}(\Omega) = \{v : \ v \in L^1(K), \ \text{ for all compact set } K \subset \Omega\}.$$

Definition 2.9. Let $u, v \in L^1_{loc}(\Omega)$. We say v is the α-th *weak partial derivative* of u, denoted as $v = D^\alpha u$, provided that

$$\int_\Omega uD^\alpha\phi dx = (-1)^{|\alpha|}\int_\Omega v\phi dx \tag{2.11}$$

holds true for all $\phi \in C_0^\infty(\Omega)$

It is easy to see that if a weak derivative exists, then it is uniquely defined up to a set of measure zero. Furthermore, for any $u \in C^{|\alpha|}(\Omega)$, the weak derivative $D^\alpha u$ exists and equals the classic derivative.

With the above preparations, we can define the Sobolev space $W^{k,p}(\Omega)$.

Definition 2.10. The Sobolev space $W^{k,p}(\Omega)$ consists of all locally summable functions $u : \Omega \to R$, i.e., $u \in L^1_{loc}(\Omega)$, such that for each multi-index α with $|\alpha| \le k$, $D^\alpha u$ exists in the weak sense and belongs to $L^p(\Omega)$. Moreover, for any $u \in W^{k,p}(\Omega)$, its norm is defined to be

$$||u||_{W^{k,p}(\Omega)} = \begin{cases} (\sum_{|\alpha|\le k}\int_\Omega |D^\alpha u|^p dx)^{1/p} & \text{if } 1 \le p < \infty, \\ \sum_{|\alpha|\le k} \text{ess sup}_\Omega |D^\alpha u| & \text{if } p = \infty. \end{cases}$$

It is known that $W^{k,p}(\Omega)$ is a Banach space (see e.g., [51, p. 28]).

We like to remark that a function belonging to a Sobolev space can be discontinuous and/or unbounded. A simple example is $u(x) = |x|^{-\alpha}$, which is unbounded at $x = 0$ for $\alpha > 0$. However, $u(x)$ belongs to some Sobolev space.

Lemma 2.2. *Let $u(x) = |x|^{-\alpha}$ $(x \neq 0)$ defined in the open unit ball $\Omega = B(0,1)$ in $\mathscr{R}^d$. Then $u \in W^{1,p}(\Omega)$ if and only if $\alpha < \frac{d-p}{p}$.*

Proof. Note that u is smooth away from 0, and

$$u_{x_i}(x) = -\alpha x_i |x|^{-(\alpha+2)} \quad (x \neq 0).$$

For any $\phi \in C_0^\infty(\Omega)$ and fixed $\epsilon > 0$, we have

$$\int_{\Omega \setminus B(0,\epsilon)} u\phi_{x_i} dx = \int_{\partial B(0,\epsilon)} u\phi n_i ds - \int_{\Omega \setminus B(0,\epsilon)} u_{x_i}\phi dx,$$

where n_i denotes the unit inward normal on $\partial B(0,\epsilon)$.

If $\alpha + 1 < d$, then $|Du(x)| = \frac{\alpha}{|x|^{\alpha+1}} \in L^1(\Omega)$, in which case,

$$|\int_{\partial B(0,\epsilon)} u\phi n_i ds| \leq ||\phi||_{L^\infty} \int_{\partial B(0,\epsilon)} \epsilon^{-\alpha} ds \leq C\epsilon^{d-1-\alpha} \to 0 \text{ as } \epsilon \to 0.$$

Hence for any $0 \leq \alpha < d - 1$, we have

$$\int_\Omega u\phi_{x_i} dx = -\int_\Omega u_{x_i}\phi dx \quad \forall\ \phi \in C_0^\infty(\Omega).$$

Similarly, it is easy to see that

$$|Du(x)| = \frac{\alpha}{|x|^{\alpha+1}} \in L^p(\Omega) \text{ if and only if } (\alpha+1)p < d,$$

which concludes the proof. □

Definition 2.11. Given a subset $S \subset X$, the *closure* of S in X (usually denoted as $\overline{S}$) is the set of all limits of convergent subsequence of S using the X norm. Furthermore, if $\overline{S} = X$, we say that the subset S is *dense* in X.

Definition 2.12. $W_0^{k,p}(\Omega)$ is denoted as the *closure* of $C_0^\infty(\Omega)$ in $W^{k,p}(\Omega)$. When $p = 2$, we usually write

$$H^k(\Omega) = W^{k,2}(\Omega) \quad (\text{or } H_0^k(\Omega) = W_0^{k,2}(\Omega)),$$

since $H^k(\Omega)$ is a Hilbert space.

Definition 2.13. A function $f : \Omega \to \mathscr{R}$ is called *Lipschitz continuous* if

$$|f(x) - f(y)| \leq L|x - y|$$

for some constant $L > 0$ and all $x, y \in \Omega$.

Definition 2.14. Let Ω be an open and bounded domain of $\mathscr{R}^d (d \geq 2)$ with boundary $\partial\Omega$. We say that Ω is *Lipschitz* (or $\partial\Omega$ is a Lipschitz boundary), if there exists a finite open cover $U_1, \cdots, U_m$ of $\partial\Omega$ such that for $j = 1, \cdots, m$:

(i) $\partial\Omega \cap U_j$ is the graph of a Lipschitz function $\phi_j : \mathscr{R}^{d-1} \to \mathscr{R}$, and
(ii) $\Omega \cap U_j$ is on one side of this graph.

Namely, for any $x = (\tilde{x}, x_d) \in U_j$, where $\tilde{x} = (x_1, \cdots, x_{d-1})$, there exists a Lipschitz function ϕ_j such that $x_d = \phi_j(\tilde{x})$, $\Omega \cap U_j = \{x : x_d > \phi_j(\tilde{x})\}$ and $\partial\Omega \cap U_j = \{x : x_d = \phi_j(\tilde{x})\}$.

Definition 2.15. A normed space U is said to be *embedded* in another normed space V, denoted as $U \hookrightarrow V$, if

(i) U is a linear subspace of V;
(ii) The injection of U into V is continuous, i.e., there exists a constant $C > 0$ such that $||u||_V \leq C\,||u||_U \quad \forall\, u \in U$.

As we mentioned earlier, Sobolev spaces provide a way of quantifying the degree of smoothness of functions. The following theorem summaries some classic results [2].

Theorem 2.1. *(Sobolev embedding theorem) Suppose that Ω is an open set of R^d with a Lipschitz continuous boundary, and $1 \leq p < \infty$. Then the following embedding results hold true:*

(i) If $0 \leq kp < d$, then $W^{k,p}(\Omega) \hookrightarrow L^{p_*}(\Omega)$ for $p_* = \frac{dp}{d-kp}$.
(ii) If $kp = d$, then $W^{k,p}(\Omega) \hookrightarrow L^q(\Omega)$ for any q such that $p \leq q < \infty$.
(iii) If $kp > d$, then $W^{k,p}(\Omega) \hookrightarrow C^0(\overline{\Omega})$.

When we deal with time-dependent problems, we need some Sobolev spaces involving time. Let X denote a real Banach space with norm $||\cdot||_X$.

Definition 2.16. The space $L^p(0, T; X)$ consists of all measurable functions $u : [0, T] \to X$ with endowed norm

$$||u||_{L^p(0,T;X)} = \left(\int_0^T ||u(t)||_X^p dt\right)^{1/p} < \infty, \quad \text{if } 1 \leq p < \infty,$$

and

$$||u||_{L^\infty(0,T;X)} = \text{ess sup}_{0 \leq t \leq T} ||u(t)||_X < \infty, \quad \text{if } p = \infty.$$

Definition 2.17. The space $C(0,T;X)$ consists of all continuous functions $u : [0,T] \to X$ with

$$||u||_{C(0,T;X)} = \max_{0\le t\le T} ||u(t)||_X < \infty.$$

Definition 2.18. The Sobolev space $W^{k,p}(0,T;X)$ comprises all functions $u \in L^p(0,T;X)$ such that $\frac{d^\alpha u}{dt^\alpha}, 1 \le \alpha \le k$, exists in the weak sense and belongs to $L^p(0,T;\Omega)$. Furthermore,

$$||u||_{W^{k,p}(0,T;X)} = \begin{cases} (\int_0^T (||u(t)||_X^p + \sum_{1\le\alpha\le k} ||\frac{d^\alpha u}{dt^\alpha}||_X^p)dt)^{1/p} & \text{if } 1 \le p < \infty, \\ \text{ess sup}_{0\le t\le T}(||u(t)||_X + \sum_{1\le\alpha\le k} ||\frac{d^\alpha u}{dt^\alpha}||_X) & \text{if } p = \infty. \end{cases}$$

When $p = 2$, we denote

$$H^k(0,T;X) = W^{k,2}(0,T;X).$$

For many applications to be discussed later, we need a space of vector functions with a square-integrable divergence. Such a space is often denoted as $H(\text{div};\Omega)$, which is defined by

$$H(\text{div};\Omega) = \{\mathbf{v} \in (L^2(\Omega))^d : \nabla\cdot\mathbf{v} \in L^2(\Omega)\}, \tag{2.12}$$

with norm $||\mathbf{v}||_{H(\text{div};\Omega)} = (||\mathbf{v}||^2_{(L^2(\Omega))^d} + ||\nabla\cdot\mathbf{v}||^2_{L^2(\Omega)})^{1/2}$, where $\nabla\cdot$ is the divergence operator defined as

$$\nabla\cdot\mathbf{v} = \sum_{i=1}^{d} \frac{\partial v_i}{\partial x_i}, \quad \text{for all } \mathbf{v} \in (C_0^\infty(\Omega))^d,\ d = 2,3.$$

It is easy to prove that $H(\text{div};\Omega)$ is a Hilbert space with the inner product

$$(\mathbf{u},\mathbf{v})_{div} = \int_\Omega (\mathbf{u}\cdot\mathbf{v} + (\nabla\cdot\mathbf{u})(\nabla\cdot\mathbf{v}))d\mathbf{x}.$$

Similar to space $W_0^{k,p}(\Omega)$, we denote $H_0(\text{div};\Omega)$ as the closure of $(C_0^\infty(\Omega))^d$ with respect to the norm $||\cdot||_{H(\text{div};\Omega)}$.

Later, when we deal with Maxwell's equations, we need a space of vector functions with a square-integrable curl. In standard notation, we define this space by

$$H(\text{curl};\Omega) = \{\mathbf{u} \in (L^2(\Omega))^d : \nabla\times\mathbf{u} \in (L^2(\Omega))^d\}, \tag{2.13}$$

with norm $||\mathbf{u}||_{H(\text{curl};\Omega)} = (||\mathbf{u}||^2_{(L^2(\Omega))^d} + ||\nabla\times\mathbf{u}||^2_{(L^2(\Omega))^d})^{1/2}$, where $\nabla\times$ is the curl operator defined as

$$\nabla\times\mathbf{u} = (\frac{\partial u_3}{\partial x_2} - \frac{\partial u_2}{\partial x_3}, \frac{\partial u_1}{\partial x_3} - \frac{\partial u_3}{\partial x_1}, \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2})',$$

for any 3-D vector $\mathbf{u} = (u_1, u_2, u_3)' \in (C_0^\infty(\Omega))^d,\ d = 3$. For a 2-D vector, the curl operator becomes as

$$\nabla \times \mathbf{u} = \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2}, \quad \text{for all } \mathbf{u} = (u_1, u_2)' \in (C_0^\infty(\Omega))^2.$$

It is easy to prove that $H(\text{curl}; \Omega)$ is a Hilbert space with the inner product

$$(\mathbf{u}, \mathbf{v})_{curl} = \int_\Omega (\mathbf{u} \cdot \mathbf{v} + (\nabla \times \mathbf{u}) \cdot (\nabla \times \mathbf{v}))d\mathbf{x}.$$

Furthermore, $H_0(\text{curl}; \Omega)$ is denoted as the closure of $(C_0^\infty(\Omega))^d$ with respect to the norm $||\cdot||_{H(\text{curl};\Omega)}$.

Similarly, with higher regularity, we can define the space

$$H^\alpha(\text{curl}; \Omega) = \{\mathbf{v} \in (H^\alpha(\Omega))^d : \nabla \times \mathbf{v} \in (H^\alpha(\Omega))^d\}, \tag{2.14}$$

for any $\alpha > 0$ with norm $||\mathbf{v}||_{\alpha,\text{curl}} = (||\mathbf{v}||^2_{(H^\alpha(\Omega))^d} + ||\nabla \times \mathbf{v}||^2_{(H^\alpha(\Omega))^d})^{1/2}$. When $\alpha = 0$, $H^\alpha(\text{curl}; \Omega)$ reduces to $H(\text{curl}; \Omega)$.

2.3 Classic Finite Element Theory

To use the finite element method to solve a PDE, we have to build up a finite dimensional space of functions on the physical domain Ω. To do this, we first need to generate a finite element mesh covering the domain Ω, i.e., we need to construct a set T_h of non-overlapping elements K_i satisfying the following conditions:

(i) $\overline{\Omega} = \bigcup_{K_i \in T_h} \overline{K_i}$;
(ii) Each $K \in T_h$ is a Lipschitz domain, and has a positive measurement;
(iii) For any distinct elements K_1 and K_2 in T_h, we have $K_1 \cap K_2 = \emptyset$. In other words, any two neighboring elements have to meet at the common vertices, match exactly at a common edge or a common face.

A mesh satisfying conditions (i)–(iii) is often called *conforming mesh*. Note that in hp finite element method [97, 98, 255], condition (iii) is often violated, in which case we have the so-called *non-conforming mesh*.

2.3.1 Conforming and Non-conforming Finite Elements

After creating a mesh for Ω, we can use the element-wise defined finite elements to construct a global finite element space on Ω by lumping together all the element degrees of freedom. The key here is how to define the element degrees of freedom to guarantee the needed global smoothness for the finite element solution. For this, we need to classify finite elements into two classes: *conforming* or *non-conforming*.

Definition 2.19. Let V be a space of functions. The finite element (K, P_K, Σ_K) is said to be V *conforming* if its corresponding global finite element space is a subspace of V. Otherwise, the finite element is said to be V *nonconforming*.

It is well-known that for a finite element space to be $H^1(\Omega)$ conforming, the global finite element function has to be continuous as stated in the next lemma (cf. [78, Theorem 2.1.1] and [217, Lemma 5.3]).

Lemma 2.3. *Let K_1 and K_2 be two non-overlapping Lipschitz domains having a common interface Λ such that $\overline{K_1} \cap \overline{K_2} = \Lambda$. Assume that $u_1 \in H^1(K_1)$ and $u_2 \in H^1(K_2)$, and $u \in L^2(K_1 \cup K_2 \cup \Lambda)$ be defined by*

$$u = \begin{cases} u_1 & \text{on } K_1, \\ u_2 & \text{on } K_2. \end{cases}$$

Then $u_1 = u_2$ on Λ implies that $u \in H^1(K_1 \cup K_2 \cup \Lambda)$.

Proof. Suppose that we have a function $u \in L^2(K_1 \cup K_2 \cup \Lambda)$ defined by $u|_{K_i} = u_i, i = 1, 2$, and $u_1 = u_2$ on Λ. To prove that $u \in H^1(K_1 \cup K_2 \cup \Lambda)$, for any function $\boldsymbol{\phi} \in (C_0^\infty(K_1 \cup K_2 \cup \Lambda))^d$, using integration by parts, we have

$$\begin{aligned} &\int_{K_1 \cup K_2 \cup \Lambda} u \frac{\partial \boldsymbol{\phi}}{\partial x_i} d\mathbf{x} = \int_{K_1} u \frac{\partial \boldsymbol{\phi}}{\partial x_i} d\mathbf{x} + \int_{K_2} u \frac{\partial \boldsymbol{\phi}}{\partial x_i} d\mathbf{x} \\ &= -\int_{K_1} \frac{\partial (u|_{K_1})}{\partial x_i} \boldsymbol{\phi} d\mathbf{x} - \int_{K_2} \frac{\partial (u|_{K_2})}{\partial x_i} \boldsymbol{\phi} d\mathbf{x} + \int_{\Lambda} (u_1 \boldsymbol{\phi} \cdot \mathbf{n}_{i,1} + u_2 \boldsymbol{\phi} \cdot \mathbf{n}_{i,2}) ds, \end{aligned}$$

where $\mathbf{n}_{i,j}$ denotes the unit outward normal to ∂K_j, $j = 1, 2$, respectively.

Denote $v_i = \frac{\partial (u|_{K_l})}{\partial x_i}, i = 1, \cdots, d$, on $K_l, l = 1, 2$. Using the assumption that $u_1 = u_2$ on Λ, we see that the boundary integral term vanishes. Hence, we have

$$\int_{K_1 \cup K_2 \cup \Lambda} u \frac{\partial \boldsymbol{\phi}}{\partial x_i} d\mathbf{x} = -\int_{K_1 \cup K_2 \cup \Lambda} \mathbf{v} \cdot \boldsymbol{\phi} d\mathbf{x},$$

which shows that $u \in H^1(K_1 \cup K_2 \cup \Lambda)$ by the definition of weak derivative. □

Examples 2.1–2.5 given in Sect. 2.1 are H^1 conforming elements. Many popular nonconforming elements are illustrated in the nice paper by Carstensen and Hu [63]. Below we present two examples of non-conforming elements. The first one is the so-called rotated Q1 element.

Example 2.6. Consider a rectangle $K = [x_c - h_x, x_c + h_x] \times [y_c - h_y, y_c + h_y]$, whose four vertices are oriented counterclockwise, starting with the bottom-left vertex $V_1(x_c - h_x, y_c - h_y)$. The four mid-edge points starting from the bottom edge are denoted as $M_i, i = 1, 2, 3, 4$. The Q_1 rectangular element is defined by the following triple:

$K = \{$The rectangle with four vertices $V_i, i = 1, 2, 3, 4\}$,

$P_K =$ Polynomial with basis $1, x, y, x^2, y^2$ on K,

$\Sigma_K = \{$Function values at the mid-edge points: $u(M_i), i = 1, 2, 3, 4\}$.

The unisolvence of this finite element (K, P_K, Σ_K) can be proved as follows. Assume that an arbitrary function u of P_K on K is represented as

$$u(x, y) = a_1 + a_2(x - x_c) + a_3(y - y_c) + a_4[(x - x_c)^2 - (y - y_c)^2], \quad (2.15)$$

where the unknown coefficients a_1, a_2, a_3 and a_4 can be determined by the interpolation conditions

$$u(M_i) = u_i, i = 1, 2, 3, 4. \quad (2.16)$$

Imposing (2.16) at the four mid-edge points, we have

$$\begin{aligned} a_1 - h_y a_3 - h_y^2 a_4 &= u_1, \\ a_1 + h_y a_3 - h_y^2 a_4 &= u_3, \\ a_1 + h_x a_2 + h_x^2 a_4 &= u_2, \\ a_1 - h_x a_2 + h_x^2 a_4 &= u_4, \end{aligned}$$

which gives the following unique solution

$$\begin{aligned} a_1 &= [h_x^2(u_1 + u_3) + h_y^2(u_2 + u_4)]/2(h_x^2 + h_y^2), \\ a_2 &= (u_2 - u_4)/2h_x, \\ a_3 &= (u_3 - u_1)/2h_y, \\ a_4 &= [(u_2 + u_4) - (u_1 + u_3)]/2(h_x^2 + h_y^2). \end{aligned}$$

Another popular non-conforming element is Wilson's rectangular element.

Example 2.7. Consider the same rectangle $K = [x_c - h_x, x_c + h_x] \times [y_c - h_y, y_c + h_y]$ as Example 2.6. The Wilson element can be defined by the following triple:

$K = \{$The rectangle with four vertices $V_i, i = 1, 2, 3, 4\}$,

$P_K =$ Polynomial P_2 on K,

$\Sigma_K = \{$Function values at the vertices: $u(V_i), i = 1, 2, 3, 4,$ and mean values of the second derivatives of u over K : $\dfrac{1}{|K|}\displaystyle\int_K \frac{\partial^2 u}{\partial x^2} d\mathbf{x}, \dfrac{1}{|K|}\int_K \frac{\partial^2 u}{\partial y^2} d\mathbf{x}\}$.

The unisolvence of this finite element (K, P_K, Σ_K) can be proved as follows. For a function u of P_2 on K, we can represent it as

$$u(x, y) = a_1+a_2(x-x_c)+a_3(y-y_c)+a_4(x-x_c)(y-y_c)+a_5(x-x_c)^2+a_6(y-y_c)^2, \tag{2.17}$$

where the unknown coefficients a_i can be determined by the given degrees of freedom Σ_K, which lead to the system of linear equations:

$$\begin{aligned}
&a_1 - h_x a_2 - h_y a_3 + h_x h_y a_4 + h_x^2 a_5 + h_y^2 a_6 = u_1,\\
&a_1 + h_x a_2 - h_y a_3 - h_x h_y a_4 + h_x^2 a_5 + h_y^2 a_6 = u_2,\\
&a_1 + h_x a_2 + h_y a_3 + h_x h_y a_4 + h_x^2 a_5 + h_y^2 a_6 = u_3,\\
&a_1 - h_x a_2 + h_y a_3 - h_x h_y a_4 + h_x^2 a_5 + h_y^2 a_6 = u_4,\\
&\int_K \frac{\partial^2 u}{\partial x^2} d\mathbf{x} = 2|K| a_5,\\
&\int_K \frac{\partial^2 u}{\partial y^2} d\mathbf{x} = 2|K| a_6.
\end{aligned}$$

Solving the above system, we can obtain the following unique solution

$$\begin{aligned}
&a_5 = \frac{1}{2|K|}\int_K \frac{\partial^2 u}{\partial x^2} d\mathbf{x}, \quad a_6 = \frac{1}{2|K|}\int_K \frac{\partial^2 u}{\partial y^2} d\mathbf{x},\\
&a_1 = \frac{u_1+u_2+u_3+u_4}{4} - \frac{h_x^2}{2|K|}\int_K \frac{\partial^2 u}{\partial x^2} d\mathbf{x} - \frac{h_y^2}{2|K|}\int_K \frac{\partial^2 u}{\partial y^2} d\mathbf{x},\\
&a_2 = (u_2 - u_1 + u_3 - u_4)/4h_x,\\
&a_3 = (u_3 + u_4 - u_1 - u_2)/4h_y,\\
&a_4 = [(u_3 - u_4) - (u_2 - u_1)]/4h_x h_y.
\end{aligned}$$

2.3.2 Basic Interpolation Error Estimates

Given a finite element (K, P_K, Σ_K), we can define a local Lagrange interpolant:

$$\Pi_K^k v = \sum_{i=1}^{n} v(a_i)\phi_i,$$

where a_i are the vertices of K, $v(a_i)$ are the degrees of freedom, and ϕ_i are the shape functions. Examples include the first-order interpolant

$$\Pi_K^1 v(x, y) = \sum_{i=1}^{3} v(a_i)\lambda_i(x, y),$$

and the second-order interpolant

$$\Pi_K^2 v(x, y) = \sum_{i=1}^{3} v(a_i)\lambda_i(2\lambda_i - 1) + \sum_{1\le i<j\le 3}^{3} 4v(a_{ij})\lambda_i\lambda_j,$$

mentioned in Sect. 2.1. Examples $\Pi_K^1 v$ and $\Pi_K^2 v$ are Lagrange interpolations, which satisfy the property $\Pi_K^{1,2} v(a_i) = v(a_i), i = 1, 2, 3$. More complicated interpolants such as Example 2.7 involve other degrees of freedom.

By piecing together the local interpolants, we can define a corresponding global interpolant Π_h^k as follows:

$$(\Pi_h^k v)|_K = \Pi_K^k(v|_K) \quad \forall\ K \in T_h.$$

We assume further that each element K of T_h can be obtained as an affine mapping of a reference element $\hat{K}$, i.e.,

$$K = F_K(\hat{K}), \quad F_K(\hat{x}) = B_K\hat{x} + b_K, \tag{2.18}$$

where B_K is a $d \times d$ non-singular matrix. The rest of this section is concerned about the estimate of interpolation error $v - \Pi_h^k v$.

Lemma 2.4. *Denote* $\hat{v}(\hat{x}) = v(F_K(\hat{x}))$. *If* $v \in W^{m,p}(K), m \ge 0, p \in [1, \infty]$, *then* $\hat{v} \in W^{m,p}(\hat{K})$. *Moreover, there exists a constant* $C = C(m, p, d)$ *such that*

$$|v|_{m,p,K} \le C||B_K^{-1}||^m |det(B_K)|^{1/p} |\hat{v}|_{m,p,\hat{K}}, \quad \forall\ \hat{v} \in W^{m,p}(\hat{K}), \tag{2.19}$$

and

$$|\hat{v}|_{m,p,\hat{K}} \le C||B_K||^m |det(B_K)|^{-1/p} |v|_{m,p,K}, \quad \forall\ v \in W^{m,p}(K), \tag{2.20}$$

where $||\cdot||$ *denotes the matrix norm associated to the Euclidean norm in* R^d.

Proof. For simplicity we just show the proof of (2.19) for $p = 2$. A complete proof can be found in the classic book by Ciarlet [78, Theorem 3.1.2]. Since $C^\infty(K)$ is dense in $H^m(K)$, it is sufficient to prove (2.19) for a smooth function v.

Using the chain rule and the mapping F_K, we have

$$|v|^2_{m,2,K} = \sum_{|\alpha|=m} ||D^\alpha v||^2_{0,K} \le C||B_K^{-1}||^{2m} \sum_{|\beta|=m} \int_K |\hat{D}^\beta \hat{v}|^2 d\mathbf{x}$$
$$\le C||B_K^{-1}||^{2m} \sum_{|\beta|=m} ||\hat{D}^\beta \hat{v}||^2_{0,\hat{K}} \cdot (\det(B_K)), \tag{2.21}$$

which completes the proof of (2.19) for $p = 2$. □

To obtain an explicit bound on $||B_K||, ||B_K^{-1}||$ and $\det(B_K)$, we introduce some notation. For an element K, we denote

$$h_K = \text{diameter of the smallest sphere (or circle) containing } K,$$

and

$$\rho_K = \text{diameter of the largest sphere (or circle) inscribed in } K.$$

Similarly, $h_{\hat{K}}$ and $\rho_{\hat{K}}$ are used for the reference element $\hat{K}$.

Noting that $\det(B_K) = \text{vol}(K)/\text{vol}(\hat{K})$, we easily have

$$C_1 \rho_K^d \le |\det(B_K)| \le C_2 h_K^d, \tag{2.22}$$

where the positive constants C_1 and C_2 are independent of h_K and ρ_K.

Lemma 2.5. *The following estimates hold*

$$||B_K^{-1}|| \le \frac{h_{\hat{K}}}{\rho_K}, \quad ||B_K|| \le \frac{h_K}{\rho_{\hat{K}}}.$$

Proof. By definition, we have

$$||B_K^{-1}|| = \frac{1}{\rho_K} \sup_{|\xi|=\rho_K} |B_K^{-1}\xi|. \tag{2.23}$$

For any ξ satisfying $|\xi| = \rho_K$, we can always find two points $x, y \in K$ such that $x - y = \xi$. Note that $|B_K^{-1}\xi| = |B_K^{-1}(x - y)| = |\hat{x} - \hat{y}| \le h_{\hat{K}}$, substituting which into (2.23) completes the proof of the first inequality. The other one can be proved in a similar way. □

Lemma 2.6 ([78, Theorem 3.1.1]). *There exists a constant $C(\hat{K})$ such that*

$$\inf_{\hat{p}\in P_k(\hat{K})} ||\hat{v} + \hat{p}||_{k+1,p,\hat{K}} \le C(\hat{K})|\hat{v}|_{k+1,p,\hat{K}} \quad \forall\, \hat{v} \in W^{k+1,p}(\hat{K}).$$

With the above preparations, we can prove the following interpolation error estimates.

Theorem 2.2 ([78, Theorem 3.1.5]). *Let $(\hat{K}, \hat{P}_K, \hat{\Sigma_K})$ be a finite element. If*

$$W^{k+1,p}(\hat{K}) \hookrightarrow C^s(\hat{K}),$$
$$W^{k+1,p}(\hat{K}) \hookrightarrow W^{m,q}(\hat{K}),$$
$$P_k(\hat{K}) \subset \hat{P}_K \subset W^{m,q}(\hat{K}),$$

hold true for integers $m, k \geq 0$, and some numbers $p, q \in [1, \infty]$, where s denotes the greatest order of derivatives occurring in the degrees of freedom set $\hat{\Sigma_K}$, then there exists a constant C, depending on $\hat{K}$, $\hat{P}_K$ and $\hat{\Sigma_K}$, such that

$$|v - \Pi_K^k v|_{m,q,K} \leq C\,|det(B_K)|^{1/q-1/p} \frac{h_K^{k+1}}{\rho_K^m} |v|_{k+1,p,K} \quad \forall\, v \in W^{k+1,p}(K).$$

Proof. Noting that the shape functions $\hat{\phi}_i$ in $\hat{K}$ are given by $\hat{\phi}_i = \phi_i \circ F_K$, we obtain

$$\widehat{\Pi_K^k v} = \Pi_h^k v \circ F_K = \sum_i v(a_i)(\phi_i \circ F_K) = \sum_i v(F_K(\hat{a}_i))\hat{\phi}_i = \Pi_{\hat{K}}^k \hat{v},$$

from which and Lemmas 2.4–2.6, we have

$$\begin{aligned}
|v - \Pi_K^k v|_{m,q,K} &\leq C\,||B_K^{-1}||^m |det(B_K)|^{1/q} |\hat{v} - \widehat{\Pi_K^k v}|_{m,q,\hat{K}} \leq \frac{C}{\rho_K^m} |det(B_K)|^{1/q} |\hat{v} - \Pi_{\hat{K}}^k \hat{v}|_{m,q,\hat{K}} \\
&\leq \frac{C}{\rho_K^m} |det(B_K)|^{1/q} ||I - \Pi_{\hat{K}}^k||_{\mathscr{L}(W^{k+1,p};W^{m,q})} \inf_{\hat{p} \in P_k} ||\hat{v} + \hat{p}||_{k+1,p,\hat{K}} \\
&\leq \frac{C}{\rho_K^m} |det(B_K)|^{1/q} C\,|\hat{v}|_{k+1,p,\hat{K}} \\
&\leq \frac{C}{\rho_K^m} |det(B_K)|^{1/q} ||B_K||^{k+1} |det(B_K)|^{-1/p} |v|_{k+1,p,K} \\
&\leq \frac{C}{\rho_K^m} h_K^{k+1} |det(B_K)|^{1/q-1/p} |v|_{k+1,p,K},
\end{aligned}$$

which completes the proof. □

The parameter ρ_K can be eliminated if the finite element mesh is regular.

Definition 2.20. A triangulation T_h of Ω is called *regular* when $h = \max_{K \in T_h} h_K$ approaches zero, if there exists a constant $\sigma \geq 1$, independent of h, such that

$$\frac{h_K}{\rho_K} \leq \sigma \quad \text{for any } K \in T_h.$$

Under the regularity assumption, from Theorem 2.2 and (2.22), we can obtain the following interpolation error estimate over a physical domain Ω.

Theorem 2.3. *In addition to the assumptions of Theorem 2.2, if the mesh of Ω is regular, then*

$$|v - \Pi_h^k v|_{m,q,\Omega} \leq C h^{d(1/q-1/p)+k+1-m} |v|_{k+1,p,\Omega} \quad \forall\, v \in W^{k+1,p}(\Omega).$$

2.4 Finite Element Analysis for Elliptic Problems

In this section, we will present some basic finite element error analysis techniques developed for elliptic problems. For simplicity, we limit our discussion to conforming finite elements. More general discussions can be found in classic books such as [51, 78, 243].

2.4.1 Abstract Convergence Theory

Consider an abstract variational problem: Find $u \in V$ such that

$$A(u, v) = F(v) \quad \forall\, v \in V, \tag{2.24}$$

where V denotes a real Hilbert space and $F \in V'$. Here V' denotes the dual space of V. The following famous Lax-Milgram lemma justifies the existence and uniqueness of the solution for this problem.

Theorem 2.4 (Lax-Milgram lemma [78, p. 8]). *Let V be a real Hilbert space with norm $||\cdot||_V$, $A(\cdot,\cdot)$ be a bilinear form from $V \times V$ to R, and $F(\cdot)$ be a linear continuous functional from V to R. Furthermore, suppose that $A(\cdot,\cdot)$ is bounded:*

$$\exists \beta > 0 \text{ such that } |A(w, v)| \leq \beta ||w||_V ||v||_V \quad \text{for all } w, v \in V,$$

and coercive:

$$\exists \alpha > 0 \text{ such that } |A(v, v)| \geq \alpha ||v||_V^2 \quad \text{for all } v \in V.$$

Then, there exists a unique solution $u \in V$ to (2.24) and

$$||u||_V \leq \frac{1}{\alpha} ||F||_{V'}.$$

Assume that a family of finite dimensional subspaces V_h is constructed to approximate the infinite dimensional space V, i.e.,

$$\inf_{v_h \in V_h} ||v - v_h||_V \to 0 \text{ as } h \to 0, \quad \text{for all } v \in V.$$

The standard Galerkin FEM for solving (2.24) is: Find $u_h \in V_h$ such that

$$A(u_h, v_h) = F(v_h) \quad \forall\, v_h \in V_h. \tag{2.25}$$

From (2.24) and (2.25), we obtain the Galerkin orthogonality relation:

$$A(u - u_h, v_h) = 0 \quad \forall\, v_h \in V_h. \tag{2.26}$$

Combining (2.26) with the coercivity and continuity of $A(\cdot, \cdot)$, we have

$$\begin{aligned} \alpha ||u - u_h||_V^2 &\le A(u - u_h, u - u_h) = A(u - u_h, u - v_h) \\ &\le \beta ||u - u_h||_V ||u - v_h||_V, \end{aligned}$$

which leads to the following stability and convergence result.

Theorem 2.5 (Céa lemma). *Under the assumption of Theorem 2.4, there exists a unique solution u_h to (2.25) and*

$$||u_h||_V \le \frac{1}{\alpha} ||F||_{V'}.$$

Furthermore, if u denotes the solution to (2.24), then

$$||u - u_h||_V \le \frac{\beta}{\alpha} \inf_{v_h \in V_h} ||u - v_h||_V, \tag{2.27}$$

i.e., u_h converges to u as $h \to 0$.

2.4.2 Error Estimate for an Elliptic Problem

Here we demonstrate how the abstract convergence theory presented in last section can be used for a specific problem. Without loss of generality, let us consider the following elliptic boundary value problem

$$-\triangle u + u = f \quad \text{in } \Omega, \tag{2.28}$$

$$u = 0 \quad \text{on } \partial\Omega. \tag{2.29}$$

Multiplying (2.28) by a test function $v \in H_0^1(\Omega)$ and using the Green's formula

$$-\int_\Omega \triangle u v dx = -\int_{\partial\Omega} \frac{\partial u}{\partial n} v ds + \int_\Omega \sum_{i=1}^{d} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i}, \quad \forall\, u \in H^2(\Omega), v \in H^1(\Omega), \tag{2.30}$$

where $\frac{\partial}{\partial n} = \sum_{i=1}^{d} n_i \frac{\partial}{\partial x_i}$ is the normal derivative operator, we can obtain an equivalent variational problem: Find $u \in H_0^1(\Omega)$ such that

$$A(u, v) \equiv (\nabla u, \nabla v) + (u, v) = (f, v), \quad \forall\, v \in H_0^1(\Omega). \tag{2.31}$$

Application of the Cauchy-Schwarz inequality shows that

$$|A(u, v)| = |(\nabla u, \nabla v) + (u, v)| \le ||\nabla u||_0 ||\nabla v||_0 + ||u||_0 ||v||_0 \le 2||u||_1 ||v||_1,$$

which means that $A(\cdot, \cdot)$ is continuous on $H_0^1(\Omega) \times H_0^1(\Omega)$.

On the other hand, we easily obtain

$$A(v, v) = ||\nabla v||_0^2 + ||v||_0^2 = ||v||_1^2,$$

which proves the coercivity of $A(\cdot, \cdot)$. Therefore, by the Lax-Milgram lemma, the variational problem (2.31) has a unique solution $u \in H_0^1(\Omega)$.

To solve (2.31) by the finite element method, we construct a finite dimensional subspace V_h of $H_0^1(\Omega)$:

$$V_h = \{v_h \in H_0^1(\Omega);\ \ \forall\ \ K \in T_h, v_h|_K \in P_k(K)\}, \tag{2.32}$$

where T_h is a regular family of triangulations of Ω.

The finite element approximation $u_h \in V_h$ of (2.31) is: Find $u_h \in V_h$ such that

$$(\nabla u_h, \nabla v_h) + (u_h, v_h) = (f, v_h), \quad \forall\, v_h \in V_h.$$

For the elliptic problem (2.28) and (2.29), the following optimal error estimates hold true in both H^1 and L^2 norms.

Theorem 2.6. *Let Ω be a polygonal domain of $R^d, d = 2, 3$, with Lipschitz boundary, and T_h be a regular family of triangulations of Ω. Let V_h be defined in (2.32). If the exact solution $u \in H^s(\Omega) \cap H_0^1(\Omega), s \ge 2$, the error estimate holds*

$$||u - u_h||_1 \le C h^l ||u||_{l+1}, \quad l = \min(k, s-1), k \ge 1.$$

Suppose, furthermore, for each $e \in L^2(\Omega)$, the solution w of the adjoint problem (see (2.35) below) of (2.28) belongs to $H^2(\Omega)$ and satisfies

$$||w||_2 \le C||e||_0, \quad \forall\, e \in L^2(\Omega). \tag{2.33}$$

Then we have

$$||u - u_h||_0 \le C h^{l+1} ||u||_{l+1}, \quad l = \min(k, s-1), k \ge 1.$$

Proof. By the Céa lemma, for any $u \in H^s(\Omega) \cup H_0^1(\Omega), s \geq 2$, we have

$$||u - u_h||_1 \leq 2 \inf_{v_h \in V_h} ||u - v_h||_1 \leq 2||u - \Pi_h^k u||_1 \leq C h^l ||u||_{l+1},$$

where $l = \min(k, s-1)$, and Π_h^k is the finite element interpolation operator defined in Sect. 2.3.

To derive error estimates in the L^2-norm, we need to use the so-called Aubin-Nitsche technique (also called duality argument). Denote error $e = u - u_h$, and let w be the solution of the adjoint problem of (2.28):

$$-\triangle w + w = e \ \text{ in } \Omega, \quad w = 0 \ \text{ on } \ \partial\Omega, \tag{2.34}$$

whose variational formulation is: Find $w \in H_0^1(\Omega)$ such that

$$A(v, w) = (e, v) \quad \forall\, v \in H_0^1(\Omega). \tag{2.35}$$

Hence, by the Galerkin orthogonality relation, we have

$$\begin{aligned} ||u - u_h||_0^2 &= (e, u - u_h) = A(u - u_h, w) = A(u - u_h, w - \Pi_h^k w) \\ &\leq 2||u - u_h||_1 ||w - \Pi_h^k w||_1 \leq C h^l ||u||_{l+1} \cdot h||w||_2, \end{aligned} \tag{2.36}$$

which, along with the bound (2.33), leads to

$$||u - u_h||_0 \leq C h^{l+1} ||u||_{l+1},$$

which is optimal in the L^2-norm. This concludes the proof. □

Using more sophisticated weighted-norm technique, error estimates in the L^∞ norm can be proved [78, p. 165].

Theorem 2.7. *Under the assumption of $u \in H_0^1(\Omega) \cap W^{k+1,\infty}(\Omega), k \geq 1$, we have*

$$||u - u_h||_{\infty,\Omega} + h||\nabla(u - u_h)||_{\infty,\Omega} \leq C h^2 |\ln h||u|_{2,\infty,\Omega}, \ \ \textit{for } k = 1,$$

and

$$||u - u_h||_{\infty,\Omega} + h||\nabla(u - u_h)||_{\infty,\Omega} \leq C h^{k+1} |u|_{k+1,\infty,\Omega}, \ \ \textit{for } k \geq 2.$$

2.5 Finite Element Programming for Elliptic Problems

In this section, we introduce the basic procedures for programming a finite element method used for solving the second-order elliptic boundary value problems. We want to remark that finite element programming is quite a sophisticated task

and can be written in a stand-alone book. Readers can find more advanced and sophisticated programming algorithms in books devoted to this subject (e.g., [21, 62, 97, 107, 112, 158, 174, 240]). For example, [21] presents the package PLTMG, which solves elliptic problems using adaptive FEM in MATLAB; [174] introduces Diffpack, a sophisticated toolbox for solving PDEs in C++; [252] elaborates the adaptive finite element software ALBERTA written in ANSI-C; [107] introduces an open-source MATLAB package IFISS, which can be used to solve convection-diffusion, Stokes, and Navier-Stokes equations. Demkowicz [97] presents a self-contained hp-finite element package for solving elliptic problems and time-harmonic Maxwell's equations.

In the rest of this section, we detail the important steps for programming a finite element method in MATLAB to solve the elliptic problem:

$$-\triangle u + u = f \quad \text{in } \Omega = (0,1)^2, \tag{2.37}$$

$$u = 0 \quad \text{on } \partial\Omega, \tag{2.38}$$

by using Q_1 element. The material of this section is modified from our previous book [187, Chap. 7].

2.5.1 The Basic Steps

2.5.1.1 FEM Mesh Generation

As we mentioned earlier, we need to subdivide the physical domain Ω into small elements. For simplicity, here we generate a uniform rectangular mesh T_h for Ω. The mesh structure can be clearly described by four variables and four arrays. More specifically, we use nx and ny for the total number of elements in the x- and y-direction, respectively; ne and np for the total number of elements and the total number of nodes in T_h, respectively. We use the 1-D array $x(1:np)$ and $y(1:np)$ to represent the nodal x and y coordinates, respectively. A 2-D array conn(1:ne,1:4) is used to store the connectivity matrix for the mesh, i.e., $conn(i, 1:4)$ specifies the four node numbers of element i. A 2-D array $gbc(1:np, 1:2)$ is used to store the indicators for Dirichlet nodes in $gbc(1:np, 1)$, and the corresponding boundary values in $gbc(1:np, 2)$.

The rectangular mesh T_h can be generated using the MATLAB code *getQ1mesh.m*, which is shown below.

```
% Generate Q1 mesh on [xlow, xhigh]x[ylow, yhigh]
% nx,ny: number of elements in each direction
% x,y: 1-D array for nodal coordinates
% conn(1:ne,1:4): connectivity matrix
% ne, np: total numbers of elements, nodes generated
% gbc(1:np, 1:2): store Dirichlet node labels and values
```

```
function[x,y,conn,ne,np,gbc] ...
      = getQ1mesh(xlow,xhigh,ylow,yhigh,nx,ny)

ne = nx*ny;
np = (nx+1)*(ny+1);

% create nodal coordinates
dx=(xhigh - xlow)/nx;  dy=(yhigh - ylow)/ny;
for i = 1:(nx+1)
   for j=1:(ny+1)
      x((ny+1)*(i-1)+j) = dx*(i-1);
      y((ny+1)*(i-1)+j) = dy*(j-1);
   end
end

% form the connectivity matrix:
%  start from low-left corner countclockwisely.
for j=1:nx
   for i=1:ny
      ele = (j-1)*ny + i;
      conn(ele,1) = ele + (j-1);
      conn(ele,2) = conn(ele,1) + ny + 1;
      conn(ele,3) = conn(ele,2) + 1;
      conn(ele,4) = conn(ele,1) + 1;
   end
end

% pick out Dirichlet BC nodes
for i=1:np
    if (abs(x(i) - xlow) < 0.1*dx ...
          | abs(x(i) - xhigh) < 0.1*dx)
        gbc(i,1)=1;   % find one BC node
    elseif (abs(y(i) - ylow) < 0.1*dy ...
          | abs(y(i) - yhigh) < 0.1*dy)
        gbc(i,1)=1;   % find one BC node
    end
end
```

A simple 4×4 rectangular mesh generated with this code is shown in Fig. 2.1, where the nodal numbers and element numbers are provided. Note that the nodes of each element are oriented counterclockwisely. For example, the connectivity matrices for elements 1 and 2 are given by:

$$conn(1, 1:4) = 1, 6, 7, 2, \quad conn(2, 1:4) = 2, 7, 8, 3.$$

2.5.1.2 Forming FEM Equations

The finite element method for solving (2.37) and (2.38) is: Find $u_h \in \mathbf{v}_h \subset H_0^1(\Omega)$ such that

$$(\nabla u_h, \nabla \phi_h) + (u_h, \phi_h) = (f, \phi_h) \quad \forall\, \phi_h \in \mathbf{v}_h, \tag{2.39}$$

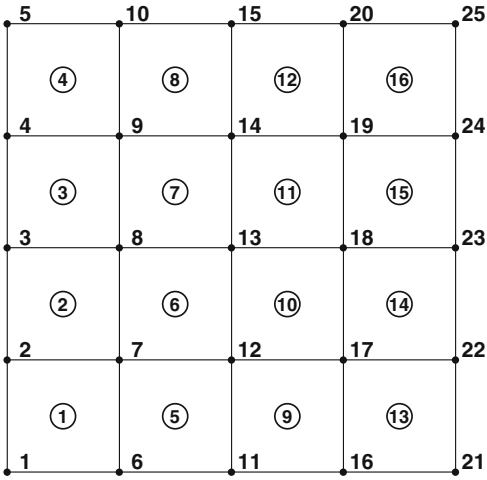

Fig. 2.1 An exemplary Q_1 element: node and element labelings

where the Q_1 finite element space is defined as

$$\mathbf{V}_h = \{v \in H_0^1(\Omega);\ \forall\ K \in T_h, v|_K \in Q_1(K) \text{ and } v|_{K\cap\partial\Omega} = 0\}.$$

On each rectangular element K, the exact solution u is approximated by

$$u_h^K(x, y) = \sum_{j=1}^{4} u_j^K \psi_j^K(x, y), \tag{2.40}$$

which leads to a 4×4 element coefficient matrix of (2.39) with entries

$$A_{ij} \equiv \int_K \nabla\psi_j^K \cdot \nabla\psi_i^K dxdy + \int_K \psi_j^K \psi_i^K dxdy,\ \ i, j = 1, \cdots, 4. \tag{2.41}$$

2.5.1.3 Calculation of Element Matrices

The calculation of A_{ij} is often carried out on a reference rectangle $\hat{K}$ with vertices

$$(\xi_1, \eta_1) = (-1, -1),\ (\xi_2, \eta_2) = (1, -1),\ (\xi_3, \eta_3) = (1, 1),\ (\xi_4, \eta_4) = (-1, 1),$$

whose corresponding shape functions

$$\hat{\psi}_i(\xi, \eta) = \frac{1}{4}(1 + \xi_i\xi)(1 + \eta_i\eta), \quad i = 1, \cdots, 4. \tag{2.42}$$

The mapping between an arbitrary rectangular element with vertices (x_i, y_i), $1 \leq i \leq 4$, and the reference element is given by

$$x = \sum_{j=1}^{4} x_j \hat{\psi}_j(\xi, \eta), \quad y = \sum_{j=1}^{4} y_j \hat{\psi}_j(\xi, \eta). \tag{2.43}$$

The basis function $\psi_j(x, y)$ on a general rectangle is defined as

$$\psi_j(x, y) = \hat{\psi}_j(\xi(x, y), \eta(x, y)),$$

from which we have

$$\frac{\partial \hat{\psi}_j}{\partial \xi} = \frac{\partial \psi_j}{\partial x}\frac{\partial x}{\partial \xi} + \frac{\partial \psi_j}{\partial y}\frac{\partial y}{\partial \xi}, \tag{2.44}$$

$$\frac{\partial \hat{\psi}_j}{\partial \eta} = \frac{\partial \psi_j}{\partial x}\frac{\partial x}{\partial \eta} + \frac{\partial \psi_j}{\partial y}\frac{\partial y}{\partial \eta}, \tag{2.45}$$

i.e.,

$$J^T \cdot \nabla \psi_j = \begin{bmatrix} \frac{\partial \hat{\psi}_j}{\partial \xi} \\ \frac{\partial \hat{\psi}_j}{\partial \eta} \end{bmatrix}, \quad j = 1, 2, 3,$$

where we denote the gradient

$$\nabla \psi_j = \begin{bmatrix} \frac{\partial \psi_j}{\partial x} \\ \frac{\partial \hat{\psi}_j}{\partial y} \end{bmatrix}$$

and the Jacobi matrix J of the mapping as

$$J \equiv \begin{bmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial x}{\partial \eta} \\ \frac{\partial y}{\partial \xi} & \frac{\partial y}{\partial \eta} \end{bmatrix}.$$

In the above, J^T denotes the transpose of J.

From (2.44) and (2.45), we see that

$$\nabla \psi_j = (J^T)^{-1} \begin{bmatrix} \frac{\partial \hat{\psi}_j}{\partial \xi} \\ \frac{\partial \hat{\psi}_j}{\partial \eta} \end{bmatrix} = \frac{1}{det(J)} \begin{bmatrix} \frac{\partial y}{\partial \eta} & -\frac{\partial y}{\partial \xi} \\ -\frac{\partial x}{\partial \eta} & \frac{\partial x}{\partial \xi} \end{bmatrix} \begin{bmatrix} \frac{\partial \hat{\psi}_j}{\partial \xi} \\ \frac{\partial \hat{\psi}_j}{\partial \eta} \end{bmatrix} \tag{2.46}$$

$$= \frac{1}{det(J)} \begin{bmatrix} (\sum_{j=1}^{4} y_j \frac{\partial \hat{\psi}_j}{\partial \eta})\frac{\partial \hat{\psi}_j}{\partial \xi} - (\sum_{j=1}^{4} y_j \frac{\partial \hat{\psi}_j}{\partial \xi})\frac{\partial \hat{\psi}_j}{\partial \eta} \\ -(\sum_{j=1}^{4} x_j \frac{\partial \hat{\psi}_j}{\partial \eta})\frac{\partial \hat{\psi}_j}{\partial \xi} + (\sum_{j=1}^{4} x_j \frac{\partial \hat{\psi}_j}{\partial \xi})\frac{\partial \hat{\psi}_j}{\partial \eta} \end{bmatrix}. \tag{2.47}$$

Since it is too complicated to evaluate A_{ij} analytically, below we use Gaussian quadrature over the reference rectangle to approximate A_{ij}. For example,

$$\int_K G(x,y)dxdy = \int_{\hat{K}} G(\hat{x},\hat{y})|J|d\xi d\eta \approx \sum_{m=1}^{N}(\sum_{n=1}^{N} G(\xi_m,\eta_n)|J|\omega_n)\omega_m,$$

where ξ_m and η_m, and ω_m and ω_n are the quadrature points and weights in the ξ and η directions, respectively. In our implementation, we use the second-order Gaussian quadrature rule, in which case

$$\xi_1 = -\frac{1}{\sqrt{3}}, \quad \xi_2 = \frac{1}{\sqrt{3}}, \quad \omega_1 = \omega_2 = 1.$$

The right-hand side vector (f,ϕ_i) is approximated as

$$(f,\phi_i) \approx (\frac{1}{4}\sum_{j=1}^{4} f_j,\phi_i), \quad i = 1,2,3,4.$$

Below is our MATLAB code *elemA.m*, which is used to calculate the element matrix and right-hand side vector.

```
function [ke,rhse] = elemA(conn,x,y,gauss,rhs,e);

% Q1 elementary stiffness matrix
ke = zeros(4,4);
rhse=zeros(4,1);
one = ones(1,4);
psiJ = [-1, +1, +1, -1]; etaJ = [-1, -1, +1, +1];

% coordinates of each element 'e'
for j=1:4
   je = conn(e,j);   % get the node label
   xe(j) = x(je); ye(j) = y(je); % get the coordinates
end

for i=1:2  % loop over gauss points in eta
  for j=1:2   % loop over gauss points in psi
     eta = gauss(i);  psi = gauss(j);
  % construct shape functions: starting at low-left corner
     NJ=0.25*(one + psi*psiJ).*(one + eta*etaJ);
  % derivatives of shape functions in reference coordinate
     NJpsi = 0.25*psiJ.*(one + eta*etaJ);     % 1x4 array
     NJeta = 0.25*etaJ.*(one + psi*psiJ);     % 1x4 array
     % derivatives of x and y wrt psi and eta
```

```
    xpsi = NJpsi*xe';  ypsi = NJpsi*ye';
    xeta = NJeta*xe';  yeta = NJeta*ye';
    Jinv = [yeta, -xeta; -ypsi, xpsi];      % 2x2 array
    jcob = xpsi*yeta - xeta*ypsi;
  % derivatives of shape functions in original coordinate
    NJdpsieta = [NJpsi; NJeta];             % 2x4 array
    NJdxy = Jinv*NJdpsieta;                 % 2x4 array
    % assemble element stiffness matrix ke: 4x4 array
    ke = ke + (NJdxy(1,:))'*(NJdxy(1,:))/jcob ...
            + (NJdxy(2,:))'*(NJdxy(2,:))/jcob ...
            + NJ(1,:)'*NJ(1,:)*jcob;
    rhse = rhse + rhs*NJ'*jcob;
  end
end
```

2.5.1.4 Assembly of Global Matrix

To obtain the global coefficient matrix, we need to assembly the contributions from each element coefficient matrix, i.e., we need to loop through all elements in the mesh, find the corresponding global nodal label for each local node, and put them in the right locations of the global coefficient matrix. A pseudo code is listed below:

```
for n=1:NE     % loop through all elements
  % computer element coefficient matrix
   for i=1:4    % loop through all nodes
      i1 = conn(n,i)
      for j=1:4
         j1=conn(n,j)
         Ag(i1,j1)=Ag(i1,j1)+Aloc(i,j)
      End
   End
End
```

After the assembly process, we need to impose the given Dirichlet boundary conditions. Suppose that after assembly we obtain a global linear system

$$A\mathbf{u} = \mathbf{b}. \tag{2.48}$$

Assume that we have to impose a Dirichlet boundary condition $u = u(x_k, y_k)$ at the k-th global node. A simple way to impose this boundary condition is as follows: First, replace each entry b_i of $\mathbf{b}$ by $b_i - A_{ik}u_k$; Then reset all entries in the k-th row and k-th column of A to 0, and the diagonal entry A_{kk} to 1; Finally, replace the k-th entry b_k of $\mathbf{b}$ by u_k.

This algorithm can be implemented by a pseudo code listed below:

```
for k=1:NG     % loop through all global nodes
    % identify all Dirichlet nodes
    If (gbc(k,1) = 1) then
       for i=1:NG
          b(i) = b(i) - Ag(i,k)*gbc(k,2)
          Ag(i,k) = 0
          Ag(k,i) = 0
       End
       Ag(k,k) = 1
       b(k) = gbc(k,2)
    End
End
```

2.5.2 A MATLAB Code for Q_1 Element

A driver function *ellip_Q1.m* developed for implementing the Q_1 element for solving the elliptic equation (2.37) is shown below:

```
%-----------------------------------------------------
% 2D Q1 FEM for solving
%     -Lap*u + u = f(x,y)
% on a rectangular domain (xlow,xhigh)x(ylow,yhigh)
% with Dirichlet BC condition: u=g  on boundary
%-----------------------------------------------------

clear all;

% readers can change the parameters to
% reset their own rectangualr domain and the mesh size
xlow = 0.0; xhigh = 1.0;
ylow = 0.0; yhigh = 1.0;
nx=20; ny=20;

% Gaussian quadrature points
gauss = [-1/sqrt(3), 1/sqrt(3)];

% generate a Q1 mesh
[x,y,conn,ne,np,gbc] = ...
    getQ1mesh(xlow,xhigh,ylow,yhigh,nx,ny);

% specify an exact solution and use it for Dirichlet BC
for i=1:np
```

```
    uex(i)=sin(pi*x(i)).*cos(pi*y(i));
    if(gbc(i,1)==1)         % find a Dirichlet node
       gbc(i,2) = uex(i);  % assign the BC value
    end
end

% initialize the coefficient matrix and the rhs vector
Ag = zeros(np); bg=zeros(np,1);

nloc = 4;   % number of nodes per element
for ie = 1:ne    % loop over all elements
   rhs= (feval(@SRC,x(conn(ie,1)),y(conn(ie,1)))...
       + feval(@SRC,x(conn(ie,2)),y(conn(ie,2)))...
       + feval(@SRC,x(conn(ie,3)),y(conn(ie,3)))...
       + feval(@SRC,x(conn(ie,4)),y(conn(ie,4))))/nloc;

   [Aloc,rhse] = elemA(conn,x,y,gauss,rhs,ie);
   % assemble local matrices into the global matrix
   for i=1:nloc;
      irow = conn(ie,i);   % global row index
      bg(irow)=bg(irow) + rhse(i);
      for j=1:nloc;
         icol = conn(ie,j);  %global column index
         Ag(irow, icol) = Ag(irow, icol) + Aloc(i,j);
      end;
   end;
end;

% impose the Dirichlet BC
for m=1:np
 if(gbc(m,1)==1)
   for i=1:np
    bg(i) = bg(i) - Ag(i,m) * gbc(m,2);
    Ag(i,m) = 0; Ag(m,i) = 0;
   end
   Ag(m,m) = 1.0; bg(m) = gbc(m,2);
  end
end

%solve the equation
ufem = Ag\bg;

% display the max pointwise error
disp('Max error='), max(ufem-uex'),
```

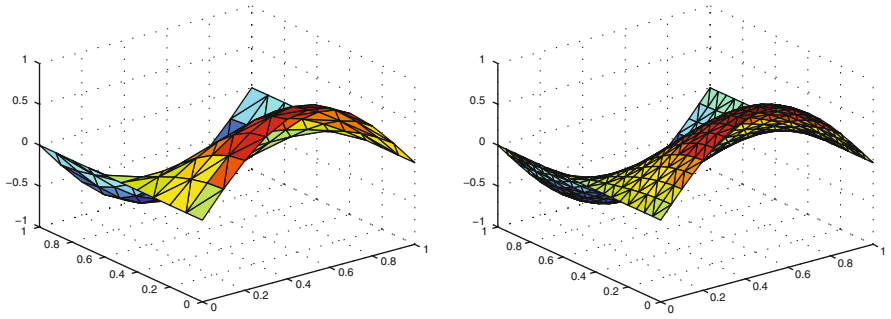

Fig. 2.2 Numerical solutions obtained on $nx = ny = 10$ (*left*), and $nx = ny = 20$ (*right*) grids

```
% plot the FEM solution
tri = delaunay(x,y);
trisurf(tri,x,y,ufem);
```

In this code, we solve the Eq. (2.37) with non-homogenous Dirichlet boundary condition $u = g$ with the exact solution

$$u = \sin \pi x \cos \pi y,$$

which leads to $f = (2\pi^2 + 1)u$.

The problem can be solved with several uniformly refined grids by just changing nx and ny in driver function *ellip_Q1.m*. For example, maximum pointwise errors obtained with $nx = ny = 10, 20, 40$ grids are $0.0084, 0.0021, 5.2583e{-}004$, respectively, which clearly shows the $O(h^2)$ convergence rate in the L^∞-norm. Exemplary numerical solutions obtained with $nx = ny = 10$ and 20 are shown in Fig. 2.2.

Chapter 3
Time-Domain Finite Element Methods for Metamaterials

In this chapter, we present several fully discrete mixed finite element methods for solving Maxwell's equations in metamaterials described by the Drude model and the Lorentz model. In Sects. 3.1 and 3.2, we respectively discuss the constructions of divergence and curl conforming finite elements, and the corresponding interpolation error estimates. These two sections are quite important, since we will use both the divergence and curl conforming finite elements for solving Maxwell's equations in the rest of the book. The material for Sects. 3.1 and 3.2 is quite classic, and we mainly follow the book by Monk (Finite element methods for Maxwell's equations. Oxford Science Publications, New York, 2003). After introducing the basic theory of divergence and curl conforming finite elements, we focus our discussion on developing some finite element methods for solving the time-dependent Maxwell's equations when metamaterials are involved. More specifically, in Sect. 3.3, we discuss the well posedness of the Drude model. Then in Sects. 3.4 and 3.5, we present detailed stability and error analysis for the Crank-Nicolson scheme and the leap-frog scheme, respectively. Finally, we extend our discussion on the well posedness, scheme development and analysis to the Lorentz model and the Drude-Lorentz model in Sects. 3.6 and 3.7, respectively.

3.1 Divergence Conforming Elements

3.1.1 Finite Element on Hexahedra and Rectangles

If a vector function has a continuous normal derivative, then such a finite element is usually called *divergence conforming*. More specifically, similar to the H^1 conforming finite elements discussed in Chap. 2, we can prove the following result.

J. Li and Y. Huang, *Time-Domain Finite Element Methods for Maxwell's Equations in Metamaterials*, Springer Series in Computational Mathematics 43,
DOI 10.1007/978-3-642-33789-5_3, © Springer-Verlag Berlin Heidelberg 2013

Lemma 3.1. *Let K_1 and K_2 be two non-overlapping Lipschitz domains having a common interface Λ such that $\overline{K_1} \cap \overline{K_2} = \Lambda$. Assume that $\mathbf{u}_1 \in H(div; K_1)$ and $\mathbf{u}_2 \in H(div; K_2)$, and $\mathbf{u} \in (L^2(K_1 \cup K_2 \cup \Lambda))^d$ be defined by*

$$\mathbf{u} = \begin{cases} \mathbf{u}_1 & \text{on } K_1, \\ \mathbf{u}_2 & \text{on } K_2. \end{cases}$$

Then $\mathbf{u}_1 \cdot \mathbf{n} = \mathbf{u}_2 \cdot \mathbf{n}$ on Λ implies that $\mathbf{u} \in H(div; K_1 \cup K_2 \cup \Lambda)$, where $\mathbf{n}$ is the unit normal vector to Λ.

Proof. Suppose that we have a function $\mathbf{u} \in (L^2(K_1 \cup K_2 \cup \Lambda))^d$ defined by $\mathbf{u}|_{K_i} = \mathbf{u}_i, i = 1, 2$, and $\mathbf{u}_1 \cdot \mathbf{n} = \mathbf{u}_2 \cdot \mathbf{n}$ on Λ. To prove that $\mathbf{u} \in H(\text{div}; K_1 \cup K_2 \cup \Lambda)$, we only need to show that $\nabla \cdot \mathbf{u} \in L^2(K_1 \cup K_2 \cup \Lambda)$. For any function $\phi \in C_0^\infty(K_1 \cup K_2 \cup \Lambda)$, using integration by parts, we have

$$\begin{aligned} &\int_{K_1 \cup K_2 \cup \Lambda} \mathbf{u} \cdot \nabla \phi d\mathbf{x} \\ &= -\int_{K_1} \nabla \cdot (\mathbf{u}|_{K_1}) \phi d\mathbf{x} - \int_{K_2} \nabla \cdot (\mathbf{u}|_{K_2}) \phi d\mathbf{x} + \int_\Lambda (\mathbf{u}_1 \cdot \mathbf{n}_1 + \mathbf{u}_2 \cdot \mathbf{n}_2) \phi ds, \end{aligned}$$

where $\mathbf{n}_1$ and $\mathbf{n}_2$ denote the unit outward normals to ∂K_1 and ∂K_2, respectively.

Denote a function v such that $v|_{K_l} = \nabla \cdot (\mathbf{u}|_{K_l}), l = 1, 2$. Using the assumption that $\mathbf{u}_1 \cdot \mathbf{n} = \mathbf{u}_2 \cdot \mathbf{n}$ on Λ, we see that the boundary integral term vanishes. Hence, we have

$$\int_{K_1 \cup K_2 \cup \Lambda} \mathbf{u} \cdot \nabla \phi d\mathbf{x} = -\int_{K_1 \cup K_2 \cup \Lambda} v \phi d\mathbf{x},$$

which shows that $\nabla \cdot \mathbf{u} \in L^2(K_1 \cup K_2 \cup \Lambda)$ by the definition of weak derivative. This concludes our proof. □

Now let us consider a divergence conforming element on a reference hexahedron.

Definition 3.1. For any integer $k \geq 1$, the Nédélec divergence conforming element is defined by the triple:

$$\begin{aligned} &\hat{K} = (0, 1)^3, \\ &P_{\hat{K}} = Q_{k,k-1,k-1} \times Q_{k-1,k,k-1} \times Q_{k-1,k-1,k}, \\ &\Sigma_{\hat{K}} = M_{\hat{f}}(\hat{\mathbf{u}}) \cup M_{\hat{K}}(\hat{\mathbf{u}}), \end{aligned}$$

where $M_{\hat{f}}(\hat{\mathbf{u}})$ is the set of degrees of freedom given on all faces $\hat{f}_i$ of $\hat{K}$, each with the outward normal $\mathbf{n}_i$:

$$M_{\hat{f}}(\hat{\mathbf{u}}) = \{ \int_{\hat{f}_i} \hat{\mathbf{u}} \cdot \hat{\mathbf{n}}_i q dA, \ \forall \ q \in Q_{k-1,k-1}(\hat{f}_i), i = 1, \cdots, 6 \} \tag{3.1}$$

and $M_{\hat{K}}(\hat{\mathbf{u}})$ is the set of degrees of freedom given on the element $\hat{K}$:

$$M_{\hat{K}}(\hat{\mathbf{u}}) = \{\int_{\hat{K}} \hat{\mathbf{u}}\cdot\hat{\mathbf{q}}dV,\ \forall\ \hat{\mathbf{q}} \in Q_{k-2,k-1,k-1}\times Q_{k-1,k-2,k-1}\times Q_{k-1,k-1,k-2}\}. \quad (3.2)$$

First we want to prove that the Nédélec element defined in Definition 3.1 is indeed unisolvent.

Theorem 3.1. *The degrees of freedom (3.1) and (3.2) uniquely determine a vector function* $\hat{\mathbf{u}} \in Q_{k,k-1,k-1} \times Q_{k-1,k,k-1} \times Q_{k-1,k-1,k}$ *on* $\hat{K} = (0,1)^3$.

Proof. Our proof follows [217].

(i) First we show that if all the face degrees of freedom in (3.1) on a face (say $\hat{f}_i$) are zero, then $\hat{\mathbf{u}}\cdot\hat{\mathbf{n}}_i = 0$ on this face. Considering that all faces are parallel to the coordinate axes, on any face $\hat{f}_i$, $\hat{\mathbf{u}}\cdot\hat{\mathbf{n}}_i \in Q_{k-1,k-1}$. Hence choosing $\hat{\mathbf{q}} = \hat{\mathbf{u}}\cdot\hat{\mathbf{n}}_i$ in (3.1) immediately leads to $\hat{\mathbf{u}}\cdot\hat{\mathbf{n}}_i = 0$.

(ii) Now let us consider the unisolvence. Note that the dimension of P_K is $3k^2(k+1)$, which equals the total number of degrees of freedom in $\Sigma_{\hat{K}}$. Hence we just need to prove that vanishing all degrees of freedom for $\hat{\mathbf{u}} \in P_{\hat{K}}$ yields $\hat{\mathbf{u}} = 0$. From Part (i), we know that $\hat{\mathbf{u}}\cdot\hat{\mathbf{n}}_i = 0$ on all faces, which implies that $\hat{\mathbf{u}}$ can be written as

$$\hat{\mathbf{u}} = (\hat{x}_1(1-\hat{x}_1)\hat{r}_1, \hat{x}_2(1-\hat{x}_2)\hat{r}_2, \hat{x}_3(1-\hat{x}_3)\hat{r}_3)^T,$$

where $\hat{r}_1 \in Q_{k-2,k-1,k-1}$, $\hat{r}_2 \in Q_{k-1,k-2,k-1}$, and $\hat{r}_3 \in Q_{k-1,k-1,k-2}$. Choosing $\hat{\mathbf{q}} = \hat{\mathbf{r}} \equiv (\hat{r}_1, \hat{r}_2, \hat{r}_3)^T$ in (3.2) shows that $\hat{\mathbf{r}} = \mathbf{0}$, which completes the proof. □

By trace theorem [2], we have $\hat{\mathbf{u}}|_{\hat{f}} \in (H^{\delta}(\hat{f}))^3 \subset (L^2(\hat{f}))^3$. Hence the degrees of freedom (3.1) and (3.2) are well defined for any $\hat{\mathbf{u}} \in (H^{\frac{1}{2}+\delta}(\hat{f}))^3, \delta > 0$.

After obtaining the basis function on the reference hexahedron $\hat{K}$, we can derive the basis function on a general element K by mapping. To make the degrees of freedom (3.1) and (3.2) invariant, we need the following special transformation

$$\mathbf{u}\circ F_K = \frac{1}{\det(B_K)} B_k\hat{\mathbf{u}}, \quad (3.3)$$

where F_K is the affine mapping defined in (2.18). For technical reasons, we assume that B_K is a diagonal matrix, i.e., the mapped element K has all edges parallel to the coordinate axes. The unit outward normal vector $\mathbf{n}$ to ∂K is obtained by the transformation [217, Eq. (5.21)]:

$$\mathbf{n}\circ F_K = \frac{1}{|B_K^{-T}\hat{\mathbf{n}}|} B_k^{-T}\hat{\mathbf{n}}, \quad (3.4)$$

where $\hat{\mathbf{n}}$ is the unit outward normal vector to $\partial\hat{K}$.

Lemma 3.2. *Suppose that* $det(B_K) > 0$ *and the function* $\mathbf{u}$ *and the normal* $\mathbf{n}$ *on* K *are obtained by the transformations (3.3) and (3.4), respectively. Then the degrees of freedom of* $\mathbf{u}$ *on* K *given by*

$$M_f(\mathbf{u}) = \{\int_{f_i} \mathbf{u}\cdot\mathbf{n}_i q dA,\ \forall\ q \in Q_{k-1,k-1}(f_i),\ i = 1,\cdots,6\}, \tag{3.5}$$

$$M_K(\mathbf{u}) = \{\int_K \mathbf{u}\cdot\mathbf{q} dV,\ \forall\ \mathbf{q}\circ F_K = B_K^{-T}\hat{\mathbf{q}},\ \textit{where}$$
$$\hat{\mathbf{q}} \in Q_{k-2,k-1,k-1}\times Q_{k-1,k-2,k-1}\times Q_{k-1,k-1,k-2}\}, \tag{3.6}$$

are identical to the degrees of freedom for $\hat{\mathbf{u}}$ *on* $\hat{K}$ *given in (3.1) and (3.2).*

Proof. (i) By the transformations (3.3) and (3.4), we have

$$\int_f \mathbf{u}\cdot\mathbf{n} q dA = \int_{\hat{f}} \frac{1}{\det(B_K)|B_K^{-T}\hat{\mathbf{n}}|}\hat{\mathbf{u}}\cdot\hat{\mathbf{n}}\hat{q}\frac{\text{area}(f)}{\text{area}(\hat{f})}d\hat{A} = \int_{\hat{f}}\hat{\mathbf{u}}\cdot\hat{\mathbf{n}}\hat{q}d\hat{A}, \tag{3.7}$$

where in the last step we used the fact that

$$\text{area}(f) = \det(B_K)|B_K^{-T}\hat{\mathbf{n}}|\text{area}(\hat{f}).$$

Equation (3.7) shows that the degrees of freedom $M_f(\mathbf{u})$ is invariant.

(ii) The invariance of $M_K(\mathbf{u})$ is easy to see by noting that

$$\begin{aligned}\int_K \mathbf{u}\cdot\mathbf{q} dV &= \int_{\hat{K}} \frac{1}{\det(B_K)} B_K\hat{\mathbf{u}}\cdot B_K^{-T}\hat{\mathbf{q}}\ \det(B_K)\, d\hat{V}\\ &= \int_{\hat{K}} \hat{\mathbf{u}}B_K^T\cdot B_K^{-T}\hat{\mathbf{q}}d\hat{V} = \int_{\hat{K}}\hat{\mathbf{u}}\cdot\hat{\mathbf{q}}d\hat{V}.\end{aligned}$$

□

Now suppose that we have a regular family of meshes of Ω denoted by T_h, and we form a global set of degrees of freedom by assembling the degrees of freedom from each element K in T_h, i.e.,

$$\Sigma = \cup_{K\in T_h}\Sigma_K.$$

If all neighboring elements match the whole common face (i.e., the face degrees of freedom match), then $\mathbf{u}\cdot\mathbf{n}$ is continuous by the proof of Part (i) in Theorem 3.1. Hence the finite element space W_h obtained by mapping the reference element in Definition 3.1 through transformation (3.3) is divergence conforming, i.e., W_h is a subset of $H(\text{div},\Omega)$. Therefore, we can write W_h explicitly as

$$W_h = \{\mathbf{u}_h \in H(\mathrm{div}, \Omega) : \mathbf{u}_h|_K \in Q_{k,k-1,k-1} \times Q_{k-1,k,k-1} \times Q_{k-1,k-1,k}, \ \forall\ K \in T_h\}. \tag{3.8}$$

Similarly, we can define divergence conforming elements on rectangles.

Definition 3.2. For any integer $k \geq 1$, a divergence conforming element can be defined by the triple:

$$\begin{aligned} \hat{K} &= (0,1)^2, \\ P_{\hat{K}} &= Q_{k,k-1} \times Q_{k-1,k}, \\ \Sigma_{\hat{K}} &= M_{\hat{f}}(\hat{\mathbf{u}}) \cup M_{\hat{K}}(\hat{\mathbf{u}}), \end{aligned}$$

where $M_{\hat{f}}(\hat{\mathbf{u}})$ is the set of degrees of freedom given on all faces $\hat{f}_i$ of $\hat{K}$, each with the outward normal $\mathbf{n}_i$:

$$M_{\hat{f}}(\hat{\mathbf{u}}) = \{\int_{\hat{f}_i} \hat{\mathbf{u}} \cdot \hat{\mathbf{n}}_i \, q \, dA, \ \forall\ q \in P_{k-1}(\hat{f}_i), i = 1, \cdots, 4\} \tag{3.9}$$

and $M_{\hat{K}}(\hat{\mathbf{u}})$ is the set of degrees of freedom given on the element $\hat{K}$:

$$M_{\hat{K}}(\hat{\mathbf{u}}) = \{\int_{\hat{K}} \hat{\mathbf{u}} \cdot \hat{\mathbf{q}} \, dV, \ \forall\ \hat{\mathbf{q}} \in Q_{k-2,k-1} \times Q_{k-1,k-2}\}. \tag{3.10}$$

Using the same technique as in the proof of Theorem 3.1, we can easily prove that the divergence element defined in Definition 3.2 is unisolvent.

Theorem 3.2. *The degrees of freedom (3.9) and (3.10) uniquely determine a vector function* $\hat{\mathbf{u}} \in Q_{k,k-1} \times Q_{k-1,k}$ *on* $\hat{K} = (0,1)^2$.

Below we present two often used divergence conforming elements: one for a cubic element; and one for a rectangular element.

Example 3.1. Choosing $k = 1$ in Definition 3.1, we know that $\mathbf{u}_{\hat{K}} \in Q_{1,0,0} \times Q_{0,1,0} \times Q_{0,0,1}$. Hence we can represent $\mathbf{u}_{\hat{K}}$ as follows:

$$\mathbf{u}_{\hat{K}} = (a_1 + b_1\hat{x}_1, a_2 + b_2\hat{x}_2, a_3 + b_3\hat{x}_3)^T,$$

where the constants can be determined by the six face degrees of freedom of (3.1).

If we label the six faces in the following order: front, right, back, left, bottom and top, then the outward normals are:

$$\begin{aligned} &\mathbf{n}_1 = (0,-1,0)', \ \ \mathbf{n}_2 = (1,0,0)', \ \ \mathbf{n}_3 = (0,1,0)', \\ &\mathbf{n}_4 = (-1,0,0)', \ \ \mathbf{n}_5 = (0,0,-1)', \ \ \mathbf{n}_6 = (0,0,1)', \end{aligned}$$

substituting which into (3.1) gives all the coefficients:

$$a_1 = -\int_{left} \mathbf{u}\cdot\mathbf{n}_4 dA,\ \ b_1 = \int_{right} \mathbf{u}\cdot\mathbf{n}_2 dA + \int_{left} \mathbf{u}\cdot\mathbf{n}_4 dA,$$
$$a_2 = -\int_{front} \mathbf{u}\cdot\mathbf{n}_1 dA,\ \ b_2 = \int_{front} \mathbf{u}\cdot\mathbf{n}_1 dA + \int_{back} \mathbf{u}\cdot\mathbf{n}_3 dA,$$
$$a_3 = -\int_{bottom} \mathbf{u}\cdot\mathbf{n}_5 dA,\ \ b_3 = \int_{bottom} \mathbf{u}\cdot\mathbf{n}_5 dA + \int_{top} \mathbf{u}\cdot\mathbf{n}_6 dA.$$

Hence we can write $\mathbf{u}_{\hat{K}}$ as

$$\begin{aligned}\mathbf{u}_{\hat{K}}(\mathbf{x}) &= \begin{pmatrix} (\hat{x}_1-1)\int_{left}\mathbf{u}\cdot\mathbf{n}_4 dA + \hat{x}_1\int_{right}\mathbf{u}\cdot\mathbf{n}_2 dA \\ (\hat{x}_2-1)\int_{front}\mathbf{u}\cdot\mathbf{n}_1 dA + \hat{x}_2\int_{back}\mathbf{u}\cdot\mathbf{n}_3 dA \\ (\hat{x}_3-1)\int_{bottom}\mathbf{u}\cdot\mathbf{n}_5 dA + \hat{x}_3\int_{top}\mathbf{u}\cdot\mathbf{n}_6 dA \end{pmatrix} \\ &= \text{`}(\int_{left}\mathbf{u}\cdot\mathbf{n}_4 dA)\mathbf{N}_{left}(\mathbf{x}) + (\int_{left}\mathbf{u}\cdot\mathbf{n}_2 dA)\mathbf{N}_{right}(\mathbf{x}) \\ &\quad + (\int_{left}\mathbf{u}\cdot\mathbf{n}_1 dA)\mathbf{N}_{front}(\mathbf{x}) + (\int_{left}\mathbf{u}\cdot\mathbf{n}_3 dA)\mathbf{N}_{back}(\mathbf{x}) \\ &\quad + (\int_{left}\mathbf{u}\cdot\mathbf{n}_5 dA)\mathbf{N}_{bottom}(\mathbf{x}) + (\int_{left}\mathbf{u}\cdot\mathbf{n}_6 dA)\mathbf{N}_{top}(\mathbf{x}),\end{aligned}$$

where the basis functions $\mathbf{N}_{**}$ are as follows:

$$\mathbf{N}_{left} = \begin{pmatrix} \hat{x}_1 - 1 \\ 0 \\ 0 \end{pmatrix},\ \mathbf{N}_{right} = \begin{pmatrix} \hat{x}_1 \\ 0 \\ 0 \end{pmatrix},\ \mathbf{N}_{front} = \begin{pmatrix} 0 \\ \hat{x}_2 - 1 \\ 0 \end{pmatrix},$$

$$\mathbf{N}_{back} = \begin{pmatrix} 0 \\ \hat{x}_2 \\ 0 \end{pmatrix},\ \mathbf{N}_{bottom} = \begin{pmatrix} 0 \\ 0 \\ \hat{x}_3 - 1 \end{pmatrix},\ \mathbf{N}_{top} = \begin{pmatrix} 0 \\ 0 \\ \hat{x}_3 \end{pmatrix}.$$

Example 3.2. For rectangular elements, we can similarly define the divergence conforming finite element space W_h:

$$W_h = \{\mathbf{u}_h \in H(\mathrm{div}, \Omega) :\ \mathbf{u}_h|_K \in Q_{k,k-1} \times Q_{k-1,k},\ \forall\ K \in T_h\}. \tag{3.11}$$

For example, consider a rectangle $K = [x_c - h_x, x_c + h_x] \times [y_c - h_y, y_c + h_y]$. Then a function $\Pi_K^d \mathbf{u} \in W_h$ with $k = 1$ can be expressed as

$$\Pi_K^d \mathbf{u}(x, y) = \sum_{j=1}^{4} (\int_{l_j} \mathbf{u}\cdot\mathbf{n}_j\, dl)\mathbf{N}_j(x, y), \tag{3.12}$$

where l_j denote the four edges of the element K, which start from the bottom edge and are oriented counterclockwise. By satisfying the interpolation condition $\int_{l_j}(\Pi_K^d\mathbf{u}-\mathbf{u})\cdot\mathbf{n}_j\,dl=0$, we can obtain the face element basis functions $\mathbf{N}_j$ as follows:

$$\mathbf{N}_1=\begin{pmatrix}0\\ \dfrac{y-(y_c+h_y)}{4h_xh_y}\end{pmatrix},\quad \mathbf{N}_2=\begin{pmatrix}\dfrac{x-(x_c-h_x)}{4h_xh_y}\\ 0\end{pmatrix},$$

$$\mathbf{N}_3=\begin{pmatrix}0\\ \dfrac{y-(y_c-h_y)}{4h_xh_y}\end{pmatrix},\quad \mathbf{N}_4=\begin{pmatrix}\dfrac{x-(x_c+h_x)}{4h_xh_y}\\ 0\end{pmatrix}.$$

It is easy to check that the basis functions $\mathbf{N}_i$ satisfy the conditions:

$$\int_{l_j}\mathbf{N}_i\cdot\mathbf{n}_j\;dl=\delta_{ij},\quad i,j=1,\cdots,4.$$

3.1.2 Interpolation Error Estimates

From the unisolvence and divergence conforming property proved in last section, we see that with sufficient regularity, there exists a well-defined $H(div;\Omega)$ interpolation operator on K denoted as Π_K^d. For example, if we assume that $\mathbf{u}\in(H^{1/2+\delta}(K))^3,\delta>0$, then there is a unique function

$$\Pi_K^d\mathbf{u}\in Q_{k,k-1,k-1}\times Q_{k-1,k,k-1}\times Q_{k-1,k-1,k}$$

such that

$$M_f(\mathbf{u}-\Pi_K^d\mathbf{u})=0\ \text{ and }\ M_K(\mathbf{u}-\Pi_K^d\mathbf{u})=0,$$

where M_f and M_K are the sets of degrees of freedom in (3.5) and (3.6), respectively. More specifically, this is equivalent to requiring that: For all faces $f_i,i=1,\cdots,6$,

$$\int_{f_i}(\mathbf{u}-\Pi_K^d\mathbf{u})\cdot\mathbf{n}_i\,q\,dA=0,\quad\forall\,q\in Q_{k-1,k-1}(f_i),\tag{3.13}$$

and

$$\int_K(\mathbf{u}-\Pi_K^d\mathbf{u})\cdot\mathbf{q}\,dV=0,\ \forall\,\mathbf{q}\circ F_K=B_K^{-T}\hat{\mathbf{q}},$$
$$\hat{\mathbf{q}}\in Q_{k-2,k-1,k-1}\times Q_{k-1,k-2,k-1}\times Q_{k-1,k-1,k-2}.\tag{3.14}$$

Before we prove the interpolation error estimate, we need to prove the following lemma, which shows that the interpolant on a general element K and the interpolation on the reference element $\hat{K}$ are closely related.

Lemma 3.3. *Suppose that* $\mathbf{u}$ *is sufficiently smooth such that* $\Pi_K^d \mathbf{u}$ *is well defined. Then under transformation (3.3), we have*

$$\widehat{\Pi_K^d \mathbf{u}} = \Pi_{\hat{K}}^d \hat{\mathbf{u}}.$$

Proof. By the definition of operator Π_K^d, we know that

$$M_f(\mathbf{u} - \Pi_K^d \mathbf{u}) = M_K(\mathbf{u} - \Pi_K^d \mathbf{u}) = 0,$$

which, along with the invariance of the degrees of freedom by transformation (3.3), leads to

$$M_{\hat{f}}(\mathbf{u} - \widehat{\Pi_K^d \mathbf{u}}) = M_{\hat{K}}(\mathbf{u} - \widehat{\Pi_K^d \mathbf{u}}) = 0.$$

Then by the unisolvence of the degrees of freedom, we obtain

$$\Pi_{\hat{K}}^d(\mathbf{u} - \widehat{\Pi_K^d \mathbf{u}}) = \Pi_{\hat{K}}^d(\hat{\mathbf{u}} - \widehat{\Pi_K^d \mathbf{u}}) = 0. \tag{3.15}$$

By the unisolvence again, we have $\Pi_{\hat{K}}^d(\widehat{\Pi_K^d \mathbf{u}}) = \widehat{\Pi_K^d \mathbf{u}}$, which together with (3.15) yields $\Pi_{\hat{K}}^d \hat{\mathbf{u}} = \widehat{\Pi_K^d \mathbf{u}}$. □

From the local interpolation operator Π_K^d, we can define a global interpolation operator

$$\Pi_h^d : (H^{\frac{1}{2}+\delta}(\Omega))^3 \to W_h, \forall\ \delta > 0,$$

element-wisely by

$$(\Pi_h^d \mathbf{u})|_K = \Pi_K^d(\mathbf{u}|_K) \quad \text{for each } K \in T_h.$$

The following theorem gives an error estimate for this interpolant.

Theorem 3.3. *Assume that* $0 < \delta < \frac{1}{2}$ *and* T_h *is a regular family of hexahedral meshes on* Ω *with faces aligning with the coordinate axes. Then if* $\mathbf{u} \in (H^s(\Omega))^3, \frac{1}{2} + \delta \le s \le k$, *there is a constant* $C > 0$ *independent of* h *and* $\mathbf{u}$ *such that*

$$||\mathbf{u} - \Pi_h^d \mathbf{u}||_{(L^2(\Omega))^3} \le Ch^s ||\mathbf{u}||_{(H^s(\Omega))^3}, \quad \frac{1}{2} + \delta \le s \le k. \tag{3.16}$$

Proof. For simplicity, here we only prove the result for integer $s = k \ge 1$. Proofs for more general cases can be found in other references (e.g., [5] for $\frac{1}{2} + \delta \le s < 1$).

As usual, we start with a local estimate on one element K. Using (3.3) and Lemma 2.5, we have

$$||\mathbf{u} - \Pi_K^d \mathbf{u}||^2_{(L^2(K))^3} = \int_K |\mathbf{u} - \Pi_K^d \mathbf{u}|^2 dV$$

$$= \int_{\hat{K}} |B_K(\hat{\mathbf{u}} - \widehat{\Pi_K^d \mathbf{u}})|^2 \frac{1}{|\det(B_K)|} d\hat{V}$$
$$\leq \frac{||B_K||^2}{|\det(B_K)|} ||\hat{\mathbf{u}} - \widehat{\Pi_K^d \mathbf{u}}||^2_{(L^2(\hat{K}))^3} \leq \frac{Ch_K^2}{|\det(B_K)|} ||\hat{\mathbf{u}} - \widehat{\Pi_K^d \mathbf{u}}||^2_{(L^2(\hat{K}))^3}.$$

By Lemma 3.3, the fact that

$$(I - \Pi_{\hat{K}}^d)\hat{\mathbf{p}} = 0 \qquad \forall\, \hat{\mathbf{p}} \in (Q_{k-1,k-1,k-1})^3,$$

and the Sobolev Embedding Theorem 2.1, we have

$$||\hat{\mathbf{u}} - \widehat{\Pi_K^d \mathbf{u}}||_{(L^2(\hat{K}))^3} = ||\hat{\mathbf{u}} - \Pi_{\hat{K}}^d \hat{\mathbf{u}}||_{(L^2(\hat{K}))^3}$$
$$= ||(I - \Pi_{\hat{K}}^d)(\hat{\mathbf{u}} + \hat{\mathbf{p}})||_{(L^2(\hat{K}))^3} \leq C\, ||\hat{\mathbf{u}} + \hat{\mathbf{p}}||_{(H^k(\hat{K}))^3}.$$

Using the vector form of Lemma 2.6 to the previous inequality, we have

$$||\hat{\mathbf{u}} - \Pi_{\hat{K}}^d \hat{\mathbf{u}}||_{(L^2(\hat{K}))^3} \leq C \inf_{\hat{\mathbf{p}} \in (Q_{k-1,k-1,k-1})^3} ||\hat{\mathbf{u}} + \hat{\mathbf{p}}||_{(H^k(\hat{K}))^3} \leq C\, |\hat{\mathbf{u}}|_{(H^k(\hat{K}))^3}.$$

Combining the above estimates gives us

$$||\mathbf{u} - \Pi_K^d \mathbf{u}||_{(L^2(K))^3} \leq \frac{Ch_K}{|\det(B_K)|^{1/2}} |\hat{\mathbf{u}}|_{(H^k(\hat{K}))^3}. \tag{3.17}$$

Using (3.3), we have

$$|\hat{\mathbf{u}}|_{(H^k(\hat{K}))^3} = \Big(\int_{\hat{K}} |\hat{\partial}_{\hat{x}}^k \hat{\mathbf{u}}|^2 d\hat{V}\Big)^{1/2}$$
$$= \Big(\int_K |\partial_x^k (B_K^{-1}(det(B_K)\mathbf{u}))\, B_K^k|^2 \frac{dV}{|\det(B_K)|}\Big)^{1/2}$$
$$\leq |\det(B_K)|^{1/2} ||B_K^{-1}|| \cdot ||B_K||^k |\mathbf{u}|_{(H^k(K))^3}.$$

Substituting the previous estimate into (3.17) and using Lemma 2.5, we obtain

$$||\mathbf{u} - \Pi_K^d \mathbf{u}||_{(L^2(K))^3} \leq \frac{Ch_K}{|\det(B_K)|^{1/2}} \frac{|\det(B_K)|^{1/2}}{\rho_K} h_K^k |\mathbf{u}|_{(H^k(K))^3}.$$

Finally, substituting the previous estimate into the identity

$$||\mathbf{u} - \Pi_K^d \mathbf{u}||^2_{(L^2(\Omega))^3} = \sum_{K \in T_h} ||\mathbf{u} - \Pi_K^d \mathbf{u}||^2_{(L^2(K))^3}$$

and using the regularity of the mesh, we complete the proof. □

3.1.3 Finite Elements on Tetrahedra and Triangles

In this section, we will introduce some divergence conforming elements on tetrahedra and triangles. For tetrahedra, we define the space

$$D_k = (P_{k-1})^3 \oplus \tilde{P}_{k-1}\mathbf{x}. \tag{3.18}$$

Recall that $\tilde{P}_{k-1}$ represents the space of homogeneous polynomial of degree $k-1$.

It is easy to check that the dimension of D_k is

$$\begin{aligned} dim(D_k) &= 3 * dim(P_{k-1}) + dim(P_{k-1}) - dim(P_{k-2}) \\ &= 4 \cdot \frac{(k+2)(k+1)k}{3!} - \frac{(k+1)k(k-1)}{3!} = \frac{1}{2}(k+1)k(k+3). \end{aligned}$$

One interesting property about D_k is that $\nabla \cdot D_k \in P_{k-1}$.

Lemma 3.4. *Let D_k be the space defined by (3.18). Then $\nabla \cdot D_k \in P_{k-1}$.*

Proof. We can express any $\mathbf{u} \in D_k$ as $\mathbf{u}(\mathbf{x}) = \mathbf{p}(\mathbf{x}) + q(\mathbf{x})\mathbf{x}$, where $\mathbf{p}(\mathbf{x}) \in (P_{k-1})^3$ and $q(\mathbf{x}) \in \tilde{P}_{k-1}$. Hence

$$\begin{aligned} \nabla \cdot (q(\mathbf{x})\mathbf{x}) &= \partial_{x_1}(qx_1) + \partial_{x_2}(qx_2) + \partial_{x_3}(qx_3) \\ &= \nabla q \cdot \mathbf{x} + 3q = (k-1)q + 3q = (k+2)q, \end{aligned}$$

where we used the fact that $\nabla q \cdot \mathbf{x} = (k-1)q$ for any $q \in \tilde{P}_{k-1}$. Thus $\nabla \cdot (q(\mathbf{x})\mathbf{x}) \in P_{k-1}$, which along with the fact that $\nabla \cdot (P_{k-1})^3 \in P_{k-2}$ concludes the proof. □

Now let us construct a divergence conforming element on a reference tetrahedron $\hat{K}$, which has four vertices as $(0,0,0), (1,0,0), (0,1,0), (0,0,1)$. We assume that the four faces are labelled as that the outward unit normals of the first three faces (i.e., left, front and bottom) are

$$\mathbf{n}_1 = (-1,0,0)', \quad \mathbf{n}_2 = (0,-1,0)', \quad \mathbf{n}_3 = (0,0,-1)'.$$

Definition 3.3. For any integer $k \geq 1$, the divergence conforming element is defined by the triple:

$$\begin{aligned} &\hat{K} = \text{the reference tetrahedron}, \\ &P_{\hat{K}} = D_k, \\ &\Sigma_{\hat{K}} = M_{\hat{f}}(\hat{\mathbf{u}}) \cup M_{\hat{K}}(\hat{\mathbf{u}}), \end{aligned}$$

where $M_{\hat{f}}(\hat{\mathbf{u}})$ and $M_{\hat{K}}(\hat{\mathbf{u}})$ are the degrees of freedom defined as follows:

$$M_{\hat{f}}(\hat{\mathbf{u}}) = \{\int_{\hat{f}_i} \hat{\mathbf{u}} \cdot \hat{\mathbf{n}}_i \hat{q} d\hat{A},\ \forall\ \hat{q} \in P_{k-1}(\hat{f}_i), i = 1, \cdots, 4\}, \tag{3.19}$$

$$M_{\hat{K}}(\hat{\mathbf{u}}) = \{\int_{\hat{K}} \hat{\mathbf{u}} \cdot \hat{\mathbf{q}} d\hat{V},\ \forall\ \hat{\mathbf{q}} \in (P_{k-2})^3\}. \tag{3.20}$$

Below we show that this element is indeed unisolvent.

Theorem 3.4. *The degrees of freedom (3.19) and (3.20) uniquely define a vector function* $\hat{\mathbf{u}} \in D_k$ *on the reference tetrahedron* $\hat{K}$.

Proof. Note that the total number of degrees of freedom in Definition 3.3 is

$$\begin{aligned} 4 * dim(P_{k-1}(f)) + 3 * dim(P_{k-2}) &= 4 \cdot \frac{1}{2}(k+1)k + 3 \cdot \frac{1}{3!}(k+1)k(k-1) \\ &= \frac{1}{2}(k+1)k(k+3), \end{aligned}$$

which is the same as $dim(D_k)$. Hence to prove the unisolvence, we only need to show that vanishing all degrees of freedom for $\hat{\mathbf{u}} \in D_k$ gives $\hat{\mathbf{u}} = 0$. For simplicity, we drop the hat sign in the rest proof.

(i) First we prove that if all degrees of freedom (3.19) on a face vanish, then $\mathbf{u} \cdot \mathbf{n} = 0$ on that face. Since $\mathbf{u} \in D_k$, we can write $\mathbf{u} = \mathbf{p} + q\mathbf{x}$ for some $\mathbf{p} \in (P_{k-1})^3$ and $q \in \tilde{P}_{k-1}$. Assume that face f contains a point $\mathbf{a}$, then for any $\mathbf{x} \in f$, we have $(\mathbf{x} - \mathbf{a}) \cdot \mathbf{n} = 0$, where $\mathbf{n}$ is the unit outward normal to f. Hence, we obtain

$$\mathbf{u} \cdot \mathbf{n} = \mathbf{p} \cdot \mathbf{n} + q\mathbf{x} \cdot \mathbf{n} = \mathbf{p} \cdot \mathbf{n} + q\mathbf{a} \cdot \mathbf{n} \in P_{k-1}.$$

Therefore, choosing $q = \mathbf{u} \cdot \mathbf{n}$ in (3.19) leads to $\mathbf{u} \cdot \mathbf{n} = 0$.

(ii) From (i), we have $\mathbf{u} \cdot \mathbf{n} = 0$ on ∂K. For any $\phi \in P_{k-1}$, using integration by parts and the assumption of vanishing degrees of freedom (3.20), we obtain

$$\int_K \nabla \cdot \mathbf{u}\phi dV = \int_{\partial K} \mathbf{u} \cdot \mathbf{n}\phi dA - \int_K \mathbf{u} \cdot \nabla\phi dV = 0.$$

Choosing $\phi = \nabla \cdot \mathbf{u} \in P_{k-1}$ (by Lemma 3.4) yields $\nabla \cdot \mathbf{u} = 0$.

On the other hand, from proof of Lemma 3.4, for any $\mathbf{u} = \mathbf{p} + q\mathbf{x}$ with some $\mathbf{p} \in (P_{k-1})^3$ and $q \in \tilde{P}_{k-1}$, we have $\nabla \cdot \mathbf{u} = \nabla \cdot \mathbf{p} + (k+2)q$, which leads to $q = -\nabla \cdot \mathbf{p}/(k+2) \in P_{k-2}$. Hence $q = 0$, which yields

$$\mathbf{u} = \mathbf{p} \in (P_{k-1})^3. \tag{3.21}$$

If $k = 1$, (3.21) along with the condition $\mathbf{u} \cdot \mathbf{n} = 0$ on ∂K immediately implies that $\mathbf{u} = 0$.

If $k \geq 2$, then (3.21) and the fact $\mathbf{u} \cdot \mathbf{n} = 0$ on ∂K imply that

$$\mathbf{u} = (x_1 r_1, x_2 r_2, x_3 r_3)^T, \text{ for some } \mathbf{r} = (r_1, r_2, r_3)^T \in (P_{k-2})^3.$$

Choosing $\mathbf{q} = \mathbf{r}$ in the vanishing degrees of freedom (3.20) shows that $\mathbf{r} = 0$, hence $\mathbf{u} = 0$, which concludes our proof. □

The divergence conforming element on a general tetrahedron K can be obtained by mapping the finite element on the reference tetrahedron $\hat{K}$ given by Definition 3.3 through the transformation (3.3).

Lemma 3.5. *Suppose that* $det(B_K) > 0$ *and the function* $\mathbf{u}$ *and the normals* $\mathbf{n}$ *on* K *are obtained by the transformations (3.3) and (3.4). Then the degrees of freedom of* $\mathbf{u}$ *on* K *given by*

$$M_f(\mathbf{u}) = \{\int_{f_i} \mathbf{u} \cdot \mathbf{n}_i q dA, \ \forall\ q \in P_{k-1}(f_i), i = 1, \cdots, 4\}, \tag{3.22}$$

$$M_K(\mathbf{u}) = \{\int_K \mathbf{u} \cdot \mathbf{q} dV, \ \forall\ \mathbf{q} \circ F_K = B_K^{-T} \hat{\mathbf{q}}, \ \hat{\mathbf{q}} \in (P_{k-2})^3\} \tag{3.23}$$

are identical to the degrees of freedom for $\hat{\mathbf{u}}$ *on* $\hat{K}$ *given in (3.19) and (3.20).*

Given a regular family of tetrahedral meshes of Ω denoted as T_h, we can define a finite element space W_h using the degrees of freedom (3.22) and (3.23). From Part (i) in the proof of Theorem 3.4, we know that if two neighboring elements share the same face degrees of freedom, then $\mathbf{u} \cdot \mathbf{n}$ is continuous across the neighboring common face. Hence the finite element space W_h is globally divergence conforming, i.e., W_h is a subset of $H(\text{div}, \Omega)$. Therefore, we can write W_h explicitly as

$$W_h = \{\mathbf{u}_h \in H(\text{div}, \Omega) : \ \mathbf{u}_h|_K \in D_k, \ \forall\ K \in T_h\}. \tag{3.24}$$

By the same technique used for hexahedral element, we can define a global interpolation operator $\Pi_h^d : (H^{1/2+\delta}(\Omega))^3 \to W_h, \delta > 0$. The same interpolation error estimate as Theorem 3.3 holds true, and the proof is exactly the same as that carried out for Theorem 3.3. Details can consult Monk's book [217].

Below we show an example for the divergence conforming element defined in Definition 3.3 on the reference tetrahedron when $k = 1$.

Example 3.3. When $k = 1$, any function $\hat{\mathbf{u}}$ in the divergence conforming element can be expressed as

$$\hat{\mathbf{u}} = \mathbf{a} + b\hat{\mathbf{x}},$$

where the coefficients $\mathbf{a} = (a_1, a_2, a_3)^T$ and b can be determined by the four face degrees of freedom $\int_{f_i} \hat{\mathbf{u}} \cdot \hat{\mathbf{n}}_i dA, i = 1, \cdots, 4$. Note that for our reference tetrahedron, the unit outward normals are given by

$$\hat{\mathbf{n}}_1 = (-1, 0, 0)', \ \hat{\mathbf{n}}_2 = (0, -1, 0)', \ \hat{\mathbf{n}}_3 = (0, 0, -1)', \ \hat{\mathbf{n}}_4 = \frac{1}{\sqrt{3}}(1, 1, 1)'.$$

The condition $\int_{f_1} \hat{\mathbf{u}} \cdot \hat{\mathbf{n}}_1 dA$ gives us

$$\int_{f_1} \hat{\mathbf{u}} \cdot \hat{\mathbf{n}}_1 dA = -\int_{x_1=0, x_2, x_3 \geq 0, x_2+x_3 \leq 1} (a_1 + bx_1) dA = -a_1 \cdot \frac{1}{2},$$

which leads to $a_1 = -2 \int_{f_1} \hat{\mathbf{u}} \cdot \hat{\mathbf{n}}_1 dA$.

By the same arguments, we obtain

$$a_2 = -2 \int_{f_2} \hat{\mathbf{u}} \cdot \hat{\mathbf{n}}_2 dA, \ \ a_3 = -2 \int_{f_3} \hat{\mathbf{u}} \cdot \hat{\mathbf{n}}_3 dA.$$

Note that face f_4 can be expressed by $x_1 + x_2 + x_3 = 1$, and has area $area(f_4) = \frac{\sqrt{3}}{2}$. Hence we have

$$\begin{aligned} \int_{f_4} \hat{\mathbf{u}} \cdot \hat{\mathbf{n}}_4 dA &= \int_{f_4} \frac{1}{\sqrt{3}} (a_1 + bx_1 + a_2 + bx_2 + a_3 + bx_3) dA \\ &= \int_{f_4} \frac{1}{\sqrt{3}} (a_1 + a_2 + a_3 + b) dA = \frac{1}{2}(a_1 + a_2 + a_3 + b), \end{aligned}$$

which leads to

$$b = 2\left(\int_{f_4} \hat{\mathbf{u}} \cdot \hat{\mathbf{n}}_4 dA + \int_{f_1} \hat{\mathbf{u}} \cdot \hat{\mathbf{n}}_1 dA + \int_{f_2} \hat{\mathbf{u}} \cdot \hat{\mathbf{n}}_2 dA + \int_{f_3} \hat{\mathbf{u}} \cdot \hat{\mathbf{n}}_3 dA\right).$$

Hence, the interpolation function on the reference element $\hat{K}$ can be written as

$$\Pi^d_{\hat{K}} \hat{\mathbf{u}} = \sum_{j=1}^{4} \left(\int_{f_j} \hat{\mathbf{u}} \cdot \mathbf{n}_j dA\right) \hat{\mathbf{N}}_j(\hat{\mathbf{x}}),$$

where the basis function $\hat{\mathbf{N}}_j$ are:

$$\hat{\mathbf{N}}_i = 2(\hat{\mathbf{x}} - \mathbf{e}_i), \ i = 1, 2, 3, \ \ \hat{\mathbf{N}}_4 = 2\hat{\mathbf{x}},$$

where $\mathbf{e}_i$ is the opposite vertex of each f_i, i.e., $\mathbf{e}_1 = (1, 0, 0)'$, $\mathbf{e}_2 = (0, 1, 0)'$, and $\mathbf{e}_3 = (0, 0, 1)'$. Moreover, it is easy to check that the basis functions satisfy the conditions $\int_{f_i} \hat{\mathbf{N}}_j \cdot \hat{\mathbf{n}}_i dA = \delta_{ij}, \ \ i, j = 1, 2, 3, 4$.

Example 3.4. Note that for triangular elements, the divergence conforming element space can still be defined using (3.24). The only difference is that we have to change

three to two in the definition of D_k given by (3.18). Below we construct the lowest-order divergence conforming triangular element.

First let us consider a reference triangle $\hat{K}$, which is formed by vertices $\hat{A}_i, i = 1, 2, 3$, where

$$\hat{A}_1 = (0, 0),\ \hat{A}_2 = (1, 0),\ \hat{A}_3 = (0, 1).$$

The unit outward normal vectors are defined as follows:

$$\hat{\mathbf{n}}_1 = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})',\ \hat{\mathbf{n}}_2 = (-1, 0)',\ \hat{\mathbf{n}}_3 = (0, -1)'.$$

The interpolation on the reference element $\hat{K}$ can be written as

$$\Pi_{\hat{K}}^d \mathbf{E} = \sum_{j=1}^{3} (\int_{l_j} \mathbf{E} \cdot \hat{\mathbf{n}}_j \, dl) \hat{\mathbf{N}}_j(\hat{x}, \hat{y}),$$

where the basis function $\hat{\mathbf{N}}_j = (a_1 + b\hat{x}, a_2 + b\hat{y})'$ satisfies the conditions

$$\int_{l_i} \hat{\mathbf{N}}_j \cdot \hat{\mathbf{n}}_i \, dl = \delta_{ij},\ \ i, j = 1, 2, 3.$$

It can be shown that the basis functions $\hat{\mathbf{N}}_j$ are:

$$\hat{\mathbf{N}}_1 = \begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix},\ \hat{\mathbf{N}}_2 = \begin{pmatrix} -1 + \hat{x} \\ \hat{y} \end{pmatrix},\ \hat{\mathbf{N}}_3 = \begin{pmatrix} \hat{x} \\ -1 + \hat{y} \end{pmatrix}.$$

Then for a general triangle K with vertices $A_i = (x_i, y_i),\ i = 1, 2, 3$, we can use the affine mapping $F_K : \hat{\mathbf{x}} \to \mathbf{x}$ defined by

$$\begin{aligned} x &= x_1 + (x_2 - x_1)\hat{x} + (x_3 - x_1)\hat{y}, \\ y &= y_1 + (y_2 - y_1)\hat{x} + (y_3 - y_1)\hat{y}, \end{aligned}$$

to map the reference element $\hat{K}$ to the element K.

Let $|K|$ be the area of K. After some lengthy algebra, we can find the inverse mapping F_K^{-1} of F_K as follows:

$$\begin{aligned} \hat{x} &= 2|K|[(y_3 - y_1)(x - x_1) - (x_3 - x_1)(y - y_1)], \\ \hat{y} &= 2|K|[-(y_2 - y_1)(x - x_1) + (x_2 - x_1)(y - y_1)], \end{aligned}$$

from which we can obtain the basis function on K defined as:

$$\mathbf{N}_i(x, y) = \hat{\mathbf{N}}_i \circ F_K^{-1}, \quad i = 1, 2, 3.$$

3.2 Curl Conforming Elements

3.2.1 Finite Element on Hexahedra and Rectangles

If a vector function has a continuous tangential component, then such a finite element is usually called *curl conforming*. Similar to the H^1 conforming finite elements, we can prove the following result.

Lemma 3.6. *Let K_1 and K_2 be two non-overlapping Lipschitz domains having a common interface Λ such that $\overline{K_1} \cap \overline{K_2} = \Lambda$. Assume that $\mathbf{u}_1 \in H(curl; K_1)$ and $\mathbf{u}_2 \in H(curl; K_2)$, and $\mathbf{u} \in (L^2(K_1 \cup K_2 \cup \Lambda))^3$ be defined by*

$$\mathbf{u} = \begin{cases} \mathbf{u}_1 & \text{on } K_1, \\ \mathbf{u}_2 & \text{on } K_2. \end{cases}$$

Then $\mathbf{u}_1 \times \mathbf{n} = \mathbf{u}_2 \times \mathbf{n}$ on Λ implies that $\mathbf{u} \in H(curl; K_1 \cup K_2 \cup \Lambda)$, where $\mathbf{n}$ is the unit normal vector to Λ.

Proof. The proof can be carried out in exactly the same way as that given for Lemma 3.1 by using the following identity: For any function $\boldsymbol{\phi} \in (C_0^\infty(K_1 \cup K_2 \cup \Lambda))^3$,

$$\begin{aligned} &\int_{K_1 \cup K_2 \cup \Lambda} \mathbf{u} \cdot \nabla \times \boldsymbol{\phi} d\mathbf{x} \\ &= \int_{K_1} \nabla \times \mathbf{u}_1 \cdot \boldsymbol{\phi} d\mathbf{x} + \int_{K_2} \nabla \times \mathbf{u}_2 \cdot \boldsymbol{\phi} d\mathbf{x} + \int_{\Lambda} (\mathbf{u}_1 \times \mathbf{n}_1 + \mathbf{u}_2 \times \mathbf{n}_2) \cdot \boldsymbol{\phi} ds, \end{aligned}$$

where $\mathbf{n}_i$ is the unit outward normal to ∂K_i, and $\mathbf{u}_i = \mathbf{u}|_{K_i}, i = 1, 2$. □

Let us consider the curl conforming elements on a reference hexahedron.

Definition 3.4. For any integer $k \geq 1$, the Nédélec curl conforming element is defined by the triple:

$$\begin{aligned} \hat{K} &= (0, 1)^3, \\ P_{\hat{K}} &= Q_{k-1,k,k} \times Q_{k,k-1,k} \times Q_{k,k,k-1}, \\ \Sigma_{\hat{K}} &= M_{\hat{e}}(\hat{\mathbf{u}}) \cup M_{\hat{f}}(\hat{\mathbf{u}}) \cup M_{\hat{K}}(\hat{\mathbf{u}}), \end{aligned}$$

where $M_{\hat{e}}(\hat{\mathbf{u}})$ is the set of degrees of freedom (DOFs) given on all edges $\hat{e}_i$ of $\hat{K}$, each with the unit tangential vector $\hat{\boldsymbol{\tau}}_i$ in the direction of $\hat{e}_i$:

$$M_{\hat{e}}(\hat{\mathbf{u}}) = \{ \int_{\hat{e}_i} \hat{\mathbf{u}} \cdot \hat{\tau}_i \hat{q} d\hat{s}, \ \forall \ \hat{q} \in P_{k-1}(\hat{e}_i), \ i = 1, \cdots, 12 \}, \tag{3.25}$$

$M_{\hat{f}}(\hat{\mathbf{u}})$ is the set of degrees of freedom given on all faces $\hat{f}_i$ of $\hat{K}$, each with the unit outward normal vector $\mathbf{n}_i$:

$$M_{\hat{f}}(\hat{\mathbf{u}}) = \{\int_{\hat{f}_i} \hat{\mathbf{u}} \times \hat{\mathbf{n}}_i \cdot \hat{\mathbf{q}} d\hat{A},\ \forall\, \hat{\mathbf{q}} \in Q_{k-2,k-1}(\hat{f}_i) \times Q_{k-1,k-2}(\hat{f}_i), i = 1, \cdots, 6\}, \tag{3.26}$$

and $M_{\hat{K}}(\hat{\mathbf{u}})$ is the set of degrees of freedom given on the element $\hat{K}$:

$$M_{\hat{K}}(\hat{\mathbf{u}}) = \{\int_{\hat{K}} \hat{\mathbf{u}} \cdot \hat{\mathbf{q}} d\hat{V},\ \forall\, \hat{\mathbf{q}} \in Q_{k-1,k-2,k-2} \times Q_{k-2,k-1,k-2} \times Q_{k-2,k-2,k-1}\}. \tag{3.27}$$

Hence we have a total of $12k$ edge DOFs, $6 \cdot 2(k-1)k$ face DOFs, and $3 \cdot k(k-1)^2$ element DOFs. It is easy to see that

$$\dim(P_{\hat{K}}) = 12k + 6 \cdot 2(k-1)k + 3 \cdot k(k-1)^2 = 3k(k+1)^2.$$

First we want to prove that the element defined in Definition 3.4 is indeed unisolvent.

Theorem 3.5. *The degrees of freedom (3.25)–(3.27) uniquely determine a vector function* $\mathbf{u} \in Q_{k-1,k,k} \times Q_{k,k-1,k} \times Q_{k,k,k-1}$ *on* $\hat{K} = (0,1)^3$.

Proof. (i) First we show that if all the face degrees of freedom (3.25) and (3.26) on a face (say $\hat{f}_i$) are zero, then $\hat{\mathbf{u}} \times \hat{\mathbf{n}}_i = \mathbf{0}$ on this face. Without loss of generality, let us consider face $\hat{x}_1 = 0$. On this face, noting that $\hat{\mathbf{u}} \times \hat{\mathbf{n}}_1 = -\hat{u}_3\mathbf{j} + \hat{u}_2\mathbf{k}$, hence the tangential components of $\hat{\mathbf{u}}$ on $\hat{x}_1 = 0$ are:

$$\hat{u}_3 \in Q_{k,k-1}(\hat{x}_2, \hat{x}_3), \quad \hat{u}_2 \in Q_{k-1,k}(\hat{x}_2, \hat{x}_3).$$

Thus on every edge of face $\hat{x}_1 = 0$, we have $\hat{\mathbf{u}} \cdot \hat{\tau} \in P_{k-1}$. Then choosing $q = \hat{\mathbf{u}} \cdot \hat{\tau}$ in (3.25) leads to $\hat{\mathbf{u}} \cdot \hat{\tau} = 0$ on each edge of this face.

Furthermore, because $\hat{\mathbf{u}} \cdot \hat{\tau} = 0$ on each edge of face $\hat{x}_1 = 0$, we know that the tangential components of $\hat{\tau}$ on this face can be written as:

$$\hat{u}_2 = \hat{x}_3(1 - \hat{x}_3)\hat{v}_2, \quad \hat{v}_2 \in Q_{k-1,k-2}(\hat{x}_2, \hat{x}_3),$$
$$\hat{u}_3 = \hat{x}_2(1 - \hat{x}_2)\hat{v}_3, \quad \hat{v}_3 \in Q_{k-2,k-1}(\hat{x}_2, \hat{x}_3).$$

Hence on $\hat{x}_1 = 0$, choosing $\hat{\mathbf{q}} = (-\hat{v}_3, \hat{v}_2)$ in (3.26) shows that $\hat{v}_2 = \hat{v}_3 = 0$ on this face, i.e., $\hat{\mathbf{u}} \times \hat{\mathbf{n}} = \mathbf{0}$ on face $\hat{x}_1 = 0$, which proves the curl conformity.

(ii) Now let us consider the unisolvence. Note that the dimension of $P_{\hat{K}}$ is $3k(k+1)^2$, which is same as the total number of degrees of freedom in $\Sigma_{\hat{K}}$. Hence we just need to prove that vanishing all degrees of freedom for $\hat{\mathbf{u}} \in P_{\hat{K}}$ yields $\hat{\mathbf{u}} = 0$. From Part (i), we know that $\hat{\mathbf{u}} \times \hat{\mathbf{n}}_i = \mathbf{0}$ on all faces, which implies that $\hat{\mathbf{u}}$ can be written as

$$\hat{\mathbf{u}} = (\hat{x}_2(1-\hat{x}_2)\hat{x}_3(1-\hat{x}_3)\hat{r}_1, \hat{x}_1(1-\hat{x}_1)\hat{x}_3(1-\hat{x}_3)\hat{r}_2, \hat{x}_1(1-\hat{x}_1)\hat{x}_2(1-\hat{x}_2)\hat{r}_3)^T,$$

where $\hat{r}_1 \in Q_{k-1,k-2,k-2}$, $\hat{r}_2 \in Q_{k-2,k-1,k-2}$, and $\hat{r}_3 \in Q_{k-2,k-2,k-1}$. Choosing $\hat{\mathbf{q}} = \hat{\mathbf{r}} \equiv (\hat{r}_1, \hat{r}_2, \hat{r}_3)^T$ in (3.27) shows that $\hat{\mathbf{r}} = 0$, which completes the proof. □

After obtaining the basis function on the reference element $\hat{K}$, we can derive the basis function on a general element K through mapping. To make the degrees of freedom (3.25)–(3.27) invariant, we need the following special transformation

$$\mathbf{u} \circ F_K = B_k^{-T} \hat{\mathbf{u}}, \tag{3.28}$$

where F_K is the affine mapping defined in (2.18). For technical reasons, we assume that B_K is a diagonal matrix, hence the mapped element K has all edges parallel to the coordinate axes. The unit outward normal vector $\mathbf{n}$ to K is obtained by the transformation (3.4), and the unit tangential vector $\boldsymbol{\tau}$ along edge e of K is given by:

$$\boldsymbol{\tau} = B_K \hat{\boldsymbol{\tau}} / |B_K \hat{\boldsymbol{\tau}}|, \tag{3.29}$$

where $\hat{\boldsymbol{\tau}}$ is a unit tangential vector along edge $\hat{e}$ of $\hat{K}$. Note that (3.29) can be seen as follows: a tangent vector $\hat{\boldsymbol{\tau}} = \hat{\mathbf{x}}_1 - \hat{\mathbf{x}}_2$ is transformed into

$$\mathbf{x}_1 - \mathbf{x}_2 = B_K(\hat{\mathbf{x}}_1 - \hat{\mathbf{x}}_2) = B_K \hat{\boldsymbol{\tau}},$$

normalizing which leads to (3.29).

Lemma 3.7. *Suppose that* $\hat{\mathbf{u}} \in H(curl; \hat{K})$*, and* $\mathbf{u}$ *is mapped from* $\hat{\mathbf{u}}$ *by (3.28). Then* $\mathbf{u} \in H(curl; K)$ *and*

$$\nabla \times \mathbf{u} = \frac{1}{det(B_K)} B_K \hat{\nabla} \times \hat{\mathbf{u}}. \tag{3.30}$$

Proof. From (3.28), we have

$$\hat{u}_i = b_{1i}u_1 + b_{2i}u_2 + b_{3i}u_3, \quad i = 1, 2, 3.$$

From mapping (2.18), we have

$$\begin{aligned}
\frac{\partial}{\partial \hat{x}_1} &= \frac{\partial}{\partial x_1}\frac{\partial x_1}{\partial \hat{x}_1} + \frac{\partial}{\partial x_2}\frac{\partial x_2}{\partial \hat{x}_1} + \frac{\partial}{\partial x_3}\frac{\partial x_3}{\partial \hat{x}_1} = b_{11}\frac{\partial}{\partial x_1} + b_{21}\frac{\partial}{\partial x_2} + b_{31}\frac{\partial}{\partial x_3}, \\
\frac{\partial}{\partial \hat{x}_2} &= \frac{\partial}{\partial x_1}\frac{\partial x_1}{\partial \hat{x}_2} + \frac{\partial}{\partial x_2}\frac{\partial x_2}{\partial \hat{x}_2} + \frac{\partial}{\partial x_3}\frac{\partial x_3}{\partial \hat{x}_2} = b_{12}\frac{\partial}{\partial x_1} + b_{22}\frac{\partial}{\partial x_2} + b_{32}\frac{\partial}{\partial x_3}, \\
\frac{\partial}{\partial \hat{x}_3} &= \frac{\partial}{\partial x_1}\frac{\partial x_1}{\partial \hat{x}_3} + \frac{\partial}{\partial x_2}\frac{\partial x_2}{\partial \hat{x}_3} + \frac{\partial}{\partial x_3}\frac{\partial x_3}{\partial \hat{x}_3} = b_{13}\frac{\partial}{\partial x_1} + b_{23}\frac{\partial}{\partial x_2} + b_{33}\frac{\partial}{\partial x_3},
\end{aligned}$$

using which we obtain the first component of $\hat{\nabla} \times \hat{\mathbf{u}}$ as

$$
\begin{aligned}
&(\hat{\nabla} \times \hat{\mathbf{u}})_1 = \frac{\partial \hat{u}_3}{\partial \hat{x}_2} - \frac{\partial \hat{u}_2}{\partial \hat{x}_3} \\
&= \frac{\partial}{\partial \hat{x}_2}(b_{13}u_1 + b_{23}u_2 + b_{33}u_3) - \frac{\partial}{\partial \hat{x}_3}(b_{12}u_1 + b_{22}u_2 + b_{32}u_3) \\
&= b_{13}(b_{12}\frac{\partial u_1}{\partial x_1} + b_{22}\frac{\partial u_1}{\partial x_2} + b_{32}\frac{\partial u_1}{\partial x_3}) + b_{23}(b_{12}\frac{\partial u_2}{\partial x_1} + b_{22}\frac{\partial u_2}{\partial x_2} + b_{32}\frac{\partial u_2}{\partial x_3}) + \cdots \\
&= (-b_{32}b_{23} + b_{33}b_{22})(\frac{\partial u_3}{\partial x_2} - \frac{\partial u_2}{\partial x_3}) + (-b_{12}b_{33} + b_{13}b_{32})(\frac{\partial u_1}{\partial x_3} - \frac{\partial u_3}{\partial x_1}) \\
&\quad + (b_{23}b_{12} - b_{22}b_{13})(\frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2}) \\
&= det(B_K) \cdot (B_K^{-1}\nabla \times \mathbf{u})_1,
\end{aligned}
$$

where in the last step we used the fact that: The inverse of matrix A can be written as $A^{-1} = \frac{1}{det(A)}C^T$, where C is the matrix of cofactors, i.e., each element c_{ij} of C is the cofactor corresponding to element a_{ij} of A.

By the same technique, we can prove that

$$
\begin{aligned}
(\hat{\nabla} \times \hat{\mathbf{u}})_2 &= \frac{\partial \hat{u}_1}{\partial \hat{x}_3} - \frac{\partial \hat{u}_3}{\partial \hat{x}_1} = det(B_K) \cdot (B_K^{-1}\nabla \times \mathbf{u})_2, \\
(\hat{\nabla} \times \hat{\mathbf{u}})_3 &= \frac{\partial \hat{u}_2}{\partial \hat{x}_1} - \frac{\partial \hat{u}_1}{\partial \hat{x}_2} = det(B_K) \cdot (B_K^{-1}\nabla \times \mathbf{u})_3,
\end{aligned}
$$

which concludes our proof. □

A more general result

$$
(\nabla \times \mathbf{u}) \circ F_K = \frac{1}{det(dF_K)} dF_K \hat{\nabla} \times \hat{\mathbf{u}} \tag{3.31}
$$

holds true [217, Corollary 3.58], where the mapping $F_K : \hat{K} \to K$ is assumed to be continuously differentiable, invertible and surjective, i.e., F_K is not restricted to an affine mapping. Here $dF_K = dF_K(\hat{x})/d\hat{x}$ is the jacobian of the mapping. It is easy to see that for the affine mapping $F_K(\hat{x}) = B_K\hat{x} + b_K$, the jacobian $dF_K = B_K$, and (3.31) reduces to (3.30).

Lemma 3.8. *Suppose that $det(B_K) > 0$, and the function $\mathbf{u}$ and the tangential vector $\boldsymbol{\tau}$ are obtained by the transformations (3.28) and (3.29), respectively. Then the degrees of freedom of $\mathbf{u}$ on K given by*

$$
M_e(\mathbf{u}) = \{\int_{e_i} \mathbf{u} \cdot \boldsymbol{\tau}_i q ds, \ \forall\ q \in P_{k-1}(e_i),\ i = 1, \cdots, 12\},
$$

$$M_f(\mathbf{u}) = \{\int_{f_i} \mathbf{u}\times\mathbf{n}_i\cdot\mathbf{q}dA,$$
$$\forall\ \mathbf{q}\circ F_K = B_K^{-T}\hat{\mathbf{q}},\ \ \hat{\mathbf{q}}\in Q_{k-2,k-1}(\hat{f}_i)\times Q_{k-1,k-2}(\hat{f}_i),\ i=1,\cdots,6\},$$
$$M_K(\mathbf{u}) = \{\int_K \mathbf{u}\cdot\mathbf{q}dV,\ \forall\ \mathbf{q}\circ F_K = \frac{1}{det(B_K)}B_K\hat{\mathbf{q}},$$
$$\hat{\mathbf{q}}\in Q_{k-1,k-2,k-2}\times Q_{k-2,k-1,k-2}\times Q_{k-2,k-2,k-1}\},$$

are identical to the degrees of freedom for $\hat{\mathbf{u}}$ *on* $\hat{K}$ *given in (3.25)–(3.27).*

Proof. (i) By the transformations (3.28) and (3.29), we have

$$\int_e \mathbf{u}\cdot\boldsymbol{\tau}qds = \int_{\hat{e}} B_K^{-T}\hat{\mathbf{u}}\cdot\frac{1}{|B_K\hat{\boldsymbol{\tau}}|}B_K\hat{\boldsymbol{\tau}}\cdot\hat{q}\cdot\frac{ds}{d\hat{s}}d\hat{s} = \int_{\hat{e}}\hat{\mathbf{u}}\cdot\hat{\boldsymbol{\tau}}\hat{q}d\hat{s},$$

which shows that the degrees of freedom $M_e(\mathbf{u})$ are invariant.

(ii) By Green's formula and (3.30), we have

$$\int_{\partial K}\mathbf{n}\times\mathbf{u}\cdot\mathbf{q}dA = \int_K(\nabla\times\mathbf{u}\cdot\mathbf{q}-\mathbf{u}\cdot\nabla\times\mathbf{q})dV$$
$$= \int_{\hat{K}}[\frac{1}{det(B_K)}B_K\hat{\nabla}\times\hat{\mathbf{u}}\cdot B_K^{-T}\hat{\mathbf{q}} - B_K^{-T}\hat{\mathbf{u}}\cdot\frac{1}{det(B_K)}B_K\hat{\nabla}\times\hat{\mathbf{q}}]det(B_K)d\hat{V}$$
$$= \int_{\hat{K}}(\hat{\nabla}\times\hat{\mathbf{u}}\cdot\hat{\mathbf{q}}-\hat{\mathbf{u}}\cdot\hat{\nabla}\times\hat{\mathbf{q}})d\hat{V} = \int_{\partial\hat{K}}\hat{\mathbf{n}}\times\hat{\mathbf{u}}\cdot\hat{\mathbf{q}}d\hat{A},$$

which shows that the degrees of freedom $M_f(\mathbf{u})$ are invariant.

(iii) The invariance of $M_K(\mathbf{u})$ is easy to see by noting that

$$\int_K\mathbf{u}\cdot\mathbf{q}dV = \int_{\hat{K}}B_K^{-T}\hat{\mathbf{u}}\cdot\frac{1}{det(B_K)}B_K\hat{\mathbf{q}}\cdot det(B_K)d\hat{V} = \int_{\hat{K}}\hat{\mathbf{u}}\cdot\hat{\mathbf{q}}d\hat{V}.$$

□

Now suppose that we have a regular family of hexahedral meshes of Ω, denoted as T_h. We can define a curl conforming finite element space V_h on the mesh T_h by assembling the degrees of freedom from each element K in T_h, i.e.,

$$\Sigma = \cup_{K\in T_h}\left(M_e(\mathbf{u})\cup M_f(\mathbf{u})\cup M_K(\mathbf{u})\right).$$

More specifically, we can write V_h explicitly as

$$V_h = \{\mathbf{u}_h\in H(\text{curl};\Omega):\ \mathbf{u}_h|_K\in$$
$$Q_{k-1,k,k}\times Q_{k,k-1,k}\times Q_{k,k,k-1},\ \forall\ K\in T_h\}. \tag{3.32}$$

The curl conforming finite element space (3.32) can be extended similarly to rectangular elements, in which case V_h becomes:

$$V_h = \{\mathbf{u}_h \in H(\text{curl}; \Omega) : \ \mathbf{u}_h|_K \in Q_{k-1,k} \times Q_{k,k-1}, \ \forall \ K \in T_h\}. \tag{3.33}$$

On a reference rectangle $\hat{K} = (0, 1)^2$, the set of DOFs for the curl conforming element is formed by edge DOFs:

$$M_{\hat{e}}(\hat{\mathbf{u}}) = \{\int_{\hat{e}_i} \hat{\mathbf{u}} \cdot \hat{\boldsymbol{\tau}}_i \hat{q} d\hat{s}, \ \forall \ \hat{q} \in P_{k-1}(\hat{e}_i), i = 1, \cdots, 4\},$$

and element DOFs:

$$M_{\hat{K}}(\hat{\mathbf{u}}) = \{\int_{\hat{K}} \hat{\mathbf{u}} \cdot \hat{\mathbf{q}} d\hat{V}, \ \forall \ \hat{\mathbf{q}} \in Q_{k-2,k-1} \times Q_{k-1,k-2}\}.$$

Hence we have a total of $4k$ edge DOFs, and $2 \cdot (k-1)k$ element DOFs, whose summation equals

$$4k + 2 \cdot (k-1)k = 2k(k+1) = \dim(Q_{k-1,k} \times Q_{k,k-1}).$$

The DOFs on general rectangles can be defined similarly as shown in Lemma 3.8. More specifically, we only need the following DOFs:

$$M_e(\mathbf{u}) = \{\int_{e_i} \mathbf{u} \cdot \boldsymbol{\tau}_i q ds, \ \forall \ q \in P_{k-1}(e_i), \ i = 1, \cdots, 4\},$$

$$M_K(\mathbf{u}) = \{\int_K \mathbf{u} \cdot \mathbf{q} dV, \ \forall \ \mathbf{q} \circ F_K = \frac{1}{det(B_K)} B_K \hat{\mathbf{q}}, \ \hat{\mathbf{q}} \in Q_{k-2,k-1} \times Q_{k-1,k-2}\}.$$

Below we present some exemplary curl conforming finite elements.

Example 3.5. Consider a cube $K = (x_c - h_x, x_c + h_x) \times (y_c - h_y, y_c + h_y) \times (z_c - h_z, z_c + h_z)$. The lowest-order curl conforming finite element (i.e., $k = 1$ in Definition 3.4) has $\mathbf{u}_{\hat{K}} \in Q_{0,1,1} \times Q_{1,0,1} \times Q_{1,1,0}$. Hence we can represent $\mathbf{u}_{\hat{K}}$ as follows:

$$\mathbf{u}_{\hat{K}} = ((a_1 + b_1 y)(c_1 + d_1 z), (a_2 + b_2 x)(c_2 + d_2 z), (a_3 + b_3 x)(c_3 + d_3 y))^T,$$

where the constants can be determined by the 12 edge degrees of freedom of (3.25).

The 12 edges are labeled as follows:

$$l_1 : (x_c - h_x, y_c - h_y, z_c - h_z) \to (x_c + h_x, y_c - h_y, z_c - h_z),$$
$$l_2 : (x_c + h_x, y_c - h_y, z_c - h_z) \to (x_c + h_x, y_c + h_y, z_c - h_z),$$
$$l_3 : (x_c - h_x, y_c + h_y, z_c - h_z) \to (x_c + h_x, y_c + h_y, z_c - h_z),$$

$$
\begin{aligned}
&l_4 : (x_c - h_x, y_c - h_y, z_c - h_z) \to (x_c - h_x, y_c + h_y, z_c - h_z),\\
&l_5 : (x_c - h_x, y_c - h_y, z_c + h_z) \to (x_c + h_x, y_c - h_y, z_c + h_z),\\
&l_6 : (x_c + h_x, y_c - h_y, z_c + h_z) \to (x_c + h_x, y_c + h_y, z_c + h_z),\\
&l_7 : (x_c - h_x, y_c + h_y, z_c + h_z) \to (x_c + h_x, y_c + h_y, z_c + h_z),\\
&l_8 : (x_c - h_x, y_c - h_y, z_c + h_z) \to (x_c - h_x, y_c + h_y, z_c + h_z),\\
&l_9 : (x_c + h_x, y_c - h_y, z_c - h_z) \to (x_c + h_x, y_c - h_y, z_c + h_z),\\
&l_{10} : (x_c + h_x, y_c + h_y, z_c - h_z) \to (x_c + h_x, y_c + h_y, z_c + h_z),\\
&l_{11} : (x_c - h_x, y_c + h_y, z_c - h_z) \to (x_c - h_x, y_c + h_y, z_c + h_z),\\
&l_{12} : (x_c - h_x, y_c - h_y, z_c - h_z) \to (x_c - h_x, y_c - h_y, z_c + h_z).
\end{aligned}
$$

For any $\mathbf{E} \in H(\text{curl}; K)$, its curl interpolation $\Pi_K^c \mathbf{E}$ satisfying

$$\int_{l_i} (\mathbf{E} - \Pi_K^c \mathbf{E}) \cdot \boldsymbol{\tau}_i dl = 0, \quad i = 1, \cdots, 12, \tag{3.34}$$

where $\boldsymbol{\tau}_i$ is the corresponding unit tangential vector along each edge l_i.

Using (3.34) and after some algebraic calculations, we obtain

$$\Pi_K^c \mathbf{E}(x, y, z) = \sum_{j=1}^{12} \Big(\int_{l_j} \mathbf{E} \cdot \boldsymbol{\tau}_j dl \Big) \mathbf{N}_j(x, y, z),$$

where the basis functions $\mathbf{N}_j$ are given as follows:

$$
\mathbf{N}_1 = \begin{pmatrix} \dfrac{(y_c + h_y - y)(z_c + h_z - z)}{|K|} \\ 0 \\ 0 \end{pmatrix}, \quad
\mathbf{N}_2 = \begin{pmatrix} 0 \\ \dfrac{(x - x_c + h_x)(z_c + h_z - z)}{|K|} \\ 0 \end{pmatrix},
$$

$$
\mathbf{N}_3 = \begin{pmatrix} \dfrac{(y_c - h_y - y)(z_c + h_z - z)}{|K|} \\ 0 \\ 0 \end{pmatrix}, \quad
\mathbf{N}_4 = \begin{pmatrix} 0 \\ \dfrac{(x - x_c - h_x)(z_c + h_z - z)}{|K|} \\ 0 \end{pmatrix},
$$

where $|K| = 8h_x h_y h_z$ denotes the volume of K.

Other basis functions can be obtained similarly. For example,

$$
\mathbf{N}_5 = \begin{pmatrix} \dfrac{(y_c + h_y - y)(z - z_c + h_z)}{|K|} \\ 0 \\ 0 \end{pmatrix}, \quad
\mathbf{N}_9 = \begin{pmatrix} 0 \\ 0 \\ \dfrac{(x - x_c + h_x)(y_c + h_y - y)}{|K|} \end{pmatrix}.
$$

It is easy to check that the basis functions $\mathbf{N}_j$ satisfy the property

$$\int_{l_i} \mathbf{N}_j \cdot \boldsymbol{\tau}_i dl = \delta_{ij}, \ \ i, j = 1, \cdots 12.$$

Example 3.6. Consider a rectangle $K = [x_c - h_x, x_c + h_x] \times [y_c - h_y, y_c + h_y]$. For the lowest-order edge element $Q_{0,1} \times Q_{1,0}$, the interpolation $\Pi_K^c \mathbf{u}$ of any $\mathbf{u} \in H(\text{curl}; K)$ can be written as

$$\Pi_K^c \mathbf{u}(x, y) = \sum_{j=1}^{4} (\int_{l_j} \mathbf{u} \cdot \boldsymbol{\tau}_j dl) \mathbf{N}_j(x, y), \tag{3.35}$$

where l_j denote the four edges of the element, which start from the bottom and are oriented counterclockwise. Furthermore, $|l_j|$ and $\boldsymbol{\tau}_j$ represent the length of edge l_j and the unit tangent vector along l_j, respectively. The edge element basis functions $\mathbf{N}_j$ are as follows:

$$\mathbf{N}_1 = \begin{pmatrix} \dfrac{(y_c + h_y) - y}{4h_x h_y} \\ 0 \end{pmatrix}, \ \ \mathbf{N}_2 = \begin{pmatrix} 0 \\ \dfrac{x - (x_c - h_x)}{4h_x h_y} \end{pmatrix},$$

$$\mathbf{N}_3 = \begin{pmatrix} \dfrac{(y_c - h_y) - y}{4h_x h_y} \\ 0 \end{pmatrix}, \ \ \mathbf{N}_4 = \begin{pmatrix} 0 \\ \dfrac{x - (x_c + h_x)}{4h_x h_y} \end{pmatrix}.$$

Example 3.7. Consider a rectangle $K = [x_c - h_x, x_c + h_x] \times [y_c - h_y, y_c + h_y]$. For the second-order edge element $Q_{1,2} \times Q_{2,1}$, the interpolation $\Pi_K^c \mathbf{u}$ of any $\mathbf{u} \in H(\text{curl}; K)$ can be obtained by satisfying

$$\int_{l_i} (\mathbf{u} - \Pi_K^c \mathbf{u}) \cdot \boldsymbol{\tau}_i q dl = 0, \ \forall \ q \in P_1(l_i), \ i = 1, \cdots, 4,$$

$$\int_K (\mathbf{u} - \Pi_K^c \mathbf{u}) \cdot \mathbf{q} dx dy = 0, \ \forall \ \mathbf{q} \in Q_{0,1} \times Q_{1,0}.$$

Let us denote the unit tangent vectors $\boldsymbol{\tau}_j$ along l_j be:

$$\boldsymbol{\tau}_1 = (1, 0)', \ \ \boldsymbol{\tau}_2 = (0, 1)', \ \ \boldsymbol{\tau}_3 = (-1, 0)', \ \ \boldsymbol{\tau}_4 = (0, -1)'.$$

After lengthy calculations, we can write the interpolation $\Pi_K^c \mathbf{u}$ as

$$\Pi_K^c \mathbf{u}(x, y) = \sum_{j=1}^{12} c_j \mathbf{N}_j(x, y), \tag{3.36}$$

where the DOFs c_j are

$$c_j = \int_{l_j} \mathbf{u}(x,y)\cdot\boldsymbol{\tau}_j dl, \quad j = 1,\cdots,4,$$

$$c_5 = \int_{l_1} (x-x_c)\mathbf{u}(x,y)\cdot\boldsymbol{\tau}_1 dl, \quad c_6 = \int_{l_2} (y-y_c)\mathbf{u}(x,y)\cdot\boldsymbol{\tau}_2 dl,$$

$$c_7 = \int_{l_3} (x-x_c)\mathbf{u}(x,y)\cdot\boldsymbol{\tau}_3 dl, \quad c_8 = \int_{l_4} (y-y_c)\mathbf{u}(x,y)\cdot\boldsymbol{\tau}_4 dl,$$

$$c_9 = \int_K \mathbf{u}(x,y)\cdot(1,0)'dxdy, \quad c_{10} = \int_K \mathbf{u}(x,y)\cdot(0,1)'dxdy,$$

$$c_{11} = \int_K \mathbf{u}(x,y)\cdot(x-x_c,0)'dxdy, \quad c_{12} = \int_K \mathbf{u}(x,y)\cdot(0,y-y_c)'dxdy,$$

and the basis functions $\mathbf{N}_j$ can be expressed as:

$$\mathbf{N}_1 = \begin{pmatrix} \frac{[3(y-y_c)+h_y][(y-y_c)-h_y]}{8h_x h_y^2} \\ 0 \end{pmatrix}, \quad \mathbf{N}_2 = \begin{pmatrix} 0 \\ \frac{[3(x-x_c)-h_x][(x-x_c)+h_x]}{8h_x^2 h_y} \end{pmatrix},$$

$$\mathbf{N}_3 = \begin{pmatrix} \frac{-[3(y-y_c)-h_y][(y-y_c)+h_y]}{8h_x h_y^2} \\ 0 \end{pmatrix}, \quad \mathbf{N}_4 = \begin{pmatrix} 0 \\ \frac{-[3(x-x_c)+h_x][(x-x_c)-h_x]}{8h_x^2 h_y} \end{pmatrix},$$

$$\mathbf{N}_5 = \begin{pmatrix} \frac{(x-x_c)[3(y-y_c)-3h_y][3(y-y_c)+h_y]}{8h_x^3 h_y^2} \\ 0 \end{pmatrix}, \quad \mathbf{N}_6 = \begin{pmatrix} 0 \\ \frac{(y-y_c)[3(x-x_c)+3h_x][3(x-x_c)-h_x]}{8h_x^2 h_y^3} \end{pmatrix},$$

$$\mathbf{N}_7 = \begin{pmatrix} \frac{-(x-x_c)[3(y-y_c)+3h_y][3(y-y_c)-h_y]}{8h_x^3 h_y^2} \\ 0 \end{pmatrix}, \quad \mathbf{N}_8 = \begin{pmatrix} 0 \\ \frac{-(y-y_c)[3(x-x_c)-3h_x][3(x-x_c)+h_x]}{8h_x^2 h_y^3} \end{pmatrix},$$

$$\mathbf{N}_9 = \begin{pmatrix} \frac{-3[(y-y_c)-h_y][(y-y_c)+h_y]}{8h_x h_y^3} \\ 0 \end{pmatrix}, \quad \mathbf{N}_{10} = \begin{pmatrix} 0 \\ \frac{-3[(x-x_c)-h_x][(x-x_c)+h_x]}{8h_x^3 h_y} \end{pmatrix},$$

$$\mathbf{N}_{11} = \begin{pmatrix} \frac{-9(x-x_c)[(y-y_c)-h_y][(y-y_c)+h_y]}{8h_x^3 h_y^3} \\ 0 \end{pmatrix}, \quad \mathbf{N}_{12} = \begin{pmatrix} 0 \\ \frac{-9(y-y_c)[(x-x_c)+h_x][(x-x_c)-h_x]}{8h_x^3 h_y^3} \end{pmatrix}.$$

3.2.2 Interpolation Error Estimates

With sufficient regularity, there exists a well-defined $H(curl)$ interpolation operator on K denoted as Π_K^c. For example, if we assume that $\mathbf{u}, \nabla\times\mathbf{u} \in (H^{1/2+\delta}(K))^3$, $\delta > 0$, then there is a unique function

$$\Pi_K^c\mathbf{u} \in Q_{k-1,k,k} \times Q_{k,k-1,k} \times Q_{k,k,k-1}$$

such that

$$M_e(\mathbf{u} - \Pi_K^c \mathbf{u}) = 0, \quad M_f(\mathbf{u} - \Pi_K^c \mathbf{u}) = 0 \text{ and } M_K(\mathbf{u} - \Pi_K^c \mathbf{u}) = 0,$$

where M_e, M_f and M_K are the sets of degrees of freedom stated in Lemma 3.8.

Similar to the proof carried out for the $H(div)$ interpolation operator, we can easily prove the following lemma, which shows that the interpolant $\Pi_K^c \mathbf{u}$ on a general element K and the interpolation $\Pi_{\hat{K}}^c \hat{\mathbf{u}}$ on the reference element $\hat{K}$ are closely related.

Lemma 3.9. *Suppose that* $\mathbf{u}$ *is sufficiently smooth such that* $\Pi_K^c \mathbf{u}$ *is well defined. Then under transformation (3.28), we have*

$$\widehat{\Pi_K^c \mathbf{u}} = \Pi_{\hat{K}}^c \hat{\mathbf{u}}.$$

From the local interpolation operator Π_K^c, we can define a global interpolation operator

$$\Pi_h^c : (H^{\frac{1}{2}+\delta}(\Omega))^3 \to W_h, \quad \forall\, \delta > 0,$$

element-wisely by

$$(\Pi_h^c \mathbf{u})|_K = \Pi_K^c(\mathbf{u}|_K) \quad \text{for each } K \in T_h.$$

The following theorem shows that there is a close connection between the curl interpolation and divergence interpolation.

Theorem 3.6. *For the space* W_h *given by (3.8) and* V_h *given by (3.32), we have*

$$\nabla \times V_h \subset W_h.$$

Furthermore, if we assume that $\mathbf{u}$ *is smooth enough such that* $\Pi_h^c \mathbf{u}$ *and* $\Pi_h^d \nabla \times \mathbf{u}$ *are well defined, then we have*

$$\nabla \times \Pi_h^c \mathbf{u} = \Pi_h^d \nabla \times \mathbf{u}. \tag{3.37}$$

Proof. For any $\mathbf{u}_h \in V_h$, it is easy to see that

$$\nabla \times \mathbf{u}_h|_K \in Q_{k,k-1,k-1} \times Q_{k-1,k,k-1} \times Q_{k-1,k-1,k},$$

which leads to $\nabla \times V_h \subset W_h$.

Without loss of generality, we just prove that (3.37) for a reference element K (for simplicity, we drop the hat notation). Noting that $\nabla \times \Pi_h^c \mathbf{u} - \Pi_h^d \nabla \times \mathbf{u} \in W_h \subset H(\text{div}; \Omega)$, hence proof of (3.37) is equivalent to prove that the degrees of freedom given in (3.1) and (3.2) vanish for $\nabla \times \Pi_h^c \mathbf{u} - \Pi_h^d \nabla \times \mathbf{u}$.

(i) Consider a face f of K with face normal $\mathbf{n}$, and let $q \in Q_{k-1,k-1}(f)$. Using (3.5) and integration by parts, we have

$$\int_f (\nabla \times \Pi_K^c \mathbf{u} - \Pi_K^d \nabla \times \mathbf{u}) \cdot \mathbf{n} q dA = \int_f (\nabla \times \Pi_K^c \mathbf{u} - \nabla \times \mathbf{u}) \cdot \mathbf{n} q dA$$

$$= -\int_f \nabla_f \cdot (\mathbf{n} \times (\Pi_K^c \mathbf{u} - \mathbf{u})) q dA$$

$$= \int_f \mathbf{n} \times (\Pi_K^c \mathbf{u} - \mathbf{u}) \cdot \nabla_f q dA - \int_{\partial f} \mathbf{n}_{\partial f} \cdot (\mathbf{n} \times (\Pi_K^c \mathbf{u} - \mathbf{u})) q ds, \qquad (3.38)$$

where $\mathbf{n}_{\partial f}$ is the unit outward normal to ∂f on the plane f. Note that in the second equality we used an identity [217, (3.52)], and ∇_f denotes the surface gradient. The first term in (3.38) actually becomes zero by noting that $\nabla_f q \in Q_{k-2,k-1}(f) \times Q_{k-1,k-2}(f)$. Furthermore, the second term in (3.38) can be rewritten as

$$\int_{\partial f} \mathbf{n}_{\partial f} \cdot (\mathbf{n} \times (\Pi_K^c \mathbf{u} - \mathbf{u})) q ds = \int_{\partial f} (\mathbf{n}_{\partial f} \times \mathbf{n}) \cdot (\Pi_K^c \mathbf{u} - \mathbf{u}) q ds,$$

which vanishes since $q \in P_{k-1}(e)$ on each edge of f.

(ii) Let $\mathbf{q} \in Q_{k-2,k-1,k-1} \times Q_{k-1,k-2,k-1} \times Q_{k-1,k-1,k-2}$. Using (3.6) and integration by parts, we have

$$\int_K (\nabla \times \Pi_K^c \mathbf{u} - \Pi_K^d \nabla \times \mathbf{u}) \cdot \mathbf{q} dV = \int_K (\nabla \times \Pi_K^c \mathbf{u} - \nabla \times \mathbf{u}) \cdot \mathbf{q} dV$$

$$= \int_K (\Pi_K^c \mathbf{u} - \mathbf{u}) \cdot \nabla \times \mathbf{q} dV + \int_{\partial K} (\mathbf{n} \times (\Pi_K^c \mathbf{u} - \mathbf{u})) \cdot \mathbf{q} dA.$$

The right hand side vanishes by using (3.27) and (3.26), and this concludes the proof.

□

Lemma 3.10. *Suppose that* $\mathbf{v}$ *and* $\hat{\mathbf{v}}$ *are related by the transformation (3.28). Then for any* $s \geq 0$*, we have*

$$|\hat{\mathbf{v}}|_{(H^s(\hat{K}))^3} \leq C\,|det(B_K)|^{-1/2} ||B_K||^{s+1} |\mathbf{v}|_{(H^s(\hat{K}))^3},$$

$$|\hat{\nabla} \times \hat{\mathbf{v}}|_{(H^s(\hat{K}))^3} \leq C\,|det(B_K)|^{1/2} ||B_K||^{s-1} |\nabla \times \mathbf{v}|_{(H^s(\hat{K}))^3}.$$

Proof. From $\hat{\mathbf{v}} = B_K^T \mathbf{v} \circ F_K$, we have

$$\frac{\partial^\alpha \hat{\mathbf{v}}}{\partial \hat{\mathbf{x}}^\alpha} = B_K^T \frac{\partial^\alpha}{\partial \hat{\mathbf{x}}^\alpha} (\mathbf{v} \circ F_K) = B_K^T (B_K)^\alpha \frac{\partial^\alpha \mathbf{v}}{\partial \mathbf{x}^\alpha},$$

which leads to

$$||\frac{\partial^\alpha \hat{\mathbf{v}}}{\partial \hat{\mathbf{x}}^\alpha}||_{(L^2(\hat{K}))^3} = (\int_K |B_K^T (B_K)^\alpha \frac{\partial^\alpha \mathbf{v}}{\partial \mathbf{x}^\alpha}|^2 \cdot \frac{1}{det(B_K)} dV)^{1/2}$$
$$\leq C|det(B_K)|^{-1/2}||B_K||^{|\alpha|+1}||\frac{\partial^\alpha \mathbf{v}}{\partial \mathbf{x}^\alpha}||_{(L^2(K))^3},$$

and summing all multi-indices $|\alpha|_1 = s$ completes the proof of the first part.

Using the fact that $\hat{\nabla} \times \hat{\mathbf{v}} = det(B_K) B_K^{-1} \nabla \times \mathbf{v}$, we can prove the second part similarly by noting that

$$||\frac{\partial^\alpha}{\partial \hat{\mathbf{x}}^\alpha}(\hat{\nabla} \times \hat{\mathbf{v}})||_{(L^2(\hat{K}))^3}$$
$$= (\int_K |det(B_K) B_K^{-1} (B_K)^\alpha \frac{\partial^\alpha}{\partial \mathbf{x}^\alpha}(\nabla \times \mathbf{v})|^2 \cdot \frac{1}{det(B_K)} dV)^{1/2}$$
$$\leq C|det(B_K)|^{1/2}||B_K||^{|\alpha|-1}||\frac{\partial^\alpha}{\partial \mathbf{x}^\alpha}(\nabla \times \mathbf{v})||_{(L^2(K))^3}.$$

□

Now we can prove the error estimate for Π_h^c interpolation operator.

Theorem 3.7. *Assume that $0 < \delta < \frac{1}{2}$ and T_h is a regular family of hexahedral meshes on Ω with faces aligning with the coordinate axes. If $\mathbf{u}, \nabla \times \mathbf{u} \in (H^s(\Omega))^3, \frac{1}{2} + \delta \leq s \leq k$, then there is a constant $C > 0$ independent of h and $\mathbf{u}$ such that*

$$||\mathbf{u} - \Pi_h^c \mathbf{u}||_{(L^2(\Omega))^3} + ||\nabla \times (\mathbf{u} - \Pi_h^c \mathbf{u})||_{(L^2(\Omega))^3}$$
$$\leq Ch^s(||\mathbf{u}||_{(H^s(\Omega))^3} + ||\nabla \times \mathbf{u}||_{(H^s(\Omega))^3}), \quad \frac{1}{2} + \delta \leq s \leq k. \tag{3.39}$$

Proof. For simplicity, here we only prove the result for integer $s = k \geq 1$.

As usual, we start with a local estimate on one element K. By (3.28), we have

$$||\mathbf{u} - \Pi_K^c \mathbf{u}||_{(L^2(K))^3} = (\int_K |\mathbf{u} - \Pi_K^c \mathbf{u}|^2 dV)^{1/2}$$
$$= (\int_{\hat{K}} |B_K^{-T}(\hat{\mathbf{u}} - \widehat{\Pi_K^c \mathbf{u}})|^2 |\det(B_K)| d\hat{V})^{1/2}$$
$$\leq |\det(B_K)|^{1/2}||B_K^{-1}||\,||\hat{\mathbf{u}} - \widehat{\Pi_K^c \mathbf{u}}||_{(L^2(\hat{K}))^3}. \tag{3.40}$$

By Lemma 3.9 and the fact that

$$(I - \Pi_{\hat{K}}^c)\hat{\mathbf{p}} = 0 \quad \forall\, \hat{\mathbf{p}} \in (Q_{k-1,k-1,k-1})^3,$$

we have

$$||\hat{\mathbf{u}} - \widehat{\Pi_K^c \mathbf{u}}||_{(L^2(\hat{K}))^3} = ||\hat{\mathbf{u}} - \Pi_{\hat{K}}^c \hat{\mathbf{u}}||_{(L^2(\hat{K}))^3} = ||(I - \Pi_{\hat{K}}^c)(\hat{\mathbf{u}} + \hat{\mathbf{p}})||_{(L^2(\hat{K}))^3}$$
$$\leq C(||\hat{\mathbf{u}} + \hat{\mathbf{p}}||_{(H^k(\hat{K}))^3} + ||\hat{\nabla} \times (\hat{\mathbf{u}} + \hat{\mathbf{p}})||_{(H^k(\hat{K}))^3}). \tag{3.41}$$

Using the fact that [217, (5.12)]: If $\mathbf{v}, \nabla \times \mathbf{v} \in (H^s(K))^3$ for $0 \leq s \leq k$, then

$$\inf_{\phi \in Q^3_{k-1,k-1,k-1}} (||\mathbf{v} + \phi||_{(H^s(K))^3} + ||\nabla \times (\mathbf{v} + \phi)||_{(H^s(K))^3})$$
$$\leq C(|\mathbf{v}|_{(H^s(K))^3} + |\nabla \times \mathbf{v}|_{(H^s(K))^3} + |\nabla \times \mathbf{v}|_{(H^{[s]}(K))^3}),$$

where $[s]$ is the integer part of s, we obtain

$$||\hat{\mathbf{u}} - \Pi_{\hat{K}}^c \hat{\mathbf{u}}||_{(L^2(\hat{K}))^3} \leq C(|\hat{\mathbf{u}}|_{(H^k(K))^3} + |\hat{\nabla} \times \hat{\mathbf{u}}|_{(H^k(K))^3}). \tag{3.42}$$

Substituting (3.41) and (3.42) into (3.40) and using Lemma 3.10, we obtain

$$\begin{aligned} ||\mathbf{u} - \Pi_K^c \mathbf{u}||_{(L^2(K))^3} &\leq |det(B_K)|^{1/2} |B_K^{-1}| \cdot C(|\hat{\mathbf{u}}|_{(H^k(K))^3} + |\hat{\nabla} \times \hat{\mathbf{u}}|_{(H^k(K))^3}) \\ &\leq C\,|det(B_K)|^{1/2} |B_K^{-1}| \cdot (|B_K|^{k+1} |det(B_K)|^{-1/2} |\mathbf{u}|_{(H^k(K))^3} \\ &\quad + |B_K|^{k-1} |det(B_K)|^{1/2} |\nabla \times \mathbf{u}|_{(H^k(K))^3}) \\ &\leq C h_K^k (|\mathbf{u}|_{(H^k(K))^3} + |\nabla \times \mathbf{u}|_{(H^k(K))^3}), \end{aligned}$$

which completes the proof for the L_2 error estimate.

Using (3.37) and Theorem 3.3, we can prove the curl estimate:

$$||\nabla \times (\mathbf{u} - \Pi_K^c \mathbf{u})||_{(L^2(K))^3} = ||(I - \Pi_K^d) \nabla \times \mathbf{u}||_{(L^2(K))^3} \leq C h^k ||\nabla \times \mathbf{u}||_{(H^k(K))^3}.$$

□

3.2.3 *Finite Elements on Tetrahedra and Triangles*

Before we construct a curl conforming finite element on a tetrahedron, we need to define a subspace of homogeneous vector polynomials of degree k denoted by

$$\mathscr{S}_k = \{\mathbf{p} \in (\tilde{P}_k)^3 : \ \mathbf{x} \cdot \mathbf{p} = 0\}. \tag{3.43}$$

Note that $\mathbf{x} \cdot \mathbf{p} \in \tilde{P}_{k+1}$, hence the dimension of $\mathscr{S}_k$ can be calculated as follows:

$$\begin{aligned} dim(\mathscr{S}_k) &= 3dim(\tilde{P}_k) - dim(\tilde{P}_{k+1}) \\ &= 3(dim(P_k) - dim(P_{k-1})) - (dim(P_{k+1}) - dim(P_k)) \end{aligned}$$

$$= 3\left(\frac{(k+3)(k+2)(k+1)}{3!} - \frac{(k+2)(k+1)k}{3!}\right)$$
$$-\left(\frac{(k+4)(k+3)(k+2)}{3!} - \frac{(k+3)(k+2)(k+1)}{3!}\right) = (k+2)k.$$

We need another important polynomial space

$$C_k = (P_{k-1})^3 \oplus \mathscr{S}_k. \tag{3.44}$$

It is easy to check that the dimension of C_K is

$$dim(C_k) = 3dim(P_{k-1}) + dim(\mathscr{S}_k) = \frac{1}{2}(k+3)(k+2)k.$$

Now we can define the curl conforming element on the reference tetrahedron $\hat{K}$ with four vertices: $(0,0,0)$, $(1,0,0)$, $(0,1,0)$ and $(0,0,1)$.

Definition 3.5. For any integer $k \geq 1$, the Nédélec curl conforming element is defined by the triple:

$$\begin{aligned}
&\hat{K} \text{ is the reference tetrahedron,}\\
&P_{\hat{K}} = C_k,\\
&\Sigma_{\hat{K}} = M_{\hat{e}}(\hat{\mathbf{u}}) \cup M_{\hat{f}}(\hat{\mathbf{u}}) \cup M_{\hat{K}}(\hat{\mathbf{u}}),
\end{aligned}$$

where $M_{\hat{e}}(\hat{\mathbf{u}})$, $M_{\hat{f}}(\hat{\mathbf{u}})$ and $M_{\hat{K}}(\hat{\mathbf{u}})$ are the sets of degrees of freedom given on edges of $\hat{K}$, faces of $\hat{K}$, and $\hat{K}$ itself:

$$M_{\hat{e}}(\hat{\mathbf{u}}) = \{\int_{\hat{e}_i} \hat{\mathbf{u}} \cdot \hat{\tau}_i \hat{q} d\hat{s},\ \forall\ \hat{q} \in P_{k-1}(\hat{e}_i),\ i = 1, \cdots, 6\}, \tag{3.45}$$

$$M_{\hat{f}}(\hat{\mathbf{u}}) = \{\frac{1}{area(\hat{f}_i)} \int_{\hat{f}_i} \hat{\mathbf{u}} \cdot \hat{\mathbf{q}} d\hat{A},\ \forall\ \hat{\mathbf{q}} \in (P_{k-2}(\hat{f}_i))^3$$
$$\text{and } \hat{\mathbf{q}} \cdot \hat{\mathbf{n}}_i = 0,\ i = 1, \cdots, 4\}, \tag{3.46}$$

$$M_{\hat{K}}(\hat{\mathbf{u}}) = \{\int_{\hat{K}} \hat{\mathbf{u}} \cdot \hat{\mathbf{q}} d\hat{V},\ \forall\ \hat{\mathbf{q}} \in (P_{k-3}(\hat{K}))^3\}. \tag{3.47}$$

Note that the face degrees of freedom defined by (3.46) look different from the original ones given by Nédélec [222], they are actually equivalent as remarked in Monk [217, p. 129]. Note that any $\hat{\mathbf{q}} \in (P_{k-2}(\hat{f}))^3$ satisfying $\hat{\mathbf{q}} \cdot \hat{\mathbf{n}} = 0$ can be written as $\hat{\mathbf{q}} = (\hat{\mathbf{n}} \times \hat{\mathbf{q}}) \times \hat{\mathbf{n}}$, from which we have

$$\int_{\hat{f}} \hat{\mathbf{u}} \cdot \hat{\mathbf{q}} d\hat{A} = \int_{\hat{f}} \hat{\mathbf{u}} \times \hat{\mathbf{n}} \cdot \hat{\mathbf{q}} \times \hat{\mathbf{n}} d\hat{A},$$

which is equivalent to

$$\int_{\hat{f}} \hat{\mathbf{u}} \times \hat{\mathbf{n}} \cdot \hat{\mathbf{r}} d\hat{A}, \quad \hat{\mathbf{r}} \in (P_{k-2}(\hat{f}))^2,$$

since $\hat{\mathbf{q}} \times \hat{\mathbf{n}} \in (P_{k-2}(\hat{f}))^2$.

Lemma 3.11. *Suppose that* $det(B_K) > 0$ *and the function* $\mathbf{u}$ *and the tangential vector* $\boldsymbol{\tau}$ *are obtained by the transformations (3.28) and (3.29). Then the degrees of freedom of* $\mathbf{u}$ *on* K *given by*

$$\begin{aligned}
M_e(\mathbf{u}) &= \{\int_{e_i} \mathbf{u} \cdot \boldsymbol{\tau}_i q ds, \ \forall\, q \in P_{k-1}(e_i),\ i = 1, \cdots, 6\},\\
M_f(\mathbf{u}) &= \{\int_{f_i} \mathbf{u} \cdot \mathbf{q} dA, \quad \forall\, \mathbf{q} \circ F_K = B_K \hat{\mathbf{q}},\\
&\qquad \hat{\mathbf{q}} \in (P_{k-2}(\hat{f}_i))^3,\ \hat{\mathbf{q}} \cdot \hat{\mathbf{n}}_i = 0,\ i = 1, \cdots, 4\},\\
M_K(\mathbf{u}) &= \{\int_K \mathbf{u} \cdot \mathbf{q} dV,\ \forall\, \mathbf{q} \circ F_K = \frac{1}{det(B_K)} B_K \hat{\mathbf{q}},\ \hat{\mathbf{q}} \in (P_{k-3}(\hat{K}))^3\},
\end{aligned}$$

are identical to the degrees of freedom for $\hat{\mathbf{u}}$ *on* $\hat{K}$ *given in Definition 3.5.*

The proof of this lemma is very similar to that given for Lemma 3.8. Details can be found in [217, Lemma 5.34]. Similarly, the finite element given in Lemma 3.11 is curl conforming and unisolvent. Readers interested in the detailed proof can consult [217, pp. 133–134].

Furthermore, we can construct the global curl conforming finite element space on a tetrahedral mesh T_h of Ω by

$$V_h = \{\mathbf{u} \in H(curl; \Omega) : \ \mathbf{u}|_K \in C_k \ \text{ for all } K \in T_h\}. \tag{3.48}$$

If $\mathbf{u}$ is smooth enough, then on any element $K \in T_h$ we can define the element-wise interpolant $\Pi_K^c \mathbf{u} \in C_k$ satisfying

$$M_e(\mathbf{u} - \Pi_K^c \mathbf{u}) = M_f(\mathbf{u} - \Pi_K^c \mathbf{u}) = M_K(\mathbf{u} - \Pi_K^c \mathbf{u}) = 0. \tag{3.49}$$

Hence we can define the global interpolant $\Pi_h^c \mathbf{u} \in V_h$ element by element:

$$(\Pi_h^c \mathbf{u})|_K = \Pi_K^c(\mathbf{u}|_K) \quad \forall\, K \in T_h.$$

Furthermore, we can prove that the global curl interpolant $\Pi_h^c \mathbf{u}$ and the global divergence interpolant $\Pi_h^d \mathbf{u}$ defined in Sect. 3.1 satisfy the relation [217, Lemma 5.40]:

$$\nabla \times \Pi_h^c \mathbf{u} = \Pi_h^d (\nabla \times \mathbf{u}).$$

Also the same interpolation error estimate stated in Theorem 3.7 holds true. Detailed proof can be found in [217, Theorem 5.41].

The above construction can be extended to triangular elements, in which case (3.43) and (3.44) become as:

$$C_k = (P_{k-1})^2 \oplus \mathscr{S}_k, \quad \mathscr{S}_k = \{\mathbf{p} \in (\tilde{P}_k)^2 : \ \mathbf{x} \cdot \mathbf{p} = 0\}. \tag{3.50}$$

It can be seen that on triangles,

$$\dim(C_k) = 2\dim(P_{k-1}) + \dim(\mathscr{S}_k) = 2 \cdot \frac{(k+1)k}{2} + k = k(k+2).$$

Similar to Definition 3.5, the curl conforming element on a reference triangle $\hat{K}$ can be formed by the following edge and element DOFs:

$$M_{\hat{e}}(\hat{\mathbf{u}}) = \{\int_{\hat{e}_i} \hat{\mathbf{u}} \cdot \hat{\tau}_i \hat{q} d\hat{s}, \ \forall\ \hat{q} \in P_{k-1}(\hat{e}_i), \ i = 1, 2, 3\}, \tag{3.51}$$

$$M_{\hat{K}}(\hat{\mathbf{u}}) = \{\int_{\hat{K}} \hat{\mathbf{u}} \cdot \hat{\mathbf{q}} d\hat{V}, \ \forall\ \hat{\mathbf{q}} \in (P_{k-2}(\hat{K}))^2\}. \tag{3.52}$$

It is easy to see that the total edge DOFs are $3k$, and the total element DOFs are $2 \cdot \frac{k(k-1)}{2}$, whose summation is

$$3k + 2 \cdot \frac{k(k-1)}{2} = k(k+2) = \dim(C_k).$$

Moreover, from (3.50) we easily write the spaces C_1 and C_2 on $\hat{K}$ as:

$$C_1 = \langle \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \ \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \ \begin{pmatrix} \hat{y} \\ -\hat{x} \end{pmatrix} \rangle,$$

$$C_2 = (P_1(\hat{K}))^2 \oplus \langle \begin{pmatrix} \hat{x}\hat{y} \\ -\hat{x}^2 \end{pmatrix}, \ \begin{pmatrix} \hat{y}^2 \\ -\hat{x}\hat{y} \end{pmatrix} \rangle.$$

The Nédélec curl conforming element on general triangles can be obtained through transformations (3.28) and (3.29) and the degrees of freedom given by

$$M_e(\mathbf{u}) = \{\int_{e_i} \mathbf{u} \cdot \boldsymbol{\tau}_i q ds, \ \forall\ q \in P_{k-1}(e_i), \ i = 1, 2, 3\},$$

$$M_K(\mathbf{u}) = \{\int_K \mathbf{u} \cdot \mathbf{q} dV, \ \forall\ \mathbf{q} \circ F_K = \frac{1}{det(B_K)} B_K \hat{\mathbf{q}}, \ \hat{\mathbf{q}} \in (P_{k-2}(\hat{K}))^2\}.$$

Below we present two lowest-order curl conforming elements: one for tetrahedra, and another one for triangles.

Example 3.8. For $k = 1$ in (3.44), Nédélec [222] shows that

$$C_1 = \{\mathbf{u}(\mathbf{x}) = \mathbf{a} + \mathbf{b} \times \mathbf{x}, \quad \text{where } \mathbf{a}, \mathbf{b} \in R^3\},$$

where $\mathbf{a}$ and $\mathbf{b}$ are uniquely determined by the edge degrees of freedom $\int_e \mathbf{u} \times \boldsymbol{\tau}\, ds$ of K. Here K is assumed to be a general non-degenerate tetrahedron formed by vertices A_i, where $i = 1, 2, 3, 4$. From (3.44) and (3.43), we can write the space C_1 as:

$$C_1 = (P_0(\hat{K}))^3 \oplus \langle \begin{pmatrix} 0 \\ -x_3 \\ x_2 \end{pmatrix}, \begin{pmatrix} -x_3 \\ 0 \\ x_1 \end{pmatrix}, \begin{pmatrix} -x_2 \\ x_1 \\ 0 \end{pmatrix} \rangle.$$

To obtain a better form for the basis functions of C_1, we need to use the barycentric coordinate function λ_i corresponding to node A_i. More specifically, if we denote $\lambda_i = \alpha_{i0} + \alpha_{i1}x + \alpha_{i2}y + \alpha_{i3}z$, then $\lambda_i(A_j)$ satisfies

$$\lambda_i(A_j) = \delta_{i,j}, \quad i, j = 1, \cdots, 4, \tag{3.53}$$

which has a unique solution for each λ_i. For example, when $i = 1$, (3.53) can be written as follows:

$$\begin{pmatrix} 1 & x_1 & y_1 & z_1 \\ 1 & x_2 & y_2 & z_2 \\ 1 & x_3 & y_3 & z_3 \\ 1 & x_4 & y_4 & z_4 \end{pmatrix} \begin{pmatrix} \alpha_{10} \\ \alpha_{11} \\ \alpha_{12} \\ \alpha_{13} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

whose coefficient matrix has determinant as six times of the volume of K, and hence the system has a unique solution. With barycentric coordinate function λ_i, it can be shown that the basis function of C_1 with unit integral on an edge formed by vertices A_i and A_j is given by

$$\phi_{i,j} = \lambda_i \nabla \lambda_j - \lambda_j \nabla \lambda_i, \quad i, j = 1, \cdots, 4. \tag{3.54}$$

Note that elements such as C_1 depend on the edge degrees of freedom and are often called *edge elements*. C_1 is also called *Whitney element*, since Whitney [294] introduced this right framework in which to develop a finite element discretization of electromagnetic theory.

Example 3.9. Similarly, we can construct the lowest-order curl conforming element on a general triangle K formed by vertices $A_i, i = 1, 2, 3$. It can be shown that the basis function of C_1 on an edge formed by vertices A_i and A_j is given by

$$\phi_{i,j} = \lambda_i \nabla \lambda_j - \lambda_j \nabla \lambda_i, \quad i, j = 1, \cdots, 3.$$

3.3 Mathematical Analysis of the Drude Model

From Chap. 1, the governing equations used for modeling wave propagation in metamaterials with the Drude model can be written as:

$$\epsilon_0 \frac{\partial \mathbf{E}}{\partial t} = \nabla \times \mathbf{H} - \mathbf{J}, \tag{3.55}$$

$$\mu_0 \frac{\partial \mathbf{H}}{\partial t} = -\nabla \times \mathbf{E} - \mathbf{K}, \tag{3.56}$$

$$\frac{1}{\epsilon_0 \omega_{pe}^2} \frac{\partial \mathbf{J}}{\partial t} + \frac{\Gamma_e}{\epsilon_0 \omega_{pe}^2} \mathbf{J} = \mathbf{E}, \tag{3.57}$$

$$\frac{1}{\mu_0 \omega_{pm}^2} \frac{\partial \mathbf{K}}{\partial t} + \frac{\Gamma_m}{\mu_0 \omega_{pm}^2} \mathbf{K} = \mathbf{H}. \tag{3.58}$$

For simplicity, we assume that the modeling domain is $\Omega \times (0, T)$, where Ω is a bounded Lipschitz polyhedral domain in $\mathscr{R}^3$ with connected boundary $\partial\Omega$. Furthermore, we assume that the boundary of Ω is perfect conducting so that

$$\mathbf{n} \times \mathbf{E} = \mathbf{0} \quad \text{on } \partial\Omega, \tag{3.59}$$

where $\mathbf{n}$ is the unit outward normal to $\partial\Omega$. Also, the initial conditions for (3.55)–(3.58) are assumed to be as follows:

$$\mathbf{E}(\mathbf{x}, 0) = \mathbf{E}_0(\mathbf{x}), \quad \mathbf{H}(\mathbf{x}, 0) = \mathbf{H}_0(\mathbf{x}), \tag{3.60}$$

$$\mathbf{J}(\mathbf{x}, 0) = \mathbf{J}_0(\mathbf{x}), \quad \mathbf{K}(\mathbf{x}, 0) = \mathbf{K}_0(\mathbf{x}), \tag{3.61}$$

where $\mathbf{E}_0(\mathbf{x}), \mathbf{H}_0(\mathbf{x}), \mathbf{J}_0(\mathbf{x})$ and $\mathbf{K}_0(\mathbf{x})$ are some given functions.

First, we can show that the model problem (3.55)–(3.61) is stable.

Lemma 3.12. *The solution* $(\mathbf{E}, \mathbf{H}, \mathbf{J}, \mathbf{K})$ *of the problem (3.55)–(3.61) satisfies the following stability estimate:*

$$\begin{aligned} &\epsilon_0 ||\mathbf{E}(t)||_0^2 + \mu_0 ||\mathbf{H}(t)||_0^2 + \frac{1}{\mu_0 \omega_{pm}^2} ||\mathbf{K}(t)||_0^2 + \frac{1}{\epsilon_0 \omega_{pe}^2} ||\mathbf{J}(t)||_0^2 \\ &\leq \epsilon_0 ||\mathbf{E}(0)||_0^2 + \mu_0 ||\mathbf{H}(0)||_0^2 + \frac{1}{\mu_0 \omega_{pm}^2} ||\mathbf{K}(0)||_0^2 + \frac{1}{\epsilon_0 \omega_{pe}^2} ||\mathbf{J}(0)||_0^2. \end{aligned} \tag{3.62}$$

Proof. Multiplying Eq. (3.55)–(3.58) by $\mathbf{E}, \mathbf{H}, \mathbf{J}, \mathbf{K}$ and integrating over the domain Ω, respectively, adding the resultants together, and using the identity

$$\int_\Omega \nabla \times \mathbf{H} \cdot \mathbf{E} d\mathbf{x} = \int_\Omega \mathbf{H} \cdot \nabla \times \mathbf{E} d\mathbf{x} - \int_{\partial\Omega} \mathbf{H} \cdot \mathbf{n} \times \mathbf{E} ds$$

and the boundary condition (3.59), we obtain

$$\frac{1}{2}\frac{d}{dt}[\epsilon_0||\mathbf{E}(t)||_0^2+\mu_0||\mathbf{H}(t)||_0^2+\frac{1}{\mu_0\omega_{pm}^2}||\mathbf{K}(t)||_0^2+\frac{1}{\epsilon_0\omega_{pe}^2}||\mathbf{J}(t)||_0^2]$$
$$+\frac{\Gamma_m}{\mu_0\omega_{pm}^2}||\mathbf{K}(t)||_0^2+\frac{\Gamma_e}{\epsilon_0\omega_{pe}^2}||\mathbf{J}(t)||_0^2=0,$$

which easily leads to the stability estimate (3.62). □

Now we want to show that the model problem (3.55)–(3.61) exists a unique solution.

Theorem 3.8. *There exists a unique solution* $\mathbf{E}\in H_0(curl;\Omega)$ *and* $\mathbf{H}\in H(curl;\Omega)$ *for the system (3.55)–(3.61).*

Proof. Let us denote the Laplace transform of a function $f(t)$ defined for $t\geq 0$ by $\hat{f}(s)=\int_0^\infty e^{-st}f(t)dt$. Taking the Laplace transform of (3.55)–(3.58), we have

$$\epsilon_0(s\hat{\mathbf{E}}-\mathbf{E}_0)=\nabla\times\hat{\mathbf{H}}-\hat{\mathbf{J}}, \tag{3.63}$$
$$\mu_0(s\hat{\mathbf{H}}-\mathbf{H}_0)=-\nabla\times\hat{\mathbf{E}}-\hat{\mathbf{K}}, \tag{3.64}$$
$$(s+\Gamma_e)\hat{\mathbf{J}}=\mathbf{J}_0+\epsilon_0\omega_{pe}^2\hat{\mathbf{E}}, \tag{3.65}$$
$$(s+\Gamma_m)\hat{\mathbf{K}}=\mathbf{K}_0+\mu_0\omega_{pm}^2\hat{\mathbf{H}}. \tag{3.66}$$

Combining (3.63) with (3.65), we obtain

$$\epsilon_0[s(s+\Gamma_e)+\omega_{pe}^2]\hat{\mathbf{E}}=(s+\Gamma_e)\nabla\times\hat{\mathbf{H}}+\epsilon_0(s+\Gamma_e)\mathbf{E}_0-\mathbf{J}_0. \tag{3.67}$$

Similarly, combining (3.64) with (3.66), we obtain

$$\mu_0[s(s+\Gamma_m)+\omega_{pm}^2]\hat{\mathbf{H}}=\mu_0(s+\Gamma_m)\mathbf{H}_0-(s+\Gamma_m)\nabla\times\hat{\mathbf{E}}-\mathbf{K}_0,$$

whose curl gives

$$\mu_0[s(s+\Gamma_m)+\omega_{pm}^2]\nabla\times\hat{\mathbf{H}}$$
$$=\mu_0(s+\Gamma_m)\nabla\times\mathbf{H}_0-(s+\Gamma_m)\nabla\times\nabla\times\hat{\mathbf{E}}-\nabla\times\mathbf{K}_0. \tag{3.68}$$

Adding the result of (3.67) multiplied by $\mu_0[s(s+\Gamma_m)+\omega_{pm}^2]$ to the result of (3.68) multiplied by $(s+\Gamma_e)$, we have

$$\epsilon_0\mu_0[s(s+\Gamma_e)+2\omega_{pe}^2][s(s+\Gamma_m)+\omega_{pm}^2]\hat{\mathbf{E}}+(s+\Gamma_m)(s+\Gamma_e)\nabla\times\nabla\times\hat{\mathbf{E}}$$
$$=\mu_0[s(s+\Gamma_m)+\omega_{pm}^2][\epsilon_0(s+\Gamma_e)\mathbf{E}_0-\mathbf{J}_0]$$
$$+(s+\Gamma_e)[\mu_0(s+\Gamma_m)\nabla\times\mathbf{H}_0-\nabla\times\mathbf{K}_0]. \tag{3.69}$$

A weak formulation of (3.69) is: Find $\hat{\mathbf{E}} \in H_0(curl;\Omega)$ such that

$$\begin{aligned}
&\epsilon_0\mu_0[s(s+\Gamma_e)+2\omega_{pe}^2][s(s+\Gamma_m)+\omega_{pm}^2](\hat{\mathbf{E}},\boldsymbol{\phi})\\
&+(s+\Gamma_m)(s+\Gamma_e)(\nabla\times\hat{\mathbf{E}},\nabla\times\boldsymbol{\phi})\\
=\ &\mu_0[s(s+\Gamma_m)+\omega_{pm}^2](\epsilon_0(s+\Gamma_e)\mathbf{E}_0-\mathbf{J}_0,\boldsymbol{\phi})\\
&+(s+\Gamma_e)(\mu_0(s+\Gamma_m)\nabla\times\mathbf{H}_0-\nabla\times\mathbf{K}_0,\boldsymbol{\phi})\quad\forall\,\boldsymbol{\phi}\in H_0(curl;\Omega),
\end{aligned}$$

which has a unique solution by the Lax-Milgram lemma. The inverse Laplace transform of the function $\hat{\mathbf{E}}$ is the solution $\mathbf{E}$ of (3.55)–(3.61), and the uniqueness of $\mathbf{E}$ follows from the uniqueness of the Laplace transform.

Existence and uniqueness of solution $\mathbf{H}$ can be proved similarly. □

Finally, we can prove that the electric and magnetic fields also satisfy the Gauss' law if the initial fields are divergence free. More specifically, we have

Lemma 3.13. *Assume that the initial conditions are divergence free, i.e.,*

$$\nabla\cdot(\epsilon_0\mathbf{E}_0)=0,\ \ \nabla\cdot(\mu_0\mathbf{H}_0)=0,\ \ \nabla\cdot\mathbf{J}_0=0,\ \ \nabla\cdot\mathbf{K}_0=0. \tag{3.70}$$

Then for any time $t>0$, we have

$$\nabla\cdot(\epsilon_0\mathbf{E}(t))=0,\ \ \nabla\cdot(\mu_0\mathbf{H}(t))=0,\ \ \nabla\cdot\mathbf{J}(t)=0,\ \ \nabla\cdot\mathbf{K}(t)=0.$$

Proof. Taking the divergence of (3.55), we have

$$\frac{\partial}{\partial t}(\nabla\cdot(\epsilon_0\mathbf{E}))=-\nabla\cdot\mathbf{J}. \tag{3.71}$$

Then taking the divergence of (3.57), we have

$$\frac{\partial}{\partial t}(\nabla\cdot\mathbf{J})+\Gamma_e\nabla\cdot\mathbf{J}=\omega_{pe}^2\nabla\cdot(\epsilon_0\mathbf{E}). \tag{3.72}$$

Substituting (3.71) into (3.72), we obtain a second-order constant coefficient ordinary differential equation

$$(\frac{\partial^2}{\partial t^2}+\Gamma_e\frac{\partial}{\partial t}+\omega_{pe}^2)\nabla\cdot(\epsilon_0\mathbf{E})=0, \tag{3.73}$$

which has initial conditions (from (3.70) and (3.71))

$$\nabla\cdot(\epsilon_0\mathbf{E})(0)=\frac{\partial}{\partial t}(\nabla\cdot(\epsilon_0\mathbf{E}))(0)=0. \tag{3.74}$$

From the basic theory of ordinary differential equation, we know that the problem (3.73) and (3.74) only has zero solution, i.e.,

$$\nabla \cdot (\epsilon_0 \mathbf{E}(t)) = 0,$$

substituting which into (3.71) leads to $\nabla \cdot \mathbf{J}(t) = 0$.

By symmetry, we can prove $\nabla \cdot (\mu_0 \mathbf{H}(t)) = 0$ and $\nabla \cdot \mathbf{K}(t) = 0$. □

3.4 The Crank-Nicolson Scheme for the Drude Model

3.4.1 The Raviart-Thomas-Nédélec Finite Elements

To design a finite element method, we assume that Ω is partitioned by a family of regular tetrahedral (or cubic) meshes T_h with maximum mesh size h. Depending upon the regularity of the solution of the problem, we can use proper order divergence and curl conforming (often called as Raviart-Thomas-Nédélec) tetrahedral elements discussed in Sects. 3.1 and 3.2: For any $l \geq 1$,

$$\mathbf{U}_h = \{\mathbf{u}_h \in H(div; \Omega) : \ \mathbf{u}_h|_K \in (p_{l-1})^3 \oplus \tilde{p}_{l-1}\mathbf{x}, \ \ \forall\ K \in T_h\}, \tag{3.75}$$

$$\mathbf{V}_h = \{\mathbf{v}_h \in H(\text{curl}; \Omega) : \ \mathbf{v}_h|_K \in (p_{l-1})^3 \oplus S_l, \ \ \forall\ K \in T_h\}, \tag{3.76}$$

where the space

$$S_l = \{\mathbf{p} \in (\tilde{p}_l)^3, \quad \mathbf{x} \cdot \mathbf{p} = 0\},$$

or Raviart-Thomas-Nédélec cubic elements:

$$\mathbf{U}_h = \{\mathbf{u}_h \in H(div; \Omega) : \mathbf{u}_h|_K \in Q_{l,l-1,l-1} \times Q_{l-1,l,l-1} \times Q_{l-1,l-1,l}, \forall K \in T_h\},$$

$$\mathbf{V}_h = \{\mathbf{v}_h \in H(\text{curl}; \Omega) : \ \mathbf{v}_h|_K \in Q_{l-1,l,l} \times Q_{l,l-1,l} \times Q_{l,l,l-1}, \ \forall\ K \in T_h\}.$$

Recall that $\tilde{p}_k$ denotes the space of homogeneous polynomials of degree k, and $Q_{i,j,k}$ denotes the space of polynomials whose degrees are less than or equal to i, j, k in variables x, y, z, respectively. To impose the boundary condition $\mathbf{n} \times \mathbf{E} = \mathbf{0}$ on the boundary $\partial\Omega$ of Ω, we introduce a subspace of $\mathbf{V}_h$:

$$\mathbf{V}_h^0 = \{\mathbf{v} \in \mathbf{V}_h : \mathbf{v} \times \mathbf{n} = \mathbf{0} \ \text{ on } \partial\Omega\}.$$

In the error analysis below, we will often use the following fact that

$$\nabla \times \mathbf{V}_h \subset \mathbf{U}_h. \tag{3.77}$$

Also we will need two operators. The first one is the standard $L^2(\Omega)$-projection operator: For any $\mathbf{H} \in (L^2(\Omega))^d$, $P_h\mathbf{H} \in \mathbf{U}_h$ satisfies

$$(P_h\mathbf{H} - \mathbf{H}, \boldsymbol{\psi}_h) = 0, \quad \forall\ \psi_h \in \mathbf{U}_h.$$

Another one is the standard Nédélec interpolation operator Π_h^c mapped from $H(\text{curl}; \Omega)$ to $\mathbf{V}_h$. To simplify the notation, we will just use Π_h for Π_h^c in the rest of this chapter.

Recall that we have the following interpolation error estimate: For any $\mathbf{E} \in H^l(\text{curl}; \Omega)$, $1 \le l$, we have

$$||\mathbf{E} - \Pi_h\mathbf{E}||_0 + ||\nabla \times (\mathbf{E} - \Pi_h\mathbf{E})||_0 \le Ch^l||\mathbf{E}||_{l,\text{curl}}, \tag{3.78}$$

and the projection error estimate:

$$||\mathbf{H} - P_h\mathbf{H}||_0 \le Ch^l||\mathbf{H}||_l, \quad \forall\ \mathbf{H} \in (H^l(\Omega))^d,\ 0 \le l. \tag{3.79}$$

To define a fully discrete scheme, we divide the time interval $[0, T]$ into M uniform subintervals by points $t_k = k\tau$, where $\tau = \frac{T}{M}$ and $k = 0, 1, \cdots, M$.

3.4.2 The Scheme and Its Stability Analysis

Now we can formulate a Crank-Nicolson mixed finite element scheme for (3.55)–(3.58): for $k = 1, 2, \cdots, M$, find $\mathbf{E}_h^k \in \mathbf{V}_h^0, \mathbf{J}_h^k \in \mathbf{V}_h, \mathbf{H}_h^k, \mathbf{K}_h^k \in \mathbf{U}_h$ such that

$$\epsilon_0(\delta_\tau\mathbf{E}_h^k, \boldsymbol{\phi}_h) - (\overline{\mathbf{H}}_h^k, \nabla \times \boldsymbol{\phi}_h) + (\overline{\mathbf{J}}_h^k, \boldsymbol{\phi}_h) = 0, \tag{3.80}$$

$$\mu_0(\delta_\tau\mathbf{H}_h^k, \boldsymbol{\psi}_h) + (\nabla \times \overline{\mathbf{E}}_h^k, \boldsymbol{\psi}_h) + (\overline{\mathbf{K}}_h^k, \boldsymbol{\psi}_h) = 0, \tag{3.81}$$

$$\frac{1}{\epsilon_0\omega_{pe}^2}(\delta_\tau\mathbf{J}_h^k, \tilde{\boldsymbol{\phi}}_h) + \frac{\Gamma_e}{\epsilon_0\omega_{pe}^2}(\overline{\mathbf{J}}_h^k, \tilde{\boldsymbol{\phi}}_h) = (\overline{\mathbf{E}}_h^k, \tilde{\boldsymbol{\phi}}_h), \tag{3.82}$$

$$\frac{1}{\mu_0\omega_{pm}^2}(\delta_\tau\mathbf{K}_h^k, \tilde{\boldsymbol{\psi}}_h) + \frac{\Gamma_m}{\mu_0\omega_{pm}^2}(\overline{\mathbf{K}}_h^k, \tilde{\boldsymbol{\psi}}_h) = (\overline{\mathbf{H}}_h^k, \tilde{\boldsymbol{\psi}}_h), \tag{3.83}$$

for any $\boldsymbol{\phi}_h \in \mathbf{V}_h^0$, $\boldsymbol{\psi}_h \in \mathbf{U}_h$, $\tilde{\boldsymbol{\phi}}_h \in \mathbf{V}_h$, $\tilde{\boldsymbol{\psi}}_h \in \mathbf{U}_h$, subject to the initial conditions

$$\mathbf{E}_h^0(\mathbf{x}) = \Pi_h\mathbf{E}_0(\mathbf{x}), \quad \mathbf{H}_h^0(\mathbf{x}) = P_h\mathbf{H}_0(\mathbf{x}),$$
$$\mathbf{J}_h^0(\mathbf{x}) = \Pi_h\mathbf{J}_0(\mathbf{x}), \quad \mathbf{K}_h^0(\mathbf{x}) = P_h\mathbf{K}_0(\mathbf{x}).$$

In (3.80)–(3.83), we use the central difference and average operators at time lever $k + \frac{1}{2}$:

$$\delta_\tau u^k = (u^k - u^{k-1})/\tau, \quad \overline{u}^k = (u^k + u^{k-1})/2,$$

where $u^k = u(k\tau)$.

First, let us look at the scheme (3.80)–(3.83) carefully. It can be seen that (3.82) and (3.83) are equivalent to

$$\mathbf{J}_h^k = \frac{\epsilon_0\omega_{pe}^2}{2\tau^{-1}+\Gamma_e}(\mathbf{E}_h^k + \mathbf{E}_h^{k-1}) + \frac{2\tau^{-1}-\Gamma_e}{2\tau^{-1}+\Gamma_e}\mathbf{J}_h^{k-1}, \tag{3.84}$$

$$\mathbf{K}_h^k = \frac{\mu_0\omega_{pm}^2}{2\tau^{-1}+\Gamma_m}(\mathbf{H}_h^k + \mathbf{H}_h^{k-1}) + \frac{2\tau^{-1}-\Gamma_m}{2\tau^{-1}+\Gamma_m}\mathbf{K}_h^{k-1}. \tag{3.85}$$

Then substituting (3.84) and (3.85) into (3.80) and (3.81), respectively, we obtain

$$\begin{aligned}(\frac{2\epsilon_0}{\tau} + \frac{\epsilon_0\omega_{pe}^2}{2\tau^{-1}+\Gamma_e})(\mathbf{E}_h^k, \boldsymbol{\phi}_h) - (\mathbf{H}_h^k, \nabla\times\boldsymbol{\phi}_h) &= -\frac{4\tau^{-1}}{2\tau^{-1}+\Gamma_e}(\mathbf{J}_h^{k-1}, \boldsymbol{\phi}_h)\\ +(\frac{2\epsilon_0}{\tau} - \frac{\epsilon_0\omega_{pe}^2}{2\tau^{-1}+\Gamma_e})(\mathbf{E}_h^{k-1}, \boldsymbol{\phi}_h) + (\mathbf{H}_h^{k-1}, \nabla\times\boldsymbol{\phi}_h),\end{aligned} \tag{3.86}$$

$$\begin{aligned}(\frac{2\mu_0}{\tau} + \frac{\mu_0\omega_{pm}^2}{2\tau^{-1}+\Gamma_m})(\mathbf{H}_h^k, \boldsymbol{\psi}_h) + (\nabla\times\mathbf{E}_h^k, \boldsymbol{\psi}_h) &= -\frac{4\tau^{-1}}{2\tau^{-1}+\Gamma_m}(\mathbf{K}_h^{k-1}, \boldsymbol{\psi}_h)\\ +(\frac{2\mu_0}{\tau} - \frac{\mu_0\omega_{pm}^2}{2\tau^{-1}+\Gamma_m})(\mathbf{H}_h^{k-1}, \boldsymbol{\psi}_h) - (\nabla\times\mathbf{E}_h^{k-1}, \boldsymbol{\psi}_h).\end{aligned} \tag{3.87}$$

Hence, to solve the system (3.80)–(3.83) at each time step, we just need to solve the smaller system (3.86) and (3.87) for $\mathbf{E}_h^k$ and $\mathbf{H}_h^k$, then update $\mathbf{J}_h^k$ and $\mathbf{K}_h^k$ using (3.84) and (3.85).

We want to assure that the system (3.86) and (3.87) is invertible.

Lemma 3.14. *At each time step, the system (3.86) and (3.87) is uniquely solvable.*

Proof. Note that the coefficient matrix for the system (3.86) and (3.87) with the vector solution $(\mathbf{E}_h^k, \mathbf{H}_h^k)'$ can be written as

$$Q \equiv \begin{pmatrix} A & -B \\ B' & D \end{pmatrix},$$

where the matrices $A = (\frac{2\epsilon_0}{\tau} + \frac{\epsilon_0\omega_{pe}^2}{2\tau^{-1}+\Gamma_e})(\boldsymbol{\phi}_h, \boldsymbol{\phi}_h)$ and $D = (\frac{2\mu_0}{\tau} + \frac{\mu_0\omega_{pm}^2}{2\tau^{-1}+\Gamma_m})(\boldsymbol{\psi}_h, \boldsymbol{\psi}_h)$ are symmetric positive definite, and the matrix $B = (\boldsymbol{\psi}_h, \nabla\times\boldsymbol{\phi}_h)$. Here $\boldsymbol{\phi}_h$ and $\boldsymbol{\psi}_h$ are arbitrary functions from $\mathbf{V}_h^0$ and $\mathbf{U}_h$, respectively.

It is easy to check that the determinant of Q equals $\det(A)\det(D + B'A^{-1}B)$, which is obviously non-zero. Hence, Q is non-singular, which concludes the proof. □

Finally, we want to show that the scheme (3.80)–(3.83) is unconditionally stable and has a discrete stability similar to the continuous case stated in Lemma 3.12.

Lemma 3.15. *For the solution of (3.80)–(3.83), we have*

$$\epsilon_0||\mathbf{E}_h^k||_0^2+\mu_0||\mathbf{H}_h^k||_0^2+\frac{1}{\epsilon_0\omega_{pe}^2}||\mathbf{J}_h^k||_0^2+\frac{1}{\mu_0\omega_{pm}^2}||\mathbf{K}_h^k||_0^2$$
$$\le \epsilon_0||\mathbf{E}_h^0||_0^2+\mu_0||\mathbf{H}_h^0||_0^2+\frac{1}{\epsilon_0\omega_{pe}^2}||\mathbf{J}_h^0||_0^2+\frac{1}{\mu_0\omega_{pm}^2}||\mathbf{K}_h^0||_0^2.$$

Proof. Choosing $\boldsymbol{\phi}_h=\tau(\mathbf{E}_h^k+\mathbf{E}_h^{k-1}),\boldsymbol{\psi}_h=\tau(\mathbf{H}_h^k+\mathbf{H}_h^{k-1}),\tilde{\boldsymbol{\phi}}_h=\tau(\mathbf{J}_h^k+\mathbf{J}_h^{k-1})$, $\tilde{\boldsymbol{\psi}}_h=\tau(\mathbf{K}_h^k+\mathbf{K}_h^{k-1})$ in (3.80)–(3.83), respectively, and adding the resultants together, we obtain

$$\begin{aligned}&\epsilon_0(||\mathbf{E}_h^k||_0^2-\mathbf{E}_h^{k-1}||_0^2)+\mu_0(||\mathbf{H}_h^k||_0^2-||\mathbf{H}_h^{k-1}||_0^2)\\&+\frac{1}{\epsilon_0\omega_{pe}^2}(||\mathbf{J}_h^k||_0^2-||\mathbf{J}_h^{k-1}||_0^2)+\frac{\tau\Gamma_e}{2\epsilon_0\omega_{pe}^2}||\mathbf{J}_h^k+\mathbf{J}_h^{k-1}||_0^2\\&+\frac{1}{\mu_0\omega_{pm}^2}(||\mathbf{K}_h^k||_0^2-||\mathbf{K}_h^{k-1}||_0^2)+\frac{\tau\Gamma_m}{2\mu_0\omega_{pm}^2}||\mathbf{K}_h^k+\mathbf{K}_h^{k-1}||_0^2=0,\end{aligned}$$

from which it is easy to obtain the following unconditional stability

$$\begin{aligned}&\epsilon_0||\mathbf{E}_h^k||_0^2+\mu_0||\mathbf{H}_h^k||_0^2+\frac{1}{\epsilon_0\omega_{pe}^2}||\mathbf{J}_h^k||_0^2+\frac{1}{\mu_0\omega_{pm}^2}||\mathbf{K}_h^k||_0^2\\&\le \epsilon_0||\mathbf{E}_h^{k-1}||_0^2+\mu_0||\mathbf{H}_h^{k-1}||_0^2+\frac{1}{\epsilon_0\omega_{pe}^2}||\mathbf{J}_h^{k-1}||_0^2+\frac{1}{\mu_0\omega_{pm}^2}||\mathbf{K}_h^{k-1}||_0^2\\&\le\cdots\le \epsilon_0||\mathbf{E}_h^0||_0^2+\mu_0||\mathbf{H}_h^0||_0^2+\frac{1}{\epsilon_0\omega_{pe}^2}||\mathbf{J}_h^0||_0^2+\frac{1}{\mu_0\omega_{pm}^2}||\mathbf{K}_h^0||_0^2.\end{aligned}$$

□

3.4.3 The Optimal Error Estimate

In this section, we shall prove that the Crank-Nicolson scheme (3.80)–(3.83) is optimally convergent. To prove that, we need the following estimates.

Lemma 3.16. *Denote* $\overline{\mathbf{u}}^k=\frac{1}{2}(\mathbf{u}^k+\mathbf{u}^{k-1})$. *Then we have*

$$(i)\quad \|\delta_\tau\mathbf{u}^k\|_0^2=\|\frac{\mathbf{u}^k-\mathbf{u}^{k-1}}{\tau}\|_0^2\le\frac{1}{\tau}\int_{t_{k-1}}^{t_k}\|\mathbf{u}_t(t)\|_0^2dt\quad\forall\,\mathbf{u}\in H^1(0,T;(L^2(\Omega))^3),$$

$$(ii)\quad \|\overline{\mathbf{u}}^k-\frac{1}{\tau}\int_{t_{k-1}}^{t_k}\mathbf{u}(t)dt\|_0^2\le\frac{\tau^3}{4}\int_{t_{k-1}}^{t_k}\|\mathbf{u}_{tt}(t)\|_0^2dt\quad\forall\mathbf{u}\in H^2(0,T;(L^2(\Omega))^3).$$

Proof. (i) The proof follows by squaring the identity

$$\delta_\tau \mathbf{u}^k = \frac{1}{\tau}\int_{t_{k-1}}^{t_k} \mathbf{u}_t(t)dt$$

and using the Cauchy-Schwarz inequality.

(ii) Squaring both sides of the integral identity

$$\frac{1}{2}(\mathbf{u}^k + \mathbf{u}^{k-1}) - \frac{1}{\tau}\int_{t_{k-1}}^{t_k} \mathbf{u}(t)dt = \frac{1}{2\tau}\int_{t_{k-1}}^{t_k} (t - t_{k-1})(t_k - t)\mathbf{u}_{tt}(t)dt, \quad (3.88)$$

we can obtain

$$|\overline{\mathbf{u}}^k - \frac{1}{\tau}\int_{t_{k-1}}^{t_k} \mathbf{u}(t)dt|^2 \le \frac{1}{4\tau^2}(\int_{t_{k-1}}^{t_k} (t - t_{k-1})^2(t_k - t)^2 dt)(\int_{t_{k-1}}^{t_k} |\mathbf{u}_{tt}(t)|^2 dt)$$
$$\le \frac{1}{4}\tau^3 \int_{t_{k-1}}^{t_k} |\mathbf{u}_{tt}(t)|^2 dt,$$

integrating which over Ω concludes the proof.

□

Below we will often use the so-called discrete Gronwall inequality ([243, p. 14], [114]).

Theorem 3.9. *Let $f(t)$ and $g(t)$ be nonnegative functions defined on $t_j = j\tau$, $j = 0, 1, \cdots, M$, and $g(t)$ be non-decreasing. If*

$$(t_k) \le g(t_k) + r\tau \sum_{j=0}^{k-1} f(t_j),$$

where r is a positive constant, then we have

$$f(t_k) \le g(t_k)\exp(kr\tau).$$

Now we can prove the following optimal error estimate for the scheme (3.80)–(3.83).

Theorem 3.10. *Let $(\mathbf{E}^n, \mathbf{H}^n)$ and $(\mathbf{E}_h^n, \mathbf{H}_h^n)$ be the analytic and finite element solutions at time $t = t_n$, respectively. Under the regularity assumptions*

$$\mathbf{H}, \mathbf{K} \in (L^2(0, T; (H^l(\Omega))^3))^3,$$
$$\mathbf{E}, \mathbf{J} \in L^\infty(0, T; H^l(curl; \Omega)), \quad \mathbf{E}_t, \mathbf{J}_t \in (L^2(0, T; H^l(curl; \Omega)))^3,$$
$$\mathbf{H}_{tt}, \mathbf{E}_{tt}, \mathbf{K}_{tt}, \mathbf{J}_{tt}, \nabla \times \mathbf{H}_{tt}, \nabla \times \mathbf{E}_{tt} \in (L^2(0, T; (L^2(\Omega))^3))^3,$$

there exists a constant $C = C(T, \epsilon_0, \mu_0, \omega_{pe}, \omega_{pm}, \Gamma_e, \Gamma_m, \mathbf{E}, \mathbf{H}, \mathbf{K}, \mathbf{L})$, independent of both time step τ and mesh size h, such that

$$\max_{1\le n\le M}(||\mathbf{E}^n-\mathbf{E}_h^n||_0+||\mathbf{H}^n-\mathbf{H}_h^n||_0+||\mathbf{J}^n-\mathbf{J}_h^n||_0+||\mathbf{K}^n-\mathbf{K}_h^n||_0)\le C(\tau^2+h^l),$$

where $l\ge 1$ is the order of basis functions in spaces $\mathbf{U}_h$ and $\mathbf{V}_h$.

Proof. Multiplying (3.55)–(3.58) by $\frac{1}{\tau}\boldsymbol{\phi}_h\in\mathbf{V}_h^0$, $\frac{1}{\tau}\boldsymbol{\psi}_h\in\mathbf{U}_h$, $\frac{1}{\tau}\tilde{\boldsymbol{\phi}}_h\in\mathbf{V}_h$, $\frac{1}{\tau}\tilde{\boldsymbol{\psi}}_h\in\mathbf{U}_h$, respectively, integrating the resultants in time over $I^k=[t_{k-1},t_k]$ and in space over Ω, then using the Stokes' formula

$$\int_\Omega\nabla\times\mathbf{E}\cdot\boldsymbol{\psi}=\int_{\partial\Omega}n\times\mathbf{E}\cdot\boldsymbol{\psi}+\int_\Omega\mathbf{E}\cdot\nabla\times\boldsymbol{\psi},\tag{3.89}$$

we obtain

$$\epsilon_0(\delta_\tau\mathbf{E}^k,\boldsymbol{\phi}_h)-(\frac{1}{\tau}\int_{I^k}\mathbf{H}(s)ds,\nabla\times\boldsymbol{\phi}_h)+(\frac{1}{\tau}\int_{I^k}\mathbf{J}(s)ds,\boldsymbol{\phi}_h)=0,\tag{3.90}$$

$$\mu_0(\delta_\tau\mathbf{H}^k,\boldsymbol{\psi}_h)+(\nabla\times\frac{1}{\tau}\int_{I^k}\mathbf{E}(s)ds,\boldsymbol{\psi}_h)+(\frac{1}{\tau}\int_{I^k}\mathbf{K}(s)ds,\boldsymbol{\psi}_h)=0,\tag{3.91}$$

$$\frac{1}{\epsilon_0\omega_{pe}^2}(\delta_\tau\mathbf{J}^k,\tilde{\boldsymbol{\phi}}_h)+\frac{\Gamma_e}{\epsilon_0\omega_{pe}^2}(\frac{1}{\tau}\int_{I^k}\mathbf{J}(s)ds,\tilde{\boldsymbol{\phi}}_h)=(\frac{1}{\tau}\int_{I^k}\mathbf{E}(s)ds,\tilde{\boldsymbol{\phi}}_h),\tag{3.92}$$

$$\frac{1}{\mu_0\omega_{pm}^2}(\delta_\tau\mathbf{K}^k,\tilde{\boldsymbol{\psi}}_h)+\frac{\Gamma_m}{\mu_0\omega_{pm}^2}(\frac{1}{\tau}\int_{I^k}\mathbf{K}(s)ds,\tilde{\boldsymbol{\psi}}_h)=(\frac{1}{\tau}\int_{I^k}\mathbf{H}(s)ds,\tilde{\boldsymbol{\psi}}_h).\tag{3.93}$$

Denote $\xi_h^k=\Pi_h\mathbf{E}^k-\mathbf{E}_h^k$, $\eta_h^k=P_h\mathbf{H}^k-\mathbf{H}_h^k$, $\tilde{\xi}_h^k=\Pi_h\mathbf{J}^k-\mathbf{J}_h^k$, $\tilde{\eta}_h^k=P_h\mathbf{K}^k-\mathbf{K}_h^k$. Subtracting (3.80)–(3.83) from (3.90)–(3.93), respectively, we obtain the error equations

$$\begin{aligned}(i)\quad&\epsilon_0(\delta_\tau\xi_h^k,\boldsymbol{\phi}_h)-(\overline{\eta}_h^k,\nabla\times\boldsymbol{\phi}_h)=\epsilon_0(\delta_\tau(\Pi_h\mathbf{E}^k-\mathbf{E}^k),\boldsymbol{\phi}_h)\\&-(P_h\overline{\mathbf{H}}^k-\frac{1}{\tau}\int_{I^k}\mathbf{H}(s)ds,\nabla\times\boldsymbol{\phi}_h)-(\frac{1}{\tau}\int_{I^k}\mathbf{J}(s)ds-\overline{\mathbf{J}}_h^k,\boldsymbol{\phi}_h),\\(ii)\quad&\mu_0(\delta_\tau\eta_h^k,\boldsymbol{\psi}_h)+(\nabla\times\overline{\xi}_h^k,\boldsymbol{\psi}_h)=\mu_0(\delta_\tau(P_h\mathbf{H}^k-\mathbf{H}^k),\boldsymbol{\psi}_h)\\&+(\nabla\times(\Pi_h\overline{\mathbf{E}}^k-\frac{1}{\tau}\int_{I^k}\mathbf{E}(s)ds),\boldsymbol{\psi}_h)-(\frac{1}{\tau}\int_{I^k}\mathbf{K}(s)ds-\overline{\mathbf{K}}_h^k,\boldsymbol{\psi}_h),\\(iii)\quad&\frac{1}{\epsilon_0\omega_{pe}^2}(\delta_\tau\tilde{\xi}_h^k,\tilde{\boldsymbol{\phi}}_h)+\frac{\Gamma_e}{\epsilon_0\omega_{pe}^2}(\overline{\tilde{\xi}}_h^k,\tilde{\boldsymbol{\phi}}_h)=\frac{1}{\epsilon_0\omega_{pe}^2}(\delta_\tau(\Pi_h\mathbf{J}^k-\mathbf{J}^k),\tilde{\boldsymbol{\phi}}_h)\\&+\frac{\Gamma_e}{\epsilon_0\omega_{pe}^2}(\Pi_h\overline{\mathbf{J}}^k-\frac{1}{\tau}\int_{I^k}\mathbf{J}(s)ds,\tilde{\boldsymbol{\phi}}_h)+(\frac{1}{\tau}\int_{I^k}\mathbf{E}(s)ds-\overline{\mathbf{E}}_h^k,\tilde{\boldsymbol{\phi}}_h),\\(iv)\quad&\frac{1}{\mu_0\omega_{pm}^2}(\delta_\tau\tilde{\eta}_h^k,\tilde{\boldsymbol{\psi}}_h)+\frac{\Gamma_m}{\mu_0\omega_{pm}^2}(\overline{\tilde{\eta}}_h^k,\tilde{\boldsymbol{\psi}}_h)=\frac{1}{\mu_0\omega_{pm}^2}(\delta_\tau(P_h\mathbf{K}^k-\mathbf{K}^k),\tilde{\boldsymbol{\psi}}_h)\end{aligned}$$

$$+\frac{\Gamma_m}{\mu_0\omega_{pm}^2}(P_h\overline{\mathbf{K}}^k-\frac{1}{\tau}\int_{I^k}\mathbf{K}(s)ds,\tilde{\boldsymbol{\psi}}_h)+(\frac{1}{\tau}\int_{I^k}\mathbf{H}(s)ds-\overline{\mathbf{H}}_h^k,\tilde{\boldsymbol{\psi}}_h).$$

Choosing $\phi_h=\tau(\xi_h^k+\xi_h^{k-1}),\psi_h=\tau(\eta_h^k+\eta_h^{k-1}),\tilde{\phi}_h=\tau(\tilde{\xi}_h^k+\tilde{\xi}_h^{k-1}),\tilde{\psi}_h=\tau(\tilde{\eta}_h^k+\tilde{\eta}_h^{k-1})$ in the above error equations, adding the resultants together, and using the property of operator P_h, we obtain

$$\begin{aligned}
&\epsilon_0(||\xi_h^k||_0^2-||\xi_h^{k-1}||_0^2)+\mu_0(||\eta_h^k||_0^2-||\eta_h^{k-1}||_0^2)\\
&+\frac{1}{\epsilon_0\omega_{pe}^2}(||\tilde{\xi}_h^k||_0^2-||\tilde{\xi}_h^{k-1}||_0^2)+\frac{2\tau\Gamma_e}{\epsilon_0\omega_{pe}^2}||\overline{\tilde{\xi}}_h^{k+1}||_0^2\\
&+\frac{1}{\mu_0\omega_{pm}^2}(||\tilde{\eta}_h^k||_0^2-||\tilde{\eta}_h^{k-1}||_0^2)+\frac{2\tau\Gamma_m}{\mu_0\omega_{pm}^2}||\overline{\tilde{\eta}}_h^{k+1}||_0^2\\
=\;&2\tau\epsilon_0(\delta_\tau(\Pi_h\mathbf{E}^k-\mathbf{E}^k),\overline{\xi}_h^k)-2\tau(\overline{\mathbf{H}}^k-\frac{1}{\tau}\int_{I^k}\mathbf{H}(s)ds,\nabla\times\overline{\xi}_h^k)\\
&-2\tau(\frac{1}{\tau}\int_{I^k}\mathbf{J}(s)ds-\overline{\mathbf{J}}_h^k,\overline{\xi}_h^k)+2\tau(\nabla\times(\Pi_h\overline{\mathbf{E}}^k-\frac{1}{\tau}\int_{I^k}\mathbf{E}(s)ds),\overline{\eta}_h^k)\\
&-2\tau(\frac{1}{\tau}\int_{I^k}\mathbf{K}(s)ds-\overline{\mathbf{K}}_h^k,\overline{\eta}_h^k)+\frac{2\tau}{\epsilon_0\omega_{pe}^2}(\delta_\tau(\Pi_h\mathbf{J}^k-\mathbf{J}^k),\overline{\tilde{\xi}}_h^k)\\
&+\frac{2\tau\Gamma_e}{\epsilon_0\omega_{pe}^2}(\Pi_h\overline{\mathbf{J}}^k-\frac{1}{\tau}\int_{I^k}\mathbf{J}(s)ds,\overline{\tilde{\xi}}_h^k)+2\tau(\frac{1}{\tau}\int_{I^k}\mathbf{E}(s)ds-\overline{\mathbf{E}}_h^k,\overline{\tilde{\xi}}_h^k)\\
&+\frac{2\tau\Gamma_m}{\mu_0\omega_{pm}^2}(\overline{\mathbf{K}}^k-\frac{1}{\tau}\int_{I^k}\mathbf{K}(s)ds,\overline{\tilde{\eta}}_h^k)+2\tau(\frac{1}{\tau}\int_{I^k}\mathbf{H}(s)ds-\overline{\mathbf{H}}_h^k,\overline{\tilde{\eta}}_h^k)\\
=\;&\sum_{i=1}^{10}(Err)_i\,.
\end{aligned}\tag{3.94}$$

Since this is our first error analysis of numerical schemes for solving Maxwell's equations in metamaterials, below we provide detailed estimates of each $(Err)_i$ in (3.94).

Using the Cauchy-Schwarz inequality, the arithmetic-geometric mean inequality

$$(a,b)\le\delta||a||_0^2+\frac{1}{4\delta}||b||_0^2\quad\forall\,\delta>0,\tag{3.95}$$

Lemma 3.16, and the interpolation estimate (3.78), we have

$$\begin{aligned}
Err_1&\le\tau\epsilon_0(2\delta_1||\overline{\xi}_h^k||_0^2+\frac{1}{2\delta_1}||\delta_\tau(\Pi_h\mathbf{E}^k-\mathbf{E}^k)||_0^2)\\
&\le\tau\epsilon_0[\delta_1(||\xi_h^k||_0^2+||\xi_h^{k-1}||_0^2)+\frac{1}{2\delta_1\tau}\int_{I^k}||\partial_t(\Pi_h\mathbf{E}^k-\mathbf{E}^k)||_0^2dt]
\end{aligned}$$

$$\leq \tau\epsilon_0[\delta_1(||\xi_h^k||_0^2 + ||\xi_h^{k-1}||_0^2) + \frac{1}{2\delta_1\tau}\int_{I^k} Ch^{2l}||\mathbf{E}_t||_{l,\mathrm{curl}}^2 dt].$$

Similarly, we can obtain

$$\begin{aligned} Err_2 &\leq \tau[2\delta_2||\overline{\xi}_h^k||_0^2 + \frac{1}{2\delta_2}||\nabla\times(\overline{\mathbf{H}}^k - \frac{1}{\tau}\int_{I^k}\mathbf{H}(s)ds||_0^2 \\ &\leq \tau[\delta_2(||\xi_h^k||_0^2 + ||\xi_h^{k-1}||_0^2) + \frac{\tau^3}{8\delta_2}\int_{I^k}||\nabla\times\mathbf{H}_{tt}(s)||_0^2 ds]. \end{aligned}$$

$$\begin{aligned} Err_3 &= -2\tau(\frac{1}{\tau}\int_{I^k}\mathbf{J}(s)ds - \overline{\mathbf{J}}^k + \overline{\mathbf{J}}^k - \Pi_h\overline{\mathbf{J}}^k + \overline{\tilde{\xi}}_h^k, \overline{\xi}_h^k) \\ &\leq -2\tau(\overline{\tilde{\xi}}_h^k, \overline{\xi}_h^k) + \tau[2\delta_3||\overline{\xi}_h^k||_0^2 + \frac{1}{\delta_3}(||\frac{1}{\tau}\int_{I^k}\mathbf{J}(s)ds - \overline{\mathbf{J}}^k||_0^2 + ||\overline{\mathbf{J}}^k - \Pi_h\overline{\mathbf{J}}^k||_0^2)] \\ &\leq -2\tau(\overline{\tilde{\xi}}_h^k, \overline{\xi}_h^k) + \tau[\delta_3(||\xi_h^k||_0^2 + ||\xi_h^{k-1}||_0^2) \\ &\quad + \frac{\tau^3}{4\delta_3}\int_{I^k}||\mathbf{J}_{tt}(s)||_0^2 ds + Ch^{2l}||\mathbf{J}||_{L^\infty(0,T;H^l(\mathrm{curl};\Omega))}^2]. \end{aligned}$$

By the same arguments, we have

$$\begin{aligned} &Err_4 = 2\tau(\nabla\times(\Pi_h\overline{\mathbf{E}}^k - \overline{\mathbf{E}}^k) + \nabla\times(\overline{\mathbf{E}}^k - \frac{1}{\tau}\int_{I^k}\mathbf{E}(s)ds), \overline{\eta}_h^k) \\ &\leq \tau[2\delta_4||\overline{\eta}_h^k||_0^2 + \frac{1}{\delta_4}(||\nabla\times(\Pi_h\overline{\mathbf{E}}^k - \overline{\mathbf{E}}^k)||_0^2 + ||\nabla\times(\overline{\mathbf{E}}^k - \frac{1}{\tau}\int_{I^k}\mathbf{E}(s)ds)||_0^2)] \\ &\leq \tau\delta_4(||\eta_h^k||_0^2 + ||\eta_h^{k-1}||_0^2) + \frac{\tau}{\delta_4}(Ch^{2l}||\mathbf{E}||_{L^\infty(0,T;H^l(\mathrm{curl};\Omega))}^2 + \frac{\tau^3}{4}\int_{I^k}||\nabla\times\mathbf{E}_{tt}(s)||_0^2 ds), \end{aligned}$$

and

$$\begin{aligned} Err_5 &= -2\tau(\frac{1}{\tau}\int_{I^k}\mathbf{K}(s)ds - \overline{\mathbf{K}}^k + \overline{\mathbf{K}}^k - P_h\overline{\mathbf{K}}^k + \overline{\tilde{\eta}}_h^k, \overline{\eta}_h^k) \\ &\leq -2\tau(\overline{\tilde{\eta}}_h^k, \overline{\eta}_h^k) + \tau\delta_5(||\eta_h^k||_0^2 + ||\eta_h^{k-1}||_0^2) + \frac{\tau^4}{8\delta_5}\int_{I^k}||\mathbf{K}_{tt}||_0^2 ds. \end{aligned}$$

Similar to Err_1, we have

$$\begin{aligned} Err_6 &= \frac{2\tau}{\epsilon_0\omega_{pe}^2}(\delta_\tau(\Pi_h\mathbf{J}^k - \mathbf{J}^k), \overline{\tilde{\xi}}_h^k) \\ &\leq \frac{\tau}{\epsilon_0\omega_{pe}^2}[\delta_6(||\tilde{\xi}_h^k||_0^2 + ||\tilde{\xi}_h^{k-1}||_0^2) + \frac{1}{2\delta_6\tau}\int_{I^k} Ch^{2l}||\mathbf{J}_t||_{l,\mathrm{curl}}^2 dt]. \end{aligned}$$

Furthermore, we have

$$Err_7 = \frac{2\tau\Gamma_e}{\epsilon_0\omega_{pe}^2}(\Pi_h\overline{\mathbf{J}}^k - \overline{\mathbf{J}}^k + \overline{\mathbf{J}}^k - \frac{1}{\tau}\int_{I^k}\mathbf{J}(s)ds, \overline{\tilde{\xi}}_h^k)$$

$$\leq \frac{\tau\Gamma_e}{\epsilon_0\omega_{pe}^2}[\delta_7(||\tilde{\xi}_h^k||_0^2 + ||\tilde{\xi}_h^{k-1}||_0^2) + \frac{1}{\delta_7}(\frac{\tau^3}{4}\int_{I^k}||\mathbf{J}_{tt}||_0^2 ds + Ch^{2l}||\mathbf{J}||^2_{L^\infty(0,T;H^l(\mathrm{curl};\Omega))})],$$

$$\begin{aligned}
Err_8 &= 2\tau(\frac{1}{\tau}\int_{I^k}\mathbf{E}(s)ds - \overline{\mathbf{E}}^k + \overline{\mathbf{E}}^k - \Pi_h\overline{\mathbf{E}}^k + \overline{\tilde{\xi}}_h^k, \overline{\tilde{\xi}}_h^k)\\
&\leq 2\tau(\overline{\tilde{\xi}}_h^k, \overline{\tilde{\xi}}_h^k) + \tau[2\delta_8||\overline{\tilde{\xi}}_h^k||_0^2 + \frac{1}{\delta_8}(||\frac{1}{\tau}\int_{I^k}\mathbf{E}(s)ds - \overline{\mathbf{E}}^k||_0^2 + ||\overline{\mathbf{E}}^k - \Pi_h\overline{\mathbf{E}}^k||_0^2)]\\
&\leq 2\tau(\overline{\tilde{\xi}}_h^k, \overline{\tilde{\xi}}_h^k) + \tau\delta_8(||\tilde{\xi}_h^k||_0^2 + ||\tilde{\xi}_h^{k-1}||_0^2)\\
&\quad + \frac{\tau}{\delta_8}(\frac{\tau^3}{4}\int_{I^k}||\mathbf{E}_{tt}||_0^2 ds + Ch^{2l}||\mathbf{E}||^2_{L^\infty(0,T;H^l(\mathrm{curl};\Omega))}),\\
Err_9 &\leq \frac{\tau\Gamma_m}{\mu_0\omega_{pm}^2}[\delta_9(||\tilde{\eta}_h^k||_0^2 + ||\tilde{\eta}_h^{k-1}||_0^2) + \frac{\tau^3}{8\delta_9}\int_{I^k}||\mathbf{K}_{tt}||_0^2 ds],
\end{aligned}$$

and

$$\begin{aligned}
Err_{10} &= 2\tau(\frac{1}{\tau}\int_{I^k}\mathbf{H}(s)ds - \overline{\mathbf{H}}^k + \overline{\mathbf{H}}^k - P_h\overline{\mathbf{H}}^k + \overline{\eta}_h^k, \overline{\tilde{\eta}}_h^k)\\
&\leq 2\tau(\overline{\eta}_h^k, \overline{\tilde{\eta}}_h^k) + \tau[\delta_{10}(||\tilde{\eta}_h^k||_0^2 + ||\tilde{\eta}_h^{k-1}||_0^2) + \frac{\tau^3}{8\delta_{10}}\int_{I^k}||\mathbf{H}_{tt}||_0^2 ds].
\end{aligned}$$

Substituting the estimates of Err_i into (3.94), and summing up the results from $k = 1$ to n ($n \leq M-1$), and using the facts $n\tau \leq T$ and $\xi_h^0 = \eta_h^0 = \tilde{\xi}_h^0 = \tilde{\eta}_h^0 = 0$, we can obtain (details see [191])

$$\begin{aligned}
&\epsilon_0||\xi_h^n||_0^2 + \mu_0||\eta_h^n||_0^2 + \frac{1}{\epsilon_0\omega_{pe}^2}||\tilde{\xi}_h^n||_0^2 + \frac{1}{\mu_0\omega_{pm}^2}||\tilde{\eta}_h^n||_0^2\\
&\leq C\tau\sum_{k=1}^{n-1}(||\xi_h^k||_0^2 + ||\eta_h^k||_0^2 + ||\tilde{\xi}_h^k||_0^2 + ||\tilde{\eta}_h^k||_0^2) + C(h^{2l} + \tau^4),
\end{aligned}$$

which, along with the discrete Gronwall inequality, the triangle inequality, the estimates (3.78) and (3.79), completes the proof. □

3.5 The Leap-Frog Scheme for the Drude Model

3.5.1 The Leap-Frog Scheme

The Crank-Nicolson scheme discussed in last section is implicit, hence we have to solve a linear system at each time step, which is quite computationally intensive. Using a similar idea to the famous Yee scheme [299], we can construct an explicit leap-frog finite element scheme [183]: Given initial approximations $\mathbf{E}_h^0, \mathbf{K}_h^0, \mathbf{H}_h^{\frac{1}{2}}, \mathbf{J}_h^{\frac{1}{2}}$, for $k = 1, 2, \cdots$, find $\mathbf{E}_h^k \in \mathbf{V}_h^0, \mathbf{J}_h^{k+\frac{1}{2}} \in \mathbf{V}_h, \mathbf{H}_h^{k+\frac{1}{2}}, \mathbf{K}_h^k \in \mathbf{U}_h$ such that

$$\epsilon_0(\frac{\mathbf{E}_h^k - \mathbf{E}_h^{k-1}}{\tau}, \boldsymbol{\phi}_h) - (\mathbf{H}_h^{k-\frac{1}{2}}, \nabla \times \boldsymbol{\phi}_h) + (\mathbf{J}_h^{k-\frac{1}{2}}, \boldsymbol{\phi}_h) = 0, \tag{3.96}$$

$$\mu_0(\frac{\mathbf{H}_h^{k+\frac{1}{2}} - \mathbf{H}_h^{k-\frac{1}{2}}}{\tau}, \boldsymbol{\psi}_h) + (\nabla \times \mathbf{E}_h^k, \boldsymbol{\psi}_h) + (\mathbf{K}_h^k, \boldsymbol{\psi}_h) = 0, \tag{3.97}$$

$$\frac{1}{\epsilon_0\omega_{pe}^2}(\frac{\mathbf{J}_h^{k+\frac{1}{2}} - \mathbf{J}_h^{k-\frac{1}{2}}}{\tau}, \tilde{\boldsymbol{\phi}}_h) + \frac{\Gamma_e}{\epsilon_0\omega_{pe}^2}(\frac{\mathbf{J}_h^{k+\frac{1}{2}} + \mathbf{J}_h^{k-\frac{1}{2}}}{2}, \tilde{\boldsymbol{\phi}}_h) = (\mathbf{E}_h^k, \tilde{\boldsymbol{\phi}}_h), \tag{3.98}$$

$$\frac{1}{\mu_0\omega_{pm}^2}(\frac{\mathbf{K}_h^k - \mathbf{K}_h^{k-1}}{\tau}, \tilde{\boldsymbol{\psi}}_h) + \frac{\Gamma_m}{\mu_0\omega_{pm}^2}(\frac{\mathbf{K}_h^k + \mathbf{K}_h^{k-1}}{2}, \tilde{\boldsymbol{\psi}}_h) = (\mathbf{H}_h^{k-\frac{1}{2}}, \tilde{\boldsymbol{\psi}}_h), \tag{3.99}$$

for any $\boldsymbol{\phi}_h \in \mathbf{V}_h^0$, $\boldsymbol{\psi}_h \in \mathbf{U}_h$, $\tilde{\boldsymbol{\phi}}_h \in \mathbf{V}_h$, $\tilde{\boldsymbol{\psi}}_h \in \mathbf{U}_h$.

Note that (3.98) and (3.99) can be simplified to

$$\mathbf{J}_h^{k+\frac{1}{2}} = \frac{2\epsilon_0\omega_{pe}^2}{2\tau^{-1} + \Gamma_e}\mathbf{E}_h^k + \frac{2\tau^{-1} - \Gamma_e}{2\tau^{-1} + \Gamma_e}\mathbf{J}_h^{k-\frac{1}{2}}, \tag{3.100}$$

$$\mathbf{K}_h^k = \frac{2\mu_0\omega_{pm}^2}{2\tau^{-1} + \Gamma_m}\mathbf{H}_h^{k-\frac{1}{2}} + \frac{2\tau^{-1} - \Gamma_m}{2\tau^{-1} + \Gamma_m}\mathbf{K}_h^{k-1}. \tag{3.101}$$

respectively.

In practice, the above leap-frog scheme can be implemented as follows: at each time step, we first solve (3.96) for $\mathbf{E}_h^k$ and update $\mathbf{K}_h^k$ using (3.101) in parallel, then solve (3.97) for $\mathbf{H}_h^{k+\frac{1}{2}}$ and update $\mathbf{J}_h^{k+\frac{1}{2}}$ by (3.100) in parallel. Compared to the Crank-Nicolson scheme presented in the last section, the leap-frog scheme is more efficient for solving large-scale problems, since no large global coefficient matrix has to be stored and inversed. Of course, we still have to inverse two mass matrices: one for (3.96) and one for (3.98). For the lowest-order cubic (or rectangular) edge element, we can even use mass-lumping technique [217, p. 352] for the mass matrix in (3.96) to speed up the computation, in which case, the mass matrix becomes a diagonal matrix. Of course, being an explicit scheme, the leap-frog scheme has a time step constraint as we will show in next section.

3.5.2 The Stability Analysis

In this section, we shall prove that the leap-frog scheme (3.98) and (3.99) is conditionally stable and has a discrete stability similar to the continuous stability obtained in Lemma 3.12.

Lemma 3.17. *Denote* $\gamma_e = |\frac{2\tau^{-1}-\Gamma_e}{2\tau^{-1}+\Gamma_e}|$, $\gamma_m = |\frac{2\tau^{-1}-\Gamma_m}{2\tau^{-1}+\Gamma_m}|$. *For the recursively defined* $\mathbf{J}_h^{k+\frac{1}{2}}$ *and* $\mathbf{K}_h^k$, *we have*

$$(i)\ \ ||\mathbf{J}_h^{k+\frac{1}{2}}||_0^2 \le 2[\epsilon_0^2\omega_{pe}^4\tau T\sum_{l=1}^{k}||\mathbf{E}_h^l||_0^2 + \gamma_e^{2k}||\mathbf{J}_h^{\frac{1}{2}}||_0^2], \tag{3.102}$$

$$(ii)\ \ ||\mathbf{K}_h^k||_0^2 \le 2[\mu_0^2\omega_{pm}^4\tau T\sum_{l=0}^{k-1}||\mathbf{H}_h^{l+\frac{1}{2}}||_0^2 + \gamma_m^{2k}||\mathbf{K}_h^0||_0^2]. \tag{3.103}$$

Proof. From (3.100) and the triangle inequality, we have

$$\begin{aligned}
||\mathbf{J}_h^{k+\frac{1}{2}}||_0 &\le \epsilon_0\omega_{pe}^2\tau||\mathbf{E}_h^k||_0 + \gamma_e||\mathbf{J}_h^{k-\frac{1}{2}}||_0\\
&\le \epsilon_0\omega_{pe}^2\tau||\mathbf{E}_h^k||_0 + \gamma_e(\epsilon_0\omega_{pe}^2\tau||\mathbf{E}_h^{k-1}||_0 + \gamma_e||\mathbf{J}_h^{k-\frac{3}{2}}||_0)\\
&\le \cdots\\
&\le \epsilon_0\omega_{pe}^2\tau(||\mathbf{E}_h^k||_0 + \gamma_e||\mathbf{E}_h^{k-1}||_0 + \cdots + \gamma_e^{k-1}||\mathbf{E}_h^1||_0) + \gamma_e^k||\mathbf{J}_h^{\frac{1}{2}}||_0\\
&\le \epsilon_0\omega_{pe}^2\tau\sum_{l=1}^{k}||\mathbf{E}_h^l||_0 + \gamma_e^k||\mathbf{J}_h^{\frac{1}{2}}||_0,
\end{aligned} \tag{3.104}$$

where we used the fact $\gamma_e < 1$ in the last step.

Squaring both sides of (3.104), we further have

$$\begin{aligned}
||\mathbf{J}_h^{k+\frac{1}{2}}||_0^2 &\le 2[\epsilon_0^2\omega_{pe}^4\tau^2(\sum_{l=1}^{k}||\mathbf{E}_h^l||_0)^2 + \gamma_e^{2k}||\mathbf{J}_h^{\frac{1}{2}}||_0^2]\\
&\le 2[\epsilon_0^2\omega_{pe}^4\tau^2(\sum_{l=1}^{k}1^2)(\sum_{l=1}^{k}||\mathbf{E}_h^l||_0^2) + \gamma_e^{2k}||\mathbf{J}_h^{\frac{1}{2}}||_0^2]\\
&\le 2[\epsilon_0^2\omega_{pe}^4\tau T\sum_{l=1}^{k}||\mathbf{E}_h^l||_0^2 + \gamma_e^{2k}||\mathbf{J}_h^{\frac{1}{2}}||_0^2],
\end{aligned} \tag{3.105}$$

where we used the fact $k\tau \le T$ in the last step.

Similarly, from (3.101), we have

$$||\mathbf{K}_h^k||_0 \le \mu_0\omega_{pm}^2\tau||\mathbf{H}_h^{k-\frac{1}{2}}||_0 + \gamma_m||\mathbf{K}_h^{k-1}||_0,$$

using which and repeating the above procedure, we can obtain

$$||\mathbf{K}_h^k||_0^2 \le 2[\mu_0^2\omega_{pm}^4\tau T\sum_{l=0}^{k-1}||\mathbf{H}_h^{l+\frac{1}{2}}||_0^2 + \gamma_m^{2k}||\mathbf{K}_h^0||_0^2], \tag{3.106}$$

which completes the proof. □

Theorem 3.11. *Let $C_v = 1/\sqrt{\mu_0\epsilon_0}$ denote the speed of light in vacuum, and C_{inv} denote the constant from the standard inverse estimate*

$$||\nabla\times\psi_h||_0 \le C_{inv}h^{-1}||\psi_h||_0, \quad \psi_h \in \mathbf{V}_h. \tag{3.107}$$

Then under the assumption that the time step

$$\tau = \min(\frac{h}{2C_{inv}C_v}, 1), \tag{3.108}$$

the solutions of (3.96)–(3.99) satisfy the following:

$$\begin{aligned}&\epsilon_0||\mathbf{E}_h^n||_0^2 + \mu_0||\mathbf{H}_h^{n+\frac{1}{2}}||_0^2 + \frac{1}{\mu_0\omega_{pm}^2}||\mathbf{K}_h^n||_0^2 + \frac{1}{\epsilon_0\omega_{pe}^2}||\mathbf{J}_h^{n+\frac{1}{2}}||_0^2\\ &\le C\left(||\mathbf{E}_h^0||_0^2 + ||\mathbf{H}_h^{\frac{1}{2}}||_0^2 + ||\mathbf{J}_h^{\frac{1}{2}}||_0^2 + ||\mathbf{K}_h^0||_0^2\right), \quad \forall\, n \ge 1,\end{aligned}$$

where the constant $C > 1$ is independent of h and τ.

Proof. Choosing $\boldsymbol{\phi}_h = \tau(\mathbf{E}_h^k + \mathbf{E}_h^{k-1})$ in (3.96), $\boldsymbol{\psi}_h = \tau(\mathbf{H}_h^{k+\frac{1}{2}} + \mathbf{H}_h^{k-\frac{1}{2}})$ in (3.97), adding (3.96) and (3.97) together, then using the following identity

$$\begin{aligned}&-(\mathbf{H}_h^{k-\frac{1}{2}}, \nabla\times(\mathbf{E}_h^k + \mathbf{E}_h^{k-1})) + (\nabla\times\mathbf{E}_h^k, \mathbf{H}_h^{k+\frac{1}{2}} + \mathbf{H}_h^{k-\frac{1}{2}})\\ &= -(\mathbf{H}_h^{k-\frac{1}{2}}, \nabla\times\mathbf{E}_h^{k-1}) + (\nabla\times\mathbf{E}_h^k, \mathbf{H}_h^{k+\frac{1}{2}}),\end{aligned} \tag{3.109}$$

and summing the resultants from $k = 1$ to $k = n$, we obtain

$$\begin{aligned}&\epsilon_0(||\mathbf{E}_h^n||_0^2 - ||\mathbf{E}_h^0||_0^2) + \mu_0(||\mathbf{H}_h^{n+\frac{1}{2}}||_0^2 - ||\mathbf{H}_h^{\frac{1}{2}}||_0^2)\\ =\ &\tau[(\mathbf{H}_h^{\frac{1}{2}}, \nabla\times\mathbf{E}_h^0) - (\nabla\times\mathbf{E}_h^n, \mathbf{H}_h^{n+\frac{1}{2}})]\\ &-\tau\sum_{k=1}^{n}(\mathbf{J}_h^{k-\frac{1}{2}}, \mathbf{E}_h^k + \mathbf{E}_h^{k-1}) - \tau\sum_{k=1}^{n}(\mathbf{K}_h^k, \mathbf{H}_h^{k+\frac{1}{2}} + \mathbf{H}_h^{k-\frac{1}{2}}).\end{aligned} \tag{3.110}$$

By the Cauchy-Schwartz inequality and the inverse estimate (3.107), we have

$$\begin{aligned}\tau(\nabla\times\mathbf{E}_h^n,\mathbf{H}_h^{n+\frac{1}{2}}) &\le \tau\cdot C_{inv}h^{-1}||\mathbf{E}_h^n||_0||\mathbf{H}_h^{n+\frac{1}{2}}||_0\\ &= \tau\cdot C_{inv}h^{-1}\cdot C_v\sqrt{\epsilon_0}||\mathbf{E}_h^n||_0\cdot\sqrt{\mu_0}||\mathbf{H}_h^{n+\frac{1}{2}}||_0\\ &\le \delta_1\epsilon_0||\mathbf{E}_h^n||_0^2+\frac{1}{4\delta_1}(\frac{C_{inv}C_v\tau}{h})^2\mu_0||\mathbf{H}_h^{n+\frac{1}{2}}||_0^2, \quad (3.111)\end{aligned}$$

and

$$\begin{aligned}\tau\sum_{k=1}^{n}(\mathbf{J}_h^{k-\frac{1}{2}},\mathbf{E}_h^k+\mathbf{E}_h^{k-1}) &\le \tau\sum_{k=1}^{n}||\mathbf{J}_h^{k-\frac{1}{2}}||_0(||\mathbf{E}_h^k||_0+||\mathbf{E}_h^{k-1}||_0)\\ &\le \tau\sum_{k=1}^{n}[\delta_2||\mathbf{E}_h^k||_0^2+\delta_3||\mathbf{E}_h^{k-1}||_0^2+(\frac{1}{4\delta_2}+\frac{1}{4\delta_3})||\mathbf{J}_h^{k-\frac{1}{2}}||_0^2].\end{aligned}$$

Furthermore, from (3.105) and the fact that $\gamma_e<1$, we have

$$\begin{aligned}\tau\sum_{k=1}^{n}||\mathbf{J}_h^{k-\frac{1}{2}}||_0^2 &\le 2\tau\sum_{k=1}^{n}[\epsilon_0^2\omega_{pe}^4\tau T\sum_{l=1}^{k-1}||\mathbf{E}_h^l||_0^2+\gamma_e^{2(k-1)}||\mathbf{J}_h^{\frac{1}{2}}||_0^2]\\ &\le 2[\epsilon_0^2\omega_{pe}^4\tau T^2\sum_{l=1}^{n-1}||\mathbf{E}_h^l||_0^2+T||\mathbf{J}_h^{\frac{1}{2}}||_0^2].\end{aligned}$$

Similarly, we have

$$\begin{aligned}&\tau\sum_{k=1}^{n}(\mathbf{K}_h^k,\mathbf{H}_h^{k+\frac{1}{2}}+\mathbf{H}_h^{k-\frac{1}{2}})\\ &\le \tau\sum_{k=1}^{n}[\delta_4||\mathbf{H}_h^{k+\frac{1}{2}}||_0^2+\delta_5||\mathbf{H}_h^{k-\frac{1}{2}}||_0^2+(\frac{1}{4\delta_4}+\frac{1}{4\delta_5})||\mathbf{K}_h^k||_0^2],\end{aligned}$$

and

$$\begin{aligned}\tau\sum_{k=1}^{n}||\mathbf{K}_h^k||_0^2 &\le 2\tau\sum_{k=1}^{n}[\mu_0^2\omega_{pm}^4\tau T\sum_{l=0}^{k-1}||\mathbf{H}_h^{l+\frac{1}{2}}||_0^2+\gamma_m^{2k}||\mathbf{K}_h^0||_0^2]\\ &\le 2[\mu_0^2\omega_{pm}^4\tau T^2\sum_{l=0}^{n-1}||\mathbf{H}_h^{l+\frac{1}{2}}||_0^2+T||\mathbf{K}_h^0||_0^2].\end{aligned}$$

Substituting the above estimates and the following (let $n = 0$ in (3.111))

$$\tau(\mathbf{H}_h^{\frac{1}{2}}, \nabla \times \mathbf{E}_h^0) \le \delta_1 \epsilon_0 ||\mathbf{E}_h^0||_0^2 + \frac{1}{4\delta_1}(\frac{C_{inv} C_v \tau}{h})^2 \mu_0 ||\mathbf{H}_h^{\frac{1}{2}}||_0^2$$

into (3.110), we obtain

$$\begin{aligned}
&\epsilon_0(||\mathbf{E}_h^n||_0^2 - ||\mathbf{E}_h^0||_0^2) + \mu_0(||\mathbf{H}_h^{n+\frac{1}{2}}||_0^2 - ||\mathbf{H}_h^{\frac{1}{2}}||_0^2) \\
\le\ & \frac{\tau}{2}||\nabla \times \mathbf{H}_h^{\frac{1}{2}}||_0^2 + (\frac{\tau}{2} + \tau\delta_3)||\mathbf{E}_h^0||_0^2 \\
&+(\delta_1 + \frac{\tau\delta_2}{\epsilon_0})\epsilon_0||\mathbf{E}_h^n||_0^2 + [\frac{1}{4\delta_1}(\frac{C_{inv} C_v \tau}{h})^2 + \frac{\tau\delta_4}{\mu_0}]\mu_0||\mathbf{H}_h^{n+\frac{1}{2}}||_0^2 \\
&+\tau[(\frac{1}{2\delta_2} + \frac{1}{2\delta_3})\epsilon_0^2\omega_{pe}^4 T^2 + \delta_2 + \delta_3]\sum_{l=1}^{n-1}||\mathbf{E}_h^l||_0^2 + (\frac{1}{2\delta_2} + \frac{1}{2\delta_3})T||\mathbf{J}_h^{\frac{1}{2}}||_0^2 \\
&+\tau[(\frac{1}{2\delta_4} + \frac{1}{2\delta_5})\mu_0^2\omega_{pm}^4 T^2 + \delta_4 + \delta_5]\sum_{l=0}^{n-1}||\mathbf{H}_h^{l+\frac{1}{2}}||_0^2 + (\frac{1}{2\delta_4} + \frac{1}{2\delta_5})T||\mathbf{K}_h^0||_0^2.
\end{aligned}$$

By choosing δ_i small enough and $\tau = O(h)$ such that $||\mathbf{E}_h^n||_0^2$ and $||\mathbf{H}_h^{n+\frac{1}{2}}||_0^2$ can be controlled by the left-hand side (e.g., $\delta_1 = \frac{1}{4}$, $\delta_2 = \frac{1}{4}\epsilon_0$, $\delta_4 = \frac{1}{4}\mu_0$, $\tau = \min(\frac{h}{2C_{inv}C_v}, 1)$, $\delta_3 = \epsilon_0, \delta_5 = \mu_0$), and using the discrete Gronwall inequality, we have

$$\epsilon_0||\mathbf{E}_h^n||_0^2 + \mu_0||\mathbf{H}_h^{n+\frac{1}{2}}||_0^2 \le C[||\mathbf{E}_h^0||_0^2 + ||\mathbf{H}_h^{\frac{1}{2}}||_0^2 + ||\mathbf{J}_h^{\frac{1}{2}}||_0^2 + ||\mathbf{K}_h^0||_0^2],$$

which, along with (3.105) and (3.106), concludes the proof. □

Remark 3.1. Note that when h is small enough, the time step constraint (3.108) reduces to the standard CFL condition $\tau = O(h)$, which is often imposed on explicit schemes used to solve the first-order hyperbolic systems.

A tight and accurate estimate of the constant C_{inv} in (3.107) is quite challenging, since it depends on the element shape and the order of the basis function. Below we just show a tight estimate of C_{inv} for the lowest rectangular edge element.

Lemma 3.18. *Consider a domain Ω is triangulated by a mesh T_h formed by m rectangles $K_i = [x_c^i - h_x, x_c^i + h_x] \times [y_c^i - h_y, y_c^i + h_y]$, $i = 1, \cdots, m$. Let $h = \max\{h_x, h_y\}$. Then we have*

$$C_{inv} \ge \max\{\frac{\sqrt{3}}{2}\frac{h}{h_x}, \frac{\sqrt{3}}{2}\frac{h}{h_y}\}. \tag{3.112}$$

Proof. Recall that the lowest edge element basis functions are (cf. Example 3.6):

$$N_1^i(x,y) = \begin{pmatrix} \frac{(y_c^i+h_y)-y}{4h_xh_y} \\ 0 \end{pmatrix}, \quad N_2^i(x,y) = \begin{pmatrix} 0 \\ \frac{x-(x_c^i-h_x)}{4h_xh_y} \end{pmatrix},$$
$$N_3^i(x,y) = \begin{pmatrix} \frac{(y_c^i-h_y)-y}{4h_xh_y} \\ 0 \end{pmatrix}, \quad N_4^i(x,y) = \begin{pmatrix} 0 \\ \frac{x-(x_c^i+h_x)}{4h_xh_y} \end{pmatrix},$$

where N_j^i, $j = 1,2,3,4$, start from the bottom edge and orient counterclockwisely.

It is easy to check that the 2-D curl of N_j^i satisfies

$$\int_{K_i} |\nabla \times N_j^i|^2 dxdy = \int_{K_i} |\frac{1}{4h_xh_y}|^2 dxdy = \frac{1}{4h_xh_y},$$

and N_j^i satisfies

$$\begin{aligned} \int_{K_i} |N_1^i|^2 dxdy &= \int_{K_i} (\frac{y_c^i + h_y - y}{4h_xh_y})^2 dxdy \\ &= \frac{2h_x}{(4h_xh_y)^2} \cdot \frac{-1}{3}(y_c^i + h_y - y)^3|_{y=y_c^i-h_y}^{y_c^i+h_y} = \frac{h_y}{3h_x}, \\ \int_{K_i} |N_3^i|^2 dxdy &= \frac{h_y}{3h_x}, \quad \int_{K_i} |N_2^i|^2 dxdy = |N_4^i|^2 dxdy = \frac{h_x}{3h_y}, \end{aligned}$$

from which we can see that

$$\frac{||\nabla \times N_j^i||_{0,K_i}^2}{||N_j^i||_{0,K_i}^2} = \frac{3}{4h_y^2}, \quad j = 1,3. \tag{3.113}$$

Similarly, we have

$$\frac{||\nabla \times N_j^i||_{0,K_i}^2}{||N_j^i||_{0,K_i}^2} = \frac{3}{4h_x^2}, \quad j = 2,4,$$

applying which to (3.107) we complete the proof. □

By Lemma 3.18, we should try to use shape regular meshes and avoid anisotropic meshes in practice computation, since the anisotropic mesh may have a very large C_{inv} and lead to a very small time step according to (3.108).

3.5.3 *The Optimal Error Estimate*

To carry out the error analysis for the scheme (3.96)–(3.99), we need some preliminary estimates.

Lemma 3.19. *Denote $u^j = u(\cdot, j\tau)$. For any $u \in H^2(0,T;L^2(\Omega))$, we have*

$$(i)\ \ ||u^k - \frac{1}{\tau}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}} u(s)ds||_0^2 \le \frac{\tau^3}{4}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}} ||u_{tt}(s)||_0^2 ds,$$

$$(ii)\ \ ||u^{k-\frac{1}{2}} - \frac{1}{\tau}\int_{t_{k-1}}^{t_k} u(s)ds||_0^2 \le \frac{\tau^3}{4}\int_{t_{k-1}}^{t_k} ||u_{tt}(s)||_0^2 ds,$$

$$(iii)\ \ ||\frac{1}{2}(u^{k-1}+u^k) - \frac{1}{\tau}\int_{t_{k-1}}^{t_k} u(s)ds||_0^2 \le \frac{\tau^3}{4}\int_{t_{k-1}}^{t_k} ||u_{tt}(s)||_0^2 ds,$$

$$(iv)\ \ ||\frac{1}{2}(u^{k-\frac{1}{2}}+u^{k+\frac{1}{2}}) - \frac{1}{\tau}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}} u(s)ds||_0^2 \le \frac{\tau^3}{4}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}} ||u_{tt}(s)||_0^2 ds.$$

Furthermore, for any $u \in H^1(0,T;L^2(\Omega))$, we have

$$(v)\ \ ||\delta_\tau u^{k+\frac{1}{2}}||_0^2 = ||\frac{u^{k+\frac{1}{2}}-u^{k-\frac{1}{2}}}{\tau}||_0^2 \le \frac{1}{\tau}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}} ||u_t(t)||_0^2 dt.$$

Proof. (i) Using the following integral identity

$$u(s) = u(t_k) + (s-t_k)u_t(t_k) + \int_s^{t_k}(r-s)u_{tt}(r)dr$$

we obtain

$$|u^k - \frac{1}{\tau}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}} u(s)ds|^2 = |-\frac{1}{\tau}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}}[\int_s^{t_k}(r-s)u_{tt}(r)dr]ds|^2$$

$$\le \frac{1}{\tau^2}(\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}}[\int_s^{t_k}(r-s)u_{tt}(r)dr]^2 ds)(\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}} 1^2 ds)$$

$$\le \frac{1}{\tau}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}}(\int_s^{t_k}(r-s)^2 dr)(\int_s^{t_k}|u_{tt}(r)|^2 dr)ds \le \frac{1}{4}\tau^3\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}}|u_{tt}(r)|^2 dr,$$

integrating which over Ω concludes the proof.

(ii) The proof is all the same as (i) except we use the following identity

$$u(s) = u(t_{k-\frac{1}{2}}) + (s-t_{k-\frac{1}{2}})u_t(t_{k-\frac{1}{2}}) + \int_{t_{k-\frac{1}{2}}}^{s}(s-r)u_{tt}(r)dr.$$

(iii) The proof is given in Lemma 3.16.

(iv) The proof is based on the following identity

$$\frac{1}{2}(u^{k-\frac{1}{2}}+u^{k+\frac{1}{2}})-\frac{1}{\tau}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}}u(s)ds=\frac{1}{2\tau}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}}(s-t_{k-\frac{1}{2}})(t_{k+\frac{1}{2}}-s)u_{ss}(s)ds.$$

(v) The proof is easily obtained by using $\frac{u^{k+\frac{1}{2}}-u^{k-\frac{1}{2}}}{\tau}=\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}}u_t(t)dt$.

□

Suppose that the solution is smooth enough, then we can prove the following optimal error estimate for the leap-frog scheme (3.96)–(3.99).

Theorem 3.12. *Let* $(\mathbf{E}^n,\mathbf{H}^n)$ *and* $(\mathbf{E}_h^n,\mathbf{H}_h^n)$ *be the analytic and finite element solutions at time* $t=t_n$*, respectively. Under the regularity assumptions*

$$\mathbf{H},\mathbf{K}\in L^\infty(0,T;(H^l(\Omega))^3),$$
$$\mathbf{E},\mathbf{J},\mathbf{E}_t,\mathbf{J}_t\in L^\infty(0,T;H^l(curl;\Omega)),$$
$$\mathbf{E}_{tt},\mathbf{H}_{tt},\mathbf{J}_{tt},\mathbf{K}_{tt},\nabla\times\mathbf{E}_{tt},\nabla\times\mathbf{H}_{tt}\in L^2(0,T;(L^2(\Omega))^3),$$

there exists a constant $C=C(T,\epsilon_0,\mu_0,\omega_{pe},\omega_{pm},\Gamma_e,\Gamma_m,\mathbf{E},\mathbf{H},\mathbf{J},\mathbf{K})$*, independent of both time step* τ *and mesh size* h*, such that*

$$\max_{1\le n}(||\mathbf{E}^n-\mathbf{E}_h^n||_0+||\mathbf{H}^{n+\frac{1}{2}}-\mathbf{H}_h^{n+\frac{1}{2}}||_0+||\mathbf{J}^{n+\frac{1}{2}}-\mathbf{J}_h^{n+\frac{1}{2}}||_0+||\mathbf{K}^n-\mathbf{K}_h^n||_0)$$
$$\le C(\tau^2+h^l)+C\Big(||\mathbf{E}^0-\mathbf{E}_h^0||_0+||\mathbf{H}^{\frac{1}{2}}-\mathbf{H}_h^{\frac{1}{2}}||_0+||\mathbf{J}^{\frac{1}{2}}-\mathbf{J}_h^{\frac{1}{2}}||_0+||\mathbf{K}^0-\mathbf{K}_h^0||_0\Big).$$

where $l\ge 1$ *is the order of basis functions defined in spaces* $\mathbf{U}_h$ *and* $\mathbf{V}_h$*.*

Proof. Integrating the governing equations (3.55) and (3.58) from t_{k-1} to t_k, and (3.56) and (3.57) from $t_{k-\frac{1}{2}}$ to $t_{k+\frac{1}{2}}$, then multiplying the respective resultants by $\frac{\boldsymbol{\phi}_h}{\tau}$, $\frac{\boldsymbol{\psi}_h}{\tau}$, $\frac{\tilde{\boldsymbol{\phi}}_h}{\tau}$, $\frac{\tilde{\boldsymbol{\psi}}_h}{\tau}$ and integrating over Ω, we have

$$\epsilon_0(\frac{\mathbf{E}^k-\mathbf{E}^{k-1}}{\tau},\boldsymbol{\phi}_h)-(\frac{1}{\tau}\int_{t_{k-1}}^{t_k}\mathbf{H}(s)ds,\nabla\times\boldsymbol{\phi}_h)+(\frac{1}{\tau}\int_{t_{k-1}}^{t_k}\mathbf{J}(s)ds,\boldsymbol{\phi}_h)=0, \tag{3.114}$$

$$\mu_0(\frac{\mathbf{H}^{k+\frac{1}{2}}-\mathbf{H}^{k-\frac{1}{2}}}{\tau},\boldsymbol{\psi}_h)+(\frac{1}{\tau}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}}\nabla\times\mathbf{E}(s)ds,\boldsymbol{\psi}_h)+(\frac{1}{\tau}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}}\mathbf{K}(s)ds,\boldsymbol{\psi}_h)=0, \tag{3.115}$$

$$\frac{1}{\epsilon_0\omega_{pe}^2}(\frac{\mathbf{J}^{k+\frac{1}{2}}-\mathbf{J}^{k-\frac{1}{2}}}{\tau},\tilde{\boldsymbol{\phi}}_h)+\frac{\Gamma_e}{\epsilon_0\omega_{pe}^2}(\frac{1}{\tau}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}}\mathbf{J}(s)ds,\tilde{\boldsymbol{\phi}}_h)=(\frac{1}{\tau}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}}\mathbf{E}(s)ds,\tilde{\boldsymbol{\phi}}_h), \tag{3.116}$$

$$\frac{1}{\mu_0\omega_{pm}^2}(\frac{\mathbf{K}^k-\mathbf{K}^{k-1}}{\tau},\tilde{\boldsymbol{\psi}}_h)+\frac{\Gamma_m}{\mu_0\omega_{pm}^2}(\frac{1}{\tau}\int_{t_{k-1}}^{t_k}\mathbf{K}(s)ds,\tilde{\boldsymbol{\psi}}_h)=(\frac{1}{\tau}\int_{t_{k-1}}^{t_k}\mathbf{H}(s)ds,\tilde{\boldsymbol{\psi}}_h). \tag{3.117}$$

Denote

$$\xi_h^k = \Pi_h \mathbf{E}^k - \mathbf{E}_h^k, \quad \tilde{\xi}_h^{k-\frac{1}{2}} = \Pi_h \mathbf{J}^{k-\frac{1}{2}} - \mathbf{J}_h^{k-\frac{1}{2}},$$
$$\eta_h^{k-\frac{1}{2}} = P_h \mathbf{H}^{k-\frac{1}{2}} - \mathbf{H}_h^{k-\frac{1}{2}}, \quad \tilde{\eta}_h^k = P_h \mathbf{K}^k - \mathbf{K}_h^k.$$

Subtracting (3.96)–(3.99) from (3.114)–(3.117), respectively, we obtain

$$\begin{aligned}
&\epsilon_0(\frac{\xi_h^k - \xi_h^{k-1}}{\tau}, \boldsymbol{\phi}_h) - (\eta_h^{k-\frac{1}{2}}, \nabla \times \boldsymbol{\phi}_h) \\
&= \epsilon_0(\delta_\tau(\Pi_h \mathbf{E}^k - \mathbf{E}^k), \boldsymbol{\phi}_h) - (P_h \mathbf{H}^{k-\frac{1}{2}} - \frac{1}{\tau}\int_{t_{k-1}}^{t_k} \mathbf{H}(s)ds, \nabla \times \boldsymbol{\phi}_h) \\
&\quad + (-\tilde{\xi}_h^{k-\frac{1}{2}} + \Pi_h \mathbf{J}^{k-\frac{1}{2}} - \frac{1}{\tau}\int_{t_{k-1}}^{t_k} \mathbf{J}(s)ds, \boldsymbol{\phi}_h), \qquad (3.118)
\end{aligned}$$

$$\begin{aligned}
&\mu_0(\frac{\eta_h^{k+\frac{1}{2}} - \eta_h^{k-\frac{1}{2}}}{\tau}, \boldsymbol{\psi}_h) + (\nabla \times \xi_h^k, \boldsymbol{\psi}_h) \\
&= \mu_0(\delta_\tau(P_h \mathbf{H}^{k+\frac{1}{2}} - \mathbf{H}^{k+\frac{1}{2}}), \boldsymbol{\psi}_h) + (\nabla \times (\Pi_h \mathbf{E}^k - \frac{1}{\tau}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}} \mathbf{E}(s)ds), \boldsymbol{\psi}_h) \\
&\quad + (-\tilde{\eta}_h^k + P_h \mathbf{K}^k - \frac{1}{\tau}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}} \mathbf{K}(s)ds, \boldsymbol{\psi}_h), \qquad (3.119)
\end{aligned}$$

$$\begin{aligned}
&\frac{1}{\epsilon_0 \omega_{pe}^2}(\frac{\tilde{\xi}_h^{k+\frac{1}{2}} - \tilde{\xi}_h^{k-\frac{1}{2}}}{\tau}, \tilde{\boldsymbol{\phi}}_h) + \frac{\Gamma_e}{\epsilon_0 \omega_{pe}^2}(\frac{1}{2}(\tilde{\xi}_h^{k+\frac{1}{2}} + \tilde{\xi}_h^{k-\frac{1}{2}}), \tilde{\boldsymbol{\phi}}_h) \\
&= \frac{1}{\epsilon_0 \omega_{pe}^2}(\delta_\tau(\Pi_h \mathbf{J}^{k+\frac{1}{2}} - \mathbf{J}^{k+\frac{1}{2}}), \tilde{\boldsymbol{\phi}}_h) + \frac{\Gamma_e}{\epsilon_0 \omega_{pe}^2}(\frac{1}{2}(\Pi_h \mathbf{J}^{k+\frac{1}{2}} + \Pi_h \mathbf{J}^{k-\frac{1}{2}}) \\
&\quad - \frac{1}{\tau}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}} \mathbf{J}(s)ds, \tilde{\boldsymbol{\phi}}_h) + (\frac{1}{\tau}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}} \mathbf{E}(s)ds - \Pi_h \mathbf{E}^k + \xi_h^k, \tilde{\boldsymbol{\phi}}_h), \qquad (3.120)
\end{aligned}$$

and

$$\begin{aligned}
&\frac{1}{\mu_0 \omega_{pm}^2}(\frac{\tilde{\eta}_h^k - \tilde{\eta}_h^{k-1}}{\tau}, \tilde{\boldsymbol{\psi}}_h) + \frac{\Gamma_m}{\mu_0 \omega_{pm}^2}(\frac{1}{2}(\tilde{\eta}_h^k + \tilde{\eta}_h^{k-1}), \tilde{\boldsymbol{\psi}}_h) \\
&= \frac{1}{\mu_0 \omega_{pm}^2}(\delta_\tau(P_h \mathbf{K}^k - \mathbf{K}^k), \tilde{\boldsymbol{\psi}}_h) + \frac{\Gamma_m}{\mu_0 \omega_{pm}^2}(\frac{1}{2}(P_h \mathbf{K}^k + P_h \mathbf{K}^{k-1}) \\
&\quad - \frac{1}{\tau}\int_{t_{k-1}}^{t_k} \mathbf{K}(s)ds, \tilde{\boldsymbol{\psi}}_h) + (\frac{1}{\tau}\int_{t_{k-1}}^{t_k} \mathbf{H}(s)ds - P_h \mathbf{H}^{k-\frac{1}{2}} + \eta_h^{k-\frac{1}{2}}, \tilde{\boldsymbol{\psi}}_h). \qquad (3.121)
\end{aligned}$$

Choosing

$$\boldsymbol{\phi}_h = \xi_h^k + \xi_h^{k-1},\ \ \boldsymbol{\psi}_h = \eta_h^{k+\frac{1}{2}} + \eta_h^{k-\frac{1}{2}},\ \ \tilde{\boldsymbol{\phi}}_h = \tilde{\xi}_h^{k+\frac{1}{2}} + \tilde{\xi}_h^{k-\frac{1}{2}},\ \ \tilde{\boldsymbol{\psi}}_h = \tilde{\eta}_h^k + \tilde{\eta}_h^{k-1},$$

in (3.118)–(3.121), respectively, then using the following identities:

$$\begin{aligned}&(\nabla\times\xi_h^k, \eta_h^{k+\frac{1}{2}} + \eta_h^{k-\frac{1}{2}}) - (\eta_h^{k-\frac{1}{2}}, \nabla\times(\xi_h^k + \xi_h^{k-1}))\\ &= (\nabla\times\xi_h^k, \eta_h^{k+\frac{1}{2}}) - (\nabla\times\xi_h^{k-1}, \eta_h^{k-\frac{1}{2}}),\end{aligned}$$

and

$$\begin{aligned}&-(\tilde{\xi}_h^{k-\frac{1}{2}}, \xi_h^k + \xi_h^{k-1}) - (\tilde{\eta}_h^k, \eta_h^{k+\frac{1}{2}} + \eta_h^{k-\frac{1}{2}})\\ &+(\xi_h^k, \tilde{\xi}_h^{k+\frac{1}{2}} + \tilde{\xi}_h^{k-\frac{1}{2}}) + (\eta_h^{k-\frac{1}{2}}, \tilde{\eta}_h^k + \tilde{\eta}_h^{k-1})\\ &= -(\tilde{\xi}_h^{k-\frac{1}{2}}, \xi_h^{k-1}) + (\tilde{\xi}_h^{k+\frac{1}{2}}, \xi_h^k) - (\tilde{\eta}_h^k, \eta_h^{k+\frac{1}{2}}) + (\tilde{\eta}_h^{k-1}, \eta_h^{k-\frac{1}{2}}),\end{aligned}$$

and summing up the resultants from $k = 1$ to $k = n$, we obtain

$$\begin{aligned}&\epsilon_0(||\xi_h^n||_0^2 - ||\xi_h^0||_0^2) + \mu_0(||\eta_h^{n+\frac{1}{2}}||_0^2 - ||\eta_h^{\frac{1}{2}}||_0^2)\\ &\quad+\frac{1}{\epsilon_0\omega_{pe}^2}(||\tilde{\xi}_h^{n+\frac{1}{2}}||_0^2 - ||\tilde{\xi}_h^{\frac{1}{2}}||_0^2) + \frac{1}{\mu_0\omega_{pm}^2}(||\tilde{\eta}_h^n||_0^2 - ||\tilde{\eta}_h^0||_0^2)\\ \le\ &\tau\epsilon_0\sum_{k=1}^n(\delta_\tau(\Pi_h\mathbf{E}^k - \mathbf{E}^k), \xi_h^k + \xi_h^{k-1})\\ &-\tau\sum_{k=1}^n(P_h\mathbf{H}^{k-\frac{1}{2}} - \frac{1}{\tau}\int_{t_{k-1}}^{t_k}\mathbf{H}(s)ds, \nabla\times(\xi_h^k + \xi_h^{k-1}))\\ &+\tau\sum_{k=1}^n(\Pi_h\mathbf{J}^{k-\frac{1}{2}} - \frac{1}{\tau}\int_{t_{k-1}}^{t_k}\mathbf{J}(s)ds, \xi_h^k + \xi_h^{k-1})\\ &+\tau\mu_0\sum_{k=1}^n(\delta_\tau(P_h\mathbf{H}^{k+\frac{1}{2}} - \mathbf{H}^{k+\frac{1}{2}}), \eta_h^{k+\frac{1}{2}} + \eta_h^{k-\frac{1}{2}})\\ &+\tau\sum_{k=1}^n(\nabla\times(\Pi_h\mathbf{E}^k - \frac{1}{\tau}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}}\mathbf{E}(s)ds), \eta_h^{k+\frac{1}{2}} + \eta_h^{k-\frac{1}{2}})\\ &+\tau\sum_{k=1}^n(P_h\mathbf{K}^k - \frac{1}{\tau}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}}\mathbf{K}(s)ds, \eta_h^{k+\frac{1}{2}} + \eta_h^{k-\frac{1}{2}})\end{aligned}$$

$$
\begin{aligned}
&+\frac{\tau}{\epsilon_0\omega_{pe}^2}\sum_{k=1}^{n}(\delta_\tau(\Pi_h\mathbf{J}^{k+\frac{1}{2}}-\mathbf{J}^{k+\frac{1}{2}}),\tilde{\xi}_h^{k+\frac{1}{2}}+\tilde{\xi}_h^{k-\frac{1}{2}})\\
&+\frac{\tau\Gamma_e}{\epsilon_0\omega_{pe}^2}\sum_{k=1}^{n}(\frac{1}{2}(\Pi_h\mathbf{J}^{k+\frac{1}{2}}+\Pi_h\mathbf{J}^{k-\frac{1}{2}})-\frac{1}{\tau}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}}\mathbf{J}(s)ds,\tilde{\xi}_h^{k+\frac{1}{2}}+\tilde{\xi}_h^{k-\frac{1}{2}})\\
&+\tau\sum_{k=1}^{n}(\frac{1}{\tau}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}}\mathbf{E}(s)ds-\Pi_h\mathbf{E}^k,\tilde{\xi}_h^{k+\frac{1}{2}}+\tilde{\xi}_h^{k-\frac{1}{2}})\\
&+\frac{\tau}{\mu_0\omega_{pm}^2}\sum_{k=1}^{n}(\delta_\tau(P_h\mathbf{K}^k-\mathbf{K}^k),\tilde{\eta}_h^k+\tilde{\eta}_h^{k-1})\\
&+\frac{\tau\Gamma_m}{\mu_0\omega_{pm}^2}\sum_{k=1}^{n}(\frac{1}{2}(P_h\mathbf{K}^k+P_h\mathbf{K}^{k-1})-\frac{1}{\tau}\int_{t_{k-1}}^{t_k}\mathbf{K}(s)ds,\tilde{\eta}_h^k+\tilde{\eta}_h^{k-1})\\
&+\tau\sum_{k=1}^{n}(\frac{1}{\tau}\int_{t_{k-1}}^{t_k}\mathbf{H}(s)ds-P_h\mathbf{H}^{k-\frac{1}{2}},\tilde{\eta}_h^k+\tilde{\eta}_h^{k-1})+\tau(\tilde{\xi}_h^{n+\frac{1}{2}},\xi_h^n)-\tau(\tilde{\xi}_h^{\frac{1}{2}},\xi_h^0)\\
&-\tau(\tilde{\eta}_h^n,\eta_h^{n+\frac{1}{2}})+\tau(\tilde{\eta}_h^0,\eta_h^{\frac{1}{2}})+\tau(\nabla\times\xi_h^0,\eta_h^{\frac{1}{2}})-\tau(\nabla\times\xi_h^n,\eta_h^{n+\frac{1}{2}})\\
&=\sum_{i=1}^{18}Err_i. \qquad (3.122)
\end{aligned}
$$

The proof is completed by using Lemma 3.19 and careful estimating all Err_i. Readers can consult the proof of Theorem 3.10 or the original paper [183]. □

Remark 3.2. We like to remark that a similar leap-frog scheme can be developed for Maxwell's equations in free space by dropping the constitutive equations (3.57) and (3.58), and treating **J** and **K** as fixed sources in (3.55) and (3.56). Following the same proof as carried out above, we can show that the stability and error estimate in the free space become as:

$$\epsilon_0||\mathbf{E}_h^n||_0^2+\mu_0||\mathbf{H}_h^{n+\frac{1}{2}}||_0^2\le C\,[||\mathbf{E}_h^0||_0^2+||\mathbf{H}_h^{\frac{1}{2}}||_0^2],$$

and

$$
\begin{aligned}
&\max_{1\le n}(||\mathbf{E}^n-\mathbf{E}_h^n||_0+||\mathbf{H}^{n+\frac{1}{2}}-\mathbf{H}_h^{n+\frac{1}{2}}||_0)\\
&\le C(\tau^2+h^l)+C(||\mathbf{E}^0-\mathbf{E}_h^0||_0+||\mathbf{H}^{\frac{1}{2}}-\mathbf{H}_h^{\frac{1}{2}}||_0).
\end{aligned}
$$

3.6 Extensions to the Lorentz Model

3.6.1 The Well-Posedness of the Lorentz Model

Another popular model for describing wave propagation in metamaterials is the Lorentz model discussed in Chap. 1. Recall that the time-domain Lorentz model is described by the following governing equations:

$$\epsilon_0 \frac{\partial \mathbf{E}}{\partial t} + \frac{\partial \mathbf{P}}{\partial t} - \nabla \times \mathbf{H} = 0, \tag{3.123}$$

$$\mu_0 \frac{\partial \mathbf{H}}{\partial t} + \frac{\partial \mathbf{M}}{\partial t} + \nabla \times \mathbf{E} = 0, \tag{3.124}$$

$$\frac{1}{\epsilon_0 \omega_{pe}^2} \frac{\partial^2 \mathbf{P}}{\partial t^2} + \frac{\Gamma_e}{\epsilon_0 \omega_{pe}^2} \frac{\partial \mathbf{P}}{\partial t} + \frac{\omega_{e0}^2}{\epsilon_0 \omega_{pe}^2} \mathbf{P} - \mathbf{E} = 0, \tag{3.125}$$

$$\frac{1}{\mu_0 \omega_{pm}^2} \frac{\partial^2 \mathbf{M}}{\partial t^2} + \frac{\Gamma_m}{\mu_0 \omega_{pm}^2} \frac{\partial \mathbf{M}}{\partial t} + \frac{\omega_{m0}^2}{\mu_0 \omega_{pm}^2} \mathbf{M} - \mathbf{H} = 0. \tag{3.126}$$

Note that the governing equations written in this way are convenient to obtain the stability and error analysis elegantly. Hopefully, readers may appreciate this as we move forward.

For simplicity, we assume that the model problem (3.123)–(3.126) is complemented by a perfect conducting boundary condition (3.59), and initial conditions

$$\mathbf{E}(\mathbf{x}, 0) = \mathbf{E}_0(\mathbf{x}), \quad \mathbf{H}(\mathbf{x}, 0) = \mathbf{H}_0(\mathbf{x}), \quad \mathbf{P}(\mathbf{x}, 0) = \mathbf{P}_0(\mathbf{x}), \tag{3.127}$$

$$\mathbf{M}(\mathbf{x}, 0) = \mathbf{M}_0(\mathbf{x}), \ \frac{\partial \mathbf{P}}{\partial t}(\mathbf{x}, 0) = \mathbf{P}_1(\mathbf{x}), \ \frac{\partial \mathbf{M}}{\partial t}(\mathbf{x}, 0) = \mathbf{M}_1(\mathbf{x}), \tag{3.128}$$

where $\mathbf{E}_0, \mathbf{H}_0, \mathbf{P}_0, \mathbf{M}_0, \mathbf{P}_1$, and $\mathbf{M}_1$ are some given functions.

Denote $\mathbf{V}^*$ as the dual space of $\mathbf{V} = H_0(\text{curl}; \Omega)$. Then a weak formulation of (3.123)–(3.126) can be formed as: Find the solution $\mathbf{E} \in H^1(0, T; \mathbf{V}^*) \cap (L^2(0, T; \mathbf{V}))^3$, $\mathbf{H} \in H^1(0, T; (L^2(\Omega))^3)$, $\mathbf{P} \in H^2(0, T; \mathbf{V}^*)$, $\mathbf{M} \in H^2(0, T; (L^2(\Omega))^3)$ such that

$$\epsilon_0 (\frac{\partial \mathbf{E}}{\partial t}, \boldsymbol{\phi}) + (\frac{\partial \mathbf{P}}{\partial t}, \boldsymbol{\phi}) - (\mathbf{H}, \nabla \times \boldsymbol{\phi}) = 0, \tag{3.129}$$

$$\mu_0 (\frac{\partial \mathbf{H}}{\partial t}, \boldsymbol{\psi}) + (\frac{\partial \mathbf{M}}{\partial t}, \boldsymbol{\psi}) + (\nabla \times \mathbf{E}, \boldsymbol{\psi}) = 0, \tag{3.130}$$

$$\frac{1}{\epsilon_0 \omega_{pe}^2} [(\frac{\partial^2 \mathbf{P}}{\partial t^2}, \tilde{\boldsymbol{\phi}}) + \Gamma_e (\frac{\partial \mathbf{P}}{\partial t}, \tilde{\boldsymbol{\phi}}) + \omega_{e0}^2 (\mathbf{P}, \tilde{\boldsymbol{\phi}})] - (\mathbf{E}, \tilde{\boldsymbol{\phi}}) = 0, \tag{3.131}$$

$$\frac{1}{\mu_0 \omega_{pm}^2} [(\frac{\partial^2 \mathbf{M}}{\partial t^2}, \tilde{\boldsymbol{\psi}}) + \Gamma_m (\frac{\partial \mathbf{M}}{\partial t}, \tilde{\boldsymbol{\psi}}) + \omega_{m0}^2 (\mathbf{M}, \tilde{\boldsymbol{\psi}})] - (\mathbf{H}, \tilde{\boldsymbol{\psi}}) = 0, \tag{3.132}$$

hold true for any $\boldsymbol{\phi} \in H_0(\text{curl}; \Omega)$, $\boldsymbol{\psi} \in (L^2(\Omega))^3$, $\tilde{\boldsymbol{\phi}} \in H(\text{curl}; \Omega)$, and $\tilde{\boldsymbol{\psi}} \in (L^2(\Omega))^3$.

First, we have the following stability for the Lorentz model (3.129)–(3.132).

Lemma 3.20.

$$\begin{aligned}
&\epsilon_0||\mathbf{E}(t)||_0^2 + \mu_0||\mathbf{H}(t)||_0^2 + \frac{1}{\epsilon_0\omega_{pe}^2}||\frac{\partial \mathbf{P}}{\partial t}(t)||_0^2 \\
&\quad + \frac{1}{\mu_0\omega_{pm}^2}||\frac{\partial \mathbf{M}}{\partial t}(t)||_0^2 + \frac{\omega_{e0}^2}{\epsilon_0\omega_{pe}^2}||\mathbf{P}(t)||_0^2 + \frac{\omega_{m0}^2}{\mu_0\omega_{pm}^2}||\mathbf{M}(t)||_0^2 \\
&\leq \epsilon_0||\mathbf{E}(0)||_0^2 + \mu_0||\mathbf{H}(0)||_0^2 + \frac{1}{\epsilon_0\omega_{pe}^2}||\frac{\partial \mathbf{P}}{\partial t}(0)||_0^2 \\
&\quad + \frac{1}{\mu_0\omega_{pm}^2}||\frac{\partial \mathbf{M}}{\partial t}(0)||_0^2 + \frac{\omega_{e0}^2}{\epsilon_0\omega_{pe}^2}||\mathbf{P}(0)||_0^2 + \frac{\omega_{m0}^2}{\mu_0\omega_{pm}^2}||\mathbf{M}(0)||_0^2.
\end{aligned}$$

Proof. Choosing $\boldsymbol{\phi} = \mathbf{E}, \boldsymbol{\psi} = \mathbf{H}, \tilde{\boldsymbol{\phi}} = \frac{\partial \mathbf{P}}{\partial t}, \tilde{\boldsymbol{\psi}} = \frac{\partial \mathbf{M}}{\partial t}$ in (3.129)–(3.132) respectively, then summing up the resultants, we have

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}[\epsilon_0||\mathbf{E}||_0^2 + \frac{1}{\epsilon_0\omega_{pe}^2}||\frac{\partial \mathbf{P}}{\partial t}||_0^2 + \frac{\omega_{e0}^2}{\epsilon_0\omega_{pe}^2}||\mathbf{P}||_0^2] + \frac{\Gamma_e}{\epsilon_0\omega_{pe}^2}||\frac{\partial \mathbf{P}}{\partial t}||_0^2 \\
&+ \frac{1}{2}\frac{d}{dt}[\mu_0||\mathbf{H}||_0^2 + \frac{1}{\mu_0\omega_{pm}^2}||\frac{\partial \mathbf{M}}{\partial t}||_0^2 + \frac{\omega_{m0}^2}{\mu_0\omega_{pm}^2}||\mathbf{M}||_0^2] + \frac{\Gamma_m}{\mu_0\omega_{pm}^2}||\frac{\partial \mathbf{M}}{\partial t}||_0^2 = 0,
\end{aligned}$$

integrating which from 0 to t completes the proof. □

The existence and uniqueness of the solution for the Lorentz model (3.123)–(3.128) can be proved by following the same technique developed earlier in Theorem 3.8 for the Drude model.

Theorem 3.13. *The model (3.123)–(3.128) exists a unique solution* $\mathbf{E} \in H_0(\text{curl}; \Omega)$ *and* $\mathbf{H} \in H(\text{curl}; \Omega)$.

Proof. Taking the Laplace transform of (3.123) and (3.124), we obtain

$$\epsilon_0(s\hat{\mathbf{E}} - \mathbf{E}_0) + s\hat{\mathbf{P}} - \mathbf{P}_0 - \nabla \times \hat{\mathbf{H}} = 0 \tag{3.133}$$

$$\mu_0(s\hat{\mathbf{H}} - \mathbf{H}_0) + s\hat{\mathbf{M}} - \mathbf{M}_0 + \nabla \times \hat{\mathbf{E}} = 0. \tag{3.134}$$

Taking the Laplace transform of (3.125) and (3.126), we have

$$\hat{\mathbf{P}} = [(s + \Gamma_e)\mathbf{P}_0 + \mathbf{P}'(0) + \hat{\mathbf{E}}]/(s^2 + \Gamma_e s + \omega_{e0}^2) \tag{3.135}$$

$$\hat{\mathbf{M}} = [(s + \Gamma_m)\mathbf{M}_0 + \mathbf{M}'(0) + \hat{\mathbf{H}}]/(s^2 + \Gamma_m s + \omega_{m0}^2). \tag{3.136}$$

Substituting (3.135) into (3.133), we have

$$\begin{aligned}&(\epsilon_0 s + \frac{s}{s^2 + \Gamma_e s + \omega_{e0}^2})\hat{\mathbf{E}} - \nabla \times \hat{\mathbf{H}} \\ &= \epsilon_0 \mathbf{E}_0 + \mathbf{P}_0 - \frac{[(s + \Gamma_e)\mathbf{P}_0 + \mathbf{P}'(0)]s}{s^2 + \Gamma_e s + \omega_{e0}^2} \equiv \mathbf{f}_0, \qquad (3.137)\\ &(\mu_0 s + \frac{s}{s^2 + \Gamma_m s + \omega_{m0}^2})\hat{\mathbf{H}} + \nabla \times \hat{\mathbf{E}} \\ &= \mu_0 \mathbf{H}_0 + \mathbf{M}_0 - \frac{[(s + \Gamma_m)\mathbf{M}_0 + \mathbf{M}'(0)]s}{s^2 + \Gamma_m s + \omega_{m0}^2} \equiv \mathbf{g}_0. \qquad (3.138)\end{aligned}$$

Eliminating $\hat{\mathbf{H}}$ from (3.137) and (3.138), we obtain

$$\begin{aligned}&(\epsilon_0 s + \frac{s}{s^2 + \Gamma_e s + \omega_{e0}^2})(\mu_0 s + \frac{s}{s^2 + \Gamma_m s + \omega_{m0}^2})\hat{\mathbf{E}} + \nabla \times \nabla \times \hat{\mathbf{E}} \\ &= \nabla \times \mathbf{g}_0 + (\mu_0 s + \frac{s}{s^2 + \Gamma_m s + \omega_{m0}^2})\mathbf{f}_0, \qquad (3.139)\end{aligned}$$

whose weak formulation can be written as: Find $\hat{\mathbf{E}} \in H_0(\text{curl}; \Omega)$ such that

$$\begin{aligned}&(\epsilon_0 s + \frac{s}{s^2 + \Gamma_e s + \omega_{e0}^2})(\mu_0 s + \frac{s}{s^2 + \Gamma_m s + \omega_{m0}^2})(\hat{\mathbf{E}}, \boldsymbol{\phi}) + (\nabla \times \hat{\mathbf{E}}, \nabla \times \boldsymbol{\phi}) \\ &= (\nabla \times \mathbf{g}_0 + (\mu_0 s + \frac{s}{s^2 + \Gamma_m s + \omega_{m0}^2})\mathbf{f}_0, \boldsymbol{\phi}), \quad \forall\, \boldsymbol{\phi} \in H_0(\text{curl}; \Omega). \qquad (3.140)\end{aligned}$$

By the Lax-Milgram lemma, the Eq. (3.140) exists a unique solution $\hat{\mathbf{E}} \in H_0(\text{curl}; \Omega)$ for any $s > 0$. The existence and uniqueness of $\hat{\mathbf{H}} \in H(\text{curl}; \Omega)$ can be assured by using the same argument, in which case we just eliminate $\hat{\mathbf{E}}$ from (3.137) and (3.138). The solutions $\mathbf{E}$ and $\mathbf{H}$ are the inverse Laplace transforms of $\hat{\mathbf{E}}$ and $\hat{\mathbf{H}}$, respectively. □

3.6.2 The Crank-Nicolson Scheme and Error Analysis

3.6.2.1 The Scheme and Stability Analysis

We first consider a Crank-Nicolson type scheme: For $k = 1, 2, \cdots, N - 1$, find $\mathbf{E}_h^{k+1} \in \mathbf{V}_h^0, \mathbf{P}_h^{k+1} \in \mathbf{V}_h, \mathbf{H}_h^{k+1}, \mathbf{M}_h^{k+1} \in \mathbf{U}_h$ such that

$$\epsilon_0(\delta_{2\tau}\mathbf{E}_h^k, \boldsymbol{\phi}_h) + (\delta_{2\tau}\mathbf{P}_h^k, \boldsymbol{\phi}_h) - (\overline{\mathbf{H}}_h^k, \nabla \times \boldsymbol{\phi}_h) = 0, \qquad (3.141)$$

$$\mu_0(\delta_{2\tau}\mathbf{H}_h^k, \boldsymbol{\psi}_h) + (\delta_{2\tau}\mathbf{M}_h^k, \boldsymbol{\psi}_h) + (\nabla \times \overline{\mathbf{E}}_h^k, \boldsymbol{\psi}_h) = 0, \tag{3.142}$$

$$\frac{1}{\epsilon_0\omega_{pe}^2}[(\frac{\delta_\tau\mathbf{P}_h^{k+1} - \delta_\tau\mathbf{P}_h^k}{\tau}, \tilde{\boldsymbol{\phi}}_h) + \Gamma_e(\delta_{2\tau}\mathbf{P}_h^k, \tilde{\boldsymbol{\phi}}_h) + \omega_{e0}^2(\overline{\mathbf{P}}_h^k, \tilde{\boldsymbol{\phi}}_h)] = (\overline{\mathbf{E}}_h^k, \tilde{\boldsymbol{\phi}}_h), \tag{3.143}$$

$$\frac{1}{\mu_0\omega_{pm}^2}[(\frac{\delta_\tau\mathbf{M}_h^{k+1} - \delta_\tau\mathbf{M}_h^k}{\tau}, \tilde{\boldsymbol{\psi}}_h) + \Gamma_m(\delta_{2\tau}\mathbf{M}_h^k, \tilde{\boldsymbol{\psi}}_h) + \omega_{m0}^2(\overline{\mathbf{M}}_h^k, \tilde{\boldsymbol{\psi}}_h)] = (\overline{\mathbf{H}}_h^k, \tilde{\boldsymbol{\psi}}_h), \tag{3.144}$$

hold true for any $\boldsymbol{\phi}_h \in \mathbf{V}_h^0$, $\boldsymbol{\psi}_h \in \mathbf{U}_h$, $\tilde{\boldsymbol{\phi}}_h \in \mathbf{V}_h$ and $\tilde{\boldsymbol{\psi}}_h \in \mathbf{U}_h$. Here we denote

$$\delta_\tau u^k = (u^k - u^{k-1})/\tau, \quad \delta_{2\tau}u^k = (u^{k+1} - u^{k-1})/2\tau, \quad \overline{u}^k = (u^{k+1} + u^{k-1})/2.$$

To implement this scheme, we use the following initial approximations:

$$\mathbf{E}_h^0(\mathbf{x}) = \Pi_h\mathbf{E}_0(\mathbf{x}), \quad \mathbf{H}_h^0(\mathbf{x}) = P_h\mathbf{H}_0(\mathbf{x}), \tag{3.145}$$

$$\mathbf{P}_h^0(\mathbf{x}) = \Pi_h\mathbf{P}_0(\mathbf{x}), \quad \mathbf{M}_h^0(\mathbf{x}) = P_h\mathbf{M}_0(\mathbf{x}), \tag{3.146}$$

$$\begin{aligned}\mathbf{E}_h^1(\mathbf{x}) &= \Pi_h\mathbf{E}(\mathbf{x},\tau) \approx \Pi_h(\mathbf{E}(\mathbf{x},0) + \tau\mathbf{E}_t(\mathbf{x},0)) \\ &= \Pi_h[\mathbf{E}_0(\mathbf{x}) + \tau\epsilon_0^{-1}(\nabla \times \mathbf{H}_0(\mathbf{x}) - \mathbf{P}_1(\mathbf{x}))],\end{aligned} \tag{3.147}$$

$$\mathbf{H}_h^1(\mathbf{x}) = P_h\mathbf{H}(\mathbf{x},\tau) \approx P_h[\mathbf{H}_0(\mathbf{x}) - \tau\mu_0^{-1}(\nabla \times \mathbf{E}_0(\mathbf{x}) + \mathbf{M}_1(\mathbf{x}))], \tag{3.148}$$

$$\begin{aligned}\mathbf{P}_h^1(\mathbf{x}) &= \Pi_h\mathbf{P}(\mathbf{x},\tau) \approx \Pi_h(\mathbf{P}(\mathbf{x},0) + \tau\mathbf{P}_t(\mathbf{x},0) + \frac{\tau^2}{2}\mathbf{P}_{tt}(\mathbf{x},0)) \\ &= \Pi_h[\mathbf{P}_0(\mathbf{x}) + \tau\mathbf{P}_1(\mathbf{x}) + \frac{\tau^2}{2}(\epsilon_0\omega_{pe}^2\mathbf{E}_0 - \omega_{e0}^2\mathbf{P}_0 - \Gamma_e\mathbf{P}_1)],\end{aligned} \tag{3.149}$$

$$\mathbf{M}_h^1(\mathbf{x}) = P_h[\mathbf{M}_0(\mathbf{x}) + \tau\mathbf{M}_1(\mathbf{x}) + \frac{\tau^2}{2}(\mu_0\omega_{pm}^2\mathbf{H}_0 - \omega_{m0}^2\mathbf{M}_0 - \Gamma_m\mathbf{M}_1)]. \tag{3.150}$$

For this scheme, we have the following stability:

Lemma 3.21. *For any $k \geq 2$, we have*

$$\begin{aligned}&\epsilon_0||\mathbf{E}_h^k||_0^2 + \mu_0||\mathbf{H}_h^k||_0^2 + \frac{2}{\epsilon_0\omega_{pe}^2}||\delta_\tau\mathbf{P}_h^k||_0^2 \\ &\quad + \frac{2}{\mu_0\omega_{pm}^2}||\delta_\tau\mathbf{M}_h^k||_0^2 + \frac{\omega_{e0}^2}{\epsilon_0\omega_{pe}^2}||\mathbf{P}_h^k||_0^2 + \frac{\omega_{m0}^2}{\mu_0\omega_{pm}^2}||\mathbf{M}_h^k||_0^2 \\ \leq\ &\epsilon_0||\mathbf{E}_h^1||_0^2 + \mu_0||\mathbf{H}_h^1||_0^2 + \frac{\omega_{e0}^2}{\epsilon_0\omega_{pe}^2}||\mathbf{P}_h^1||_0^2 + \frac{\omega_{m0}^2}{\mu_0\omega_{pm}^2}||\mathbf{M}_h^1||_0^2 \\ &+\epsilon_0||\mathbf{E}_h^0||_0^2 + \mu_0||\mathbf{H}_h^0||_0^2 + \frac{2}{\epsilon_0\omega_{pe}^2}||\delta_\tau\mathbf{P}_h^1||_0^2\end{aligned}$$

$$+\frac{2}{\mu_0\omega_{pm}^2}||\delta_\tau \mathbf{M}_h^1||_0^2+\frac{\omega_{e0}^2}{\epsilon_0\omega_{pe}^2}||\mathbf{P}_h^0||_0^2+\frac{\omega_{m0}^2}{\mu_0\omega_{pm}^2}||\mathbf{M}_h^0||_0^2.$$

Proof. Choosing $\boldsymbol{\phi}_h=\tau(\mathbf{E}_h^{k+1}+\mathbf{E}_h^{k-1})$, $\boldsymbol{\psi}_h=\tau(\mathbf{H}_h^{k+1}+\mathbf{H}_h^{k-1})$, $\tilde{\boldsymbol{\phi}}_h=\tau(\delta_\tau\mathbf{P}_h^{k+1}+\delta_\tau\mathbf{P}_h^k)$, $\tilde{\boldsymbol{\psi}}_h=\tau(\delta_\tau\mathbf{M}_h^{k+1}+\delta_\tau\mathbf{M}_h^k)$ in (3.141)–(3.144), respectively, and adding the resultants together, we have

$$\begin{aligned}
&\frac{\epsilon_0}{2}(||\mathbf{E}_h^{k+1}||_0^2-||\mathbf{E}_h^{k-1}||_0^2)+\frac{\mu_0}{2}(||\mathbf{H}_h^{k+1}||_0^2-||\mathbf{H}_h^{k-1}||_0^2)\\
&+\frac{1}{\epsilon_0\omega_{pe}^2}(||\delta_\tau\mathbf{P}_h^{k+1}||_0^2-||\delta_\tau\mathbf{P}_h^k||_0^2)+\frac{1}{\mu_0\omega_{pm}^2}(||\delta_\tau\mathbf{M}_h^{k+1}||_0^2-||\delta_\tau\mathbf{M}_h^k||_0^2)\\
&+\frac{\omega_{e0}^2}{2\epsilon_0\omega_{pe}^2}(||\mathbf{P}_h^{k+1}||_0^2-||\mathbf{P}_h^{k-1}||_0^2)+\frac{\omega_{m0}^2}{2\mu_0\omega_{pm}^2}(||\mathbf{M}_h^{k+1}||_0^2-||\mathbf{M}_h^{k-1}||_0^2)\le 0.
\end{aligned}$$

Summing up the above estimate from $k=1$ to $k=n-1$, and using the identity

$$\sum_{k=1}^{n}(a_{k+1}^2-a_{k-1}^2)=a_{n+1}^2+a_n^2-a_1^2-a_0^2,$$

we can easily see that the proof completes. □

Let us look at the scheme carefully. From (3.143), we have

$$\mathbf{P}_h^{k+1}=a_1(\mathbf{E}_h^{k+1}+\mathbf{E}_h^{k-1})+a_2\mathbf{P}_h^k-a_3\mathbf{P}_h^{k-1},\tag{3.151}$$

where

$$a_1=\frac{\epsilon_0\omega_{pe}^2\tau^2}{2+\tau\Gamma_e+\tau^2\omega_{e0}^2},\ a_2=\frac{4}{2+\tau\Gamma_e+\tau^2\omega_{e0}^2},\ a_3=\frac{2-\tau\Gamma_e+\tau^2\omega_{e0}^2}{2+\tau\Gamma_e+\tau^2\omega_{e0}^2}.$$

Similarly, from (3.144), we have

$$\mathbf{M}_h^{k+1}=\tilde{a}_1(\mathbf{H}_h^{k+1}+\mathbf{H}_h^{k-1})+\tilde{a}_2\mathbf{M}_h^k-\tilde{a}_3\mathbf{M}_h^{k-1},\tag{3.152}$$

where

$$\tilde{a}_1=\frac{\mu_0\omega_{pm}^2\tau^2}{2+\tau\Gamma_m+\tau^2\omega_{m0}^2},\ \tilde{a}_2=\frac{4}{2+\tau\Gamma_m+\tau^2\omega_{m0}^2},\ \tilde{a}_3=\frac{2-\tau\Gamma_m+\tau^2\omega_{m0}^2}{2+\tau\Gamma_m+\tau^2\omega_{m0}^2}.$$

Substituting (3.151) and (3.152) into (3.141) and (3.142), respectively, we obtain

$$\begin{aligned}
&(\epsilon_0+a_1)(\mathbf{E}_h^{k+1},\boldsymbol{\phi}_h)-\tau(\mathbf{H}_h^{k+1},\nabla\times\boldsymbol{\phi}_h)=(\epsilon_0-a_1)(\mathbf{E}_h^{k-1},\boldsymbol{\phi}_h)\\
&-a_2(\mathbf{P}_h^k,\boldsymbol{\phi}_h)+(1+a_3)(\mathbf{P}_h^{k-1},\boldsymbol{\phi}_h)+\tau(\mathbf{H}_h^{k-1},\nabla\times\boldsymbol{\phi}_h),
\end{aligned}\tag{3.153}$$

and

$$(\mu_0+\tilde{a}_1)(\mathbf{H}_h^{k+1},\boldsymbol{\psi}_h)+\tau(\nabla\times\mathbf{E}_h^{k+1},\boldsymbol{\psi}_h)=(\mu_0-\tilde{a}_1)(\mathbf{H}_h^{k-1},\boldsymbol{\psi}_h)$$
$$-\tilde{a}_2(\mathbf{M}_h^{k},\boldsymbol{\psi}_h)+(1+\tilde{a}_3)(\mathbf{M}_h^{k-1},\boldsymbol{\psi}_h)-\tau(\nabla\times\mathbf{E}_h^{k-1},\boldsymbol{\psi}_h),\tag{3.154}$$

whose system's coefficient matrix can be written as $\begin{pmatrix} A & -B' \\ B & D \end{pmatrix}$, where matrices $A=(\epsilon_0+a_1)(\boldsymbol{\phi}_h,\boldsymbol{\phi}_h)$ and $D=(\mu_0+\tilde{a}_1)(\boldsymbol{\psi}_h,\boldsymbol{\psi}_h)$ are symmetric positive definite, and matrix $B=\tau(\nabla\times\boldsymbol{\phi}_h,\boldsymbol{\psi}_h)$. Here B' denotes the transpose of B. It is easy to see that the determinant of the coefficient matrix is

$$\begin{vmatrix} A & -B' \\ B & D \end{vmatrix}=|A|\cdot\begin{vmatrix} I & -A^{-1}B' \\ B & D \end{vmatrix}=|A|\cdot\begin{vmatrix} I & -A^{-1}B' \\ 0 & D+BA^{-1}B' \end{vmatrix}=|A|\cdot|D+BA^{-1}B'|,$$

which is non-zero. Hence in practical implementation of the scheme (3.141)–(3.144), at each time step we can solve (3.153) and (3.154) for $(\mathbf{E}_h^{k+1},\mathbf{H}_h^{k+1})$, then update $\mathbf{P}_h^{k+1}$ and $\mathbf{M}_h^{k+1}$ using (3.151) and (3.152), respectively.

3.6.2.2 The Optimal Error Estimate

Before proving the optimal error estimate for the Crank-Nicolson scheme (3.141)–(3.144), we need some estimates.

Lemma 3.22 ([184, Lemma 4.3]). *For any $p(x,t)\in H^3(0,T;L^2(\Omega))$ and $k\geq 1$, we have*

$$||(\frac{p^{k+1}+p^k}{2}-p^{k+\frac{1}{2}})-(\frac{p^k+p^{k-1}}{2}-p^{k-\frac{1}{2}})||_0^2\leq\frac{\tau^5}{8}\int_{t_{k-1}}^{t_{k+1}}||p_{s^3}(s)||_0^2ds.$$

Proof. By Taylor expansions, it is easy to see the following identity is true:

$$\frac{p(t)+p(t-\tau)}{2}-p(t-\frac{\tau}{2})=\frac{1}{2}[\int_{t-\frac{\tau}{2}}^{t}(t-s)p_{s^2}(s)ds+\int_{t-\tau}^{t-\frac{\tau}{2}}(s-t+\tau)p_{s^2}(s)ds],$$

applying which to p_t we obtain

$$\begin{aligned}&||(\frac{p^{k+1}+p^k}{2}-p^{k+\frac{1}{2}})-(\frac{p^k+p^{k-1}}{2}-p^{k-\frac{1}{2}})||_0^2\\&=||\int_{t_k}^{t_{k+1}}\frac{d}{dt}(\frac{p(t)+p(t-\tau)}{2}-p(t-\frac{\tau}{2}))dt||_0^2\\&=||\frac{1}{2}\int_{t_k}^{t_{k+1}}[\int_{t-\frac{\tau}{2}}^{t}(t-s)p_{s^3}(s)ds+\int_{t-\tau}^{t-\frac{\tau}{2}}(s-t+\tau)p_{s^3}(s)ds]dt||_0^2\end{aligned}$$

$$\leq \frac{\tau}{4}\int_{t_k}^{t_{k+1}} 2[||\int_{t-\frac{\tau}{2}}^{t}(t-s)p_{s^3}(s)ds||_0^2 + ||\int_{t-\tau}^{t-\frac{\tau}{2}}(s-t+\tau)p_{s^3}(s)ds||_0^2]dt$$

$$\leq \frac{\tau}{2}\int_{t_k}^{t_{k+1}} [\frac{\tau}{2}||\int_{t-\frac{\tau}{2}}^{t}(t-s)^2||p_{s^3}(s)||_0^2 ds + \frac{\tau}{2}\int_{t-\tau}^{t-\frac{\tau}{2}}(s-t+\tau)^2||p_{s^3}(s)||_0^2 ds]dt$$

$$\leq \frac{\tau^3}{2}\cdot\frac{\tau^2}{4}\int_{t_{k-1}}^{t_{k+1}}||p_{s^3}(s)||_0^2 ds,$$

from which the proof completes. □

By similar arguments, the following two estimates can be proved (cf. [184]).

Lemma 3.23. *For any $p(x,t) \in H^4(0,T;L^2(\Omega))$ and $k \geq 1$, we have*

$$||(\delta_\tau p^{k+1} - p_t^{k+\frac{1}{2}}) - (\delta_\tau p^k - p_t^{k-\frac{1}{2}})||_0^2 \leq \frac{\tau^5}{32}\int_{t_{k-1}}^{t_{k+1}}||p_{s^4}(s)||_0^2 ds.$$

Lemma 3.24. *For any $p(x,t) \in H^2(0,T;L^2(\Omega))$ and $k \geq 1$, we have*

$$||\frac{p^{k+1}+p^{k-1}}{2} - p^k||_0^2 \leq \tau^3\int_{t_{k-1}}^{t_{k+1}}||p_{s^2}(s)||_0^2 ds.$$

With the above estimates and proper regularity assumptions, we can prove the following optimal error estimate.

Theorem 3.14. *Let $(\mathbf{E}^m, \mathbf{H}^m, \mathbf{P}^m, \mathbf{M}^m)$ and $(\mathbf{E}_h^m, \mathbf{H}_h^m, \mathbf{P}_h^m, \mathbf{M}_h^m)$ be the analytic and numerical solutions of (3.129)–(3.132) and (3.141)–(3.144) at time t_m, respectively. Under the regularity assumptions*

$$\mathbf{E} \in H^1(0,T;H^l(curl;\Omega)) \cap L^\infty(0,T;H^l(curl;\Omega)) \cap H^2(0,T;H(curl;\Omega)),$$
$$\mathbf{P} \in H^2(0,T;H^l(curl;\Omega)) \cap L^\infty(0,T;H^l(curl;\Omega)) \cap H^4(0,T;(L^2(\Omega))^3),$$
$$\mathbf{H} \in H^2(0,T;H(curl;\Omega)) \cap L^\infty(0,T;(H^l(\Omega))^3),$$
$$\mathbf{M} \in H^4(0,T;(L^2(\Omega))^3) \cap L^\infty(0,T;(H^l(\Omega))^3),$$

there exists a constant $C > 0$ independent of mesh size h and time step τ, such that

$$\epsilon_0||\mathbf{E}^n - \mathbf{E}_h^n||_0^2 + \frac{\omega_{e0}^2}{\epsilon_0\omega_{pe}^2}||\mathbf{P}^n - \mathbf{P}_h^n||_0^2 + \mu_0||\mathbf{H}^n - \mathbf{H}_h^n||_0^2 + \frac{\omega_{m0}^2}{\mu_0\omega_{pm}^2}||\mathbf{M}^n - \mathbf{M}_h^n||_0^2$$
$$\leq C(h^{2l} + \tau^4) + C(||\xi_h^0||_0^2 + ||\tilde{\xi}_h^0||_0^2 + ||\eta_h^0||_0^2 + ||\tilde{\eta}_h^0||_0^2),$$
$$+C[||\xi_h^1||_0^2 + ||\delta_\tau\tilde{\xi}_h^1||_0^2 + ||\tilde{\xi}_h^1||_0^2 + ||\eta_h^1||_0^2 + ||\delta_\tau\tilde{\eta}_h^1||_0^2 + ||\tilde{\eta}_h^1||_0^2],$$

where $\xi_h^k = \Pi_h \mathbf{E}^k - \mathbf{E}_h^k$, $\eta_h^k = P_h \mathbf{H}^k - \mathbf{H}_h^k$, $\tilde{\xi}_h^k = \Pi_h \mathbf{P}^k - \mathbf{P}_h^k$, $\tilde{\eta}_h^k = P_h \mathbf{M}^k - \mathbf{M}_h^k$, *and* $l \geq 1$ *is the order of the basis functions in spaces* $\mathbf{U}_h$ *and* $\mathbf{V}_h$.

Remark 3.3. Note that the initial approximations (3.145)–(3.150) yield the initial errors

$$||\xi_h^0||_0 = ||\tilde{\xi}_h^0||_0 = ||\eta_h^0||_0 = ||\tilde{\eta}_h^0||_0 = 0,$$
$$||\xi_h^1||_0 = ||\delta_\tau \tilde{\xi}_h^1||_0 = ||\tilde{\xi}_h^1||_0 = ||\eta_h^1||_0 = ||\delta_\tau \tilde{\eta}_h^1||_0 = ||\tilde{\eta}_h^1||_0 = O(h^l + \tau^2),$$

from which we have the optimal error estimate

$$||\mathbf{E} - \mathbf{E}_h^n||_0 + ||\mathbf{P} - \mathbf{P}_h^n||_0 + ||\mathbf{H} - \mathbf{H}_h^n||_0 + ||\mathbf{M} - \mathbf{M}_h^n||_0 \leq C(h^l + \tau^2).$$

Proof. Integrating (3.129) and (3.130) in time from $t = t_{k-1}$ to t_{k+1} and dividing all by 2τ, and integrating (3.131) and (3.132) in time from $t = t_{k-\frac{1}{2}}$ to $t = t_{k+\frac{1}{2}}$ and then dividing by τ, we have

$$\epsilon_0(\delta_{2\tau}\mathbf{E}^k, \boldsymbol{\phi}) + (\delta_{2\tau}\mathbf{P}^k, \boldsymbol{\phi}) - (\frac{1}{2\tau}\int_{t_{k-1}}^{t_{k+1}} \mathbf{H}(s)ds, \nabla \times \boldsymbol{\phi}) = 0, \tag{3.155}$$

$$\mu_0(\delta_{2\tau}\mathbf{H}^k, \boldsymbol{\psi}) + (\delta_{2\tau}\mathbf{M}^k, \boldsymbol{\psi}) + (\frac{1}{2\tau}\int_{t_{k-1}}^{t_{k+1}} \nabla \times \mathbf{E}(s)ds, \boldsymbol{\psi}) = 0, \tag{3.156}$$

$$\begin{aligned}&\frac{1}{\tau\epsilon_0\omega_{pe}^2}(\frac{\partial \mathbf{P}^{k+\frac{1}{2}}}{\partial t} - \frac{\partial \mathbf{P}^{k-\frac{1}{2}}}{\partial t}, \tilde{\boldsymbol{\phi}}) + \frac{\Gamma_e}{\epsilon_0\omega_{pe}^2}(\frac{\mathbf{P}^{k+\frac{1}{2}} - \mathbf{P}^{k-\frac{1}{2}}}{\tau}, \tilde{\boldsymbol{\phi}})\\ &+\frac{\omega_{e0}^2}{\epsilon_0\omega_{pe}^2}(\frac{1}{\tau}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}} \mathbf{P}(s)ds, \tilde{\boldsymbol{\phi}}) - (\frac{1}{\tau}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}} \mathbf{E}(s)ds, \tilde{\boldsymbol{\phi}}) = 0,\end{aligned} \tag{3.157}$$

$$\begin{aligned}&\frac{1}{\tau\mu_0\omega_{pm}^2}(\frac{\partial \mathbf{M}^{k+\frac{1}{2}}}{\partial t} - \frac{\partial \mathbf{M}^{k-\frac{1}{2}}}{\partial t}, \tilde{\boldsymbol{\psi}}) + \frac{\Gamma_m}{\mu_0\omega_{pm}^2}(\frac{\mathbf{M}^{k+\frac{1}{2}} - \mathbf{M}^{k-\frac{1}{2}}}{\tau}, \tilde{\boldsymbol{\psi}})\\ &+\frac{\omega_{m0}^2}{\mu_0\omega_{pm}^2}(\frac{1}{\tau}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}} \mathbf{M}(s)ds, \tilde{\boldsymbol{\psi}}) - (\frac{1}{\tau}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}} \mathbf{H}(s)ds, \tilde{\boldsymbol{\psi}}) = 0.\end{aligned} \tag{3.158}$$

Denote $\xi_h^k = \Pi_h \mathbf{E}^k - \mathbf{E}_h^k$, $\eta_h^k = P_h \mathbf{H}^k - \mathbf{H}_h^k$, $\tilde{\xi}_h^k = \Pi_h \mathbf{P}^k - \mathbf{P}_h^k$, $\tilde{\eta}_h^k = P_h \mathbf{M}^k - \mathbf{M}_h^k$. Subtracting (3.141)–(3.144) from (3.155)–(3.158) with $\boldsymbol{\phi} = \boldsymbol{\phi}_h$, $\boldsymbol{\psi} = \boldsymbol{\psi}_h$, $\tilde{\boldsymbol{\phi}} = \tilde{\boldsymbol{\phi}}_h$, and $\tilde{\boldsymbol{\psi}} = \tilde{\boldsymbol{\psi}}_h$, using the property of operator P_h, we obtain the error equations

$$\begin{aligned}(i)\quad &\epsilon_0(\frac{\xi_h^{k+1} - \xi_h^{k-1}}{2\tau}, \boldsymbol{\phi}_h) + (\frac{\tilde{\xi}_h^{k+1} - \tilde{\xi}_h^{k-1}}{2\tau}, \boldsymbol{\phi}_h) - \frac{1}{2}(\eta_h^{k+1} + \eta_h^{k-1}, \nabla \times \boldsymbol{\phi}_h)\\ &= \epsilon_0(\frac{(\Pi_h \mathbf{E}^{k+1} - \Pi_h \mathbf{E}^{k-1}) - (\mathbf{E}^{k+1} - \mathbf{E}^{k-1})}{2\tau}, \boldsymbol{\phi}_h)\end{aligned}$$

$$+(\frac{(\Pi_h\mathbf{P}^{k+1}-\Pi_h\mathbf{P}^{k-1})-(\mathbf{P}^{k+1}-\mathbf{P}^{k-1})}{2\tau},\boldsymbol{\phi}_h)$$
$$-(\frac{1}{2}(\mathbf{H}^{k+1}+\mathbf{H}^{k-1})-\frac{1}{2\tau}\int_{t_{k-1}}^{t_{k+1}}\mathbf{H}(s)ds,\nabla\times\boldsymbol{\phi}_h), \tag{3.159}$$

$$(ii)\quad \mu_0(\frac{\eta_h^{k+1}-\eta_h^{k-1}}{2\tau},\boldsymbol{\psi}_h)+(\frac{\tilde{\eta}_h^{k+1}-\tilde{\eta}_h^{k-1}}{2\tau},\boldsymbol{\psi}_h)+\frac{1}{2}(\nabla\times(\xi_h^{k+1}+\xi_h^{k-1}),\boldsymbol{\psi}_h)$$
$$=(\frac{1}{2}(\nabla\times\Pi_h\mathbf{E}^{k+1}+\nabla\times\Pi_h\mathbf{E}^{k-1})-\frac{1}{2\tau}\int_{t_{k-1}}^{t_{k+1}}\nabla\times\mathbf{E}(s)ds,\boldsymbol{\psi}_h), \tag{3.160}$$

$$(iii)\quad \frac{1}{\tau\epsilon_0\omega_{pe}^2}(\delta_\tau\tilde{\xi}_h^{k+1}-\delta_\tau\tilde{\xi}_h^{k},\tilde{\boldsymbol{\phi}}_h)+\frac{\Gamma_e}{\epsilon_0\omega_{pe}^2}(\frac{\tilde{\xi}_h^{k+1}-\tilde{\xi}_h^{k-1}}{2\tau},\tilde{\boldsymbol{\phi}}_h)$$
$$+\frac{\omega_{e0}^2}{\epsilon_0\omega_{pe}^2}(\frac{\tilde{\xi}_h^{k+1}+\tilde{\xi}_h^{k-1}}{2},\tilde{\boldsymbol{\phi}}_h)-(\frac{\xi_h^{k+1}+\xi_h^{k-1}}{2},\tilde{\boldsymbol{\phi}}_h)$$
$$=\frac{1}{\tau\epsilon_0\omega_{pe}^2}((\delta_\tau\Pi_h\mathbf{P}^{k+1}-\delta_\tau\Pi_h\mathbf{P}^{k})-(\mathbf{P}_t^{k+\frac{1}{2}}-\mathbf{P}_t^{k-\frac{1}{2}}),\tilde{\boldsymbol{\phi}}_h)$$
$$+\frac{\Gamma_e}{\epsilon_0\omega_{pe}^2}(\frac{\Pi_h\mathbf{P}^{k+1}-\Pi_h\mathbf{P}^{k-1}}{2\tau}-\frac{\mathbf{P}^{k+\frac{1}{2}}-\mathbf{P}^{k-\frac{1}{2}}}{\tau},\tilde{\boldsymbol{\phi}}_h)$$
$$+\frac{\omega_{e0}^2}{\epsilon_0\omega_{pe}^2}(\frac{\Pi_h\mathbf{P}^{k+1}+\Pi_h\mathbf{P}^{k-1}}{2}-\frac{1}{\tau}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}}\mathbf{P}(s)ds,\tilde{\boldsymbol{\phi}}_h)$$
$$-(\frac{\Pi_h\mathbf{E}^{k+1}+\Pi_h\mathbf{E}^{k-1}}{2}-\frac{1}{\tau}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}}\mathbf{E}(s)ds,\tilde{\boldsymbol{\phi}}_h), \tag{3.161}$$

$$(iv)\quad \frac{1}{\tau\mu_0\omega_{pm}^2}(\delta_\tau\tilde{\eta}_h^{k+1}-\delta_\tau\tilde{\eta}_h^{k},\tilde{\boldsymbol{\psi}}_h)+\frac{\Gamma_m}{\mu_0\omega_{pm}^2}(\frac{\tilde{\eta}_h^{k+1}-\tilde{\eta}_h^{k-1}}{2\tau},\tilde{\boldsymbol{\psi}}_h)$$
$$+\frac{\omega_{m0}^2}{\mu_0\omega_{pm}^2}(\frac{\tilde{\eta}_h^{k+1}+\tilde{\eta}_h^{k-1}}{2},\tilde{\boldsymbol{\psi}}_h)-(\frac{\eta_h^{k+1}+\eta_h^{k-1}}{2},\tilde{\boldsymbol{\psi}}_h)$$
$$=\frac{1}{\tau\mu_0\omega_{pm}^2}((\delta_\tau\mathbf{M}^{k+1}-\delta_\tau\mathbf{M}^{k})-(\mathbf{M}_t^{k+\frac{1}{2}}-\mathbf{M}_t^{k-\frac{1}{2}}),\tilde{\boldsymbol{\psi}}_h)$$
$$+\frac{\Gamma_m}{\mu_0\omega_{pm}^2}(\frac{\mathbf{M}^{k+1}-\mathbf{M}^{k-1}}{2\tau}-\frac{\mathbf{M}^{k+\frac{1}{2}}-\mathbf{M}^{k-\frac{1}{2}}}{\tau},\tilde{\boldsymbol{\psi}}_h)$$

$$+\frac{\omega_{m0}^2}{\epsilon_0\omega_{pe}^2}\left(\frac{\mathbf{M}^{k+1}+\mathbf{M}^{k-1}}{2}-\frac{1}{\tau}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}}\mathbf{M}(s)ds,\tilde{\boldsymbol{\psi}}_h\right)$$

$$-\left(\frac{\mathbf{H}^{k+1}+\mathbf{H}^{k-1}}{2}-\frac{1}{\tau}\int_{t_{k-\frac{1}{2}}}^{t_{k+\frac{1}{2}}}\mathbf{H}(s)ds,\tilde{\boldsymbol{\psi}}_h\right). \tag{3.162}$$

Choosing $\phi_h=\tau(\xi_h^{k+1}+\xi_h^{k-1})$, $\psi_h=\tau(\eta_h^{k+1}+\eta_h^{k-1})$, $\tilde{\phi}_h=\tau(\delta_\tau\tilde{\xi}_h^{k+1}+\delta_\tau\tilde{\xi}_h^{k})$, $\tilde{\psi}_h=\tau(\delta_\tau\tilde{\eta}_h^{k+1}+\delta_\tau\tilde{\eta}_h^{k})$ in the above error equations, and adding the resultants together, we obtain

$$\begin{aligned}
&\frac{\epsilon_0}{2}(||\xi_h^{k+1}||_0^2-||\xi_h^{k-1}||_0^2)+\frac{1}{\epsilon_0\omega_{pe}^2}(||\delta_\tau\tilde{\xi}_h^{k+1}||_0^2-||\delta_\tau\tilde{\xi}_h^{k}||_0^2)\\
&\quad+\frac{\omega_{e0}^2}{2\epsilon_0\omega_{pe}^2}(||\tilde{\xi}_h^{k+1}||_0^2-||\tilde{\xi}_h^{k-1}||_0^2)+\frac{\mu_0}{2}(||\eta_h^{k+1}||_0^2-||\eta_h^{k-1}||_0^2)\\
&\quad+\frac{1}{\mu_0\omega_{pm}^2}(||\delta_\tau\tilde{\eta}_h^{k+1}||_0^2-||\delta_\tau\tilde{\eta}_h^{k}||_0^2)+\frac{\omega_{m0}^2}{2\mu_0\omega_{pm}^2}(||\tilde{\eta}_h^{k+1}||_0^2-||\tilde{\eta}_h^{k-1}||_0^2)\\
&\le\sum_{i=1}^{12}Err_i,
\end{aligned} \tag{3.163}$$

where Err_i are those right hand side terms from (3.159) to (3.162).

The proof can be done by carefully estimating all Err_i. Details can be seen in the original paper [184]. □

3.6.3 Some Other Schemes

By introducing the induced electric and magnetic currents $\mathbf{J}=\frac{\partial\mathbf{P}}{\partial t}$ and $\mathbf{K}=\frac{\partial\mathbf{M}}{\partial t}$, respectively, we can rewrite the Lorentz model (3.123)–(3.126) as

$$\epsilon_0\frac{\partial\mathbf{E}}{\partial t}=\nabla\times\mathbf{H}-\mathbf{J} \tag{3.164}$$

$$\mu_0\frac{\partial\mathbf{H}}{\partial t}=-\nabla\times\mathbf{E}-\mathbf{K} \tag{3.165}$$

$$\frac{1}{\epsilon_0\omega_{pe}^2}\frac{\partial\mathbf{J}}{\partial t}+\frac{\Gamma_e}{\epsilon_0\omega_{pe}^2}\mathbf{J}=\mathbf{E}-\frac{\omega_{e0}^2}{\epsilon_0\omega_{pe}^2}\mathbf{P} \tag{3.166}$$

$$\frac{\omega_{e0}^2}{\epsilon_0\omega_{pe}^2}\frac{\partial\mathbf{P}}{\partial t}=\frac{\omega_{e0}^2}{\epsilon_0\omega_{pe}^2}\mathbf{J} \tag{3.167}$$

$$\frac{1}{\mu_0\omega_{pm}^2}\frac{\partial\mathbf{K}}{\partial t}+\frac{\Gamma_m}{\mu_0\omega_{pm}^2}\mathbf{K}=\mathbf{H}-\frac{\omega_{m0}^2}{\mu_0\omega_{pm}^2}\mathbf{M} \tag{3.168}$$

$$\frac{\omega_{m0}^2}{\mu_0\omega_{pm}^2}\frac{\partial \mathbf{M}}{\partial t} = \frac{\omega_{m0}^2}{\mu_0\omega_{pm}^2}\mathbf{K}. \tag{3.169}$$

Multiplying the above equations by $\mathbf{E}, \mathbf{H}, \mathbf{J}, \mathbf{P}, \mathbf{K}$ and $\mathbf{M}$, respectively, then integrating over Ω and summing up the resultants, we can easily obtain the following stability.

Lemma 3.25.

$$\begin{aligned}
&\epsilon_0||\mathbf{E}(t)||_0^2 + \mu_0||\mathbf{H}(t)||_0^2 + \frac{1}{\epsilon_0\omega_{pe}^2}||\mathbf{J}(t)||_0^2 \\
&\quad + \frac{1}{\mu_0\omega_{pm}^2}||\mathbf{K}(t)||_0^2 + \frac{\omega_{e0}^2}{\epsilon_0\omega_{pe}^2}||\mathbf{P}(t)||_0^2 + \frac{\omega_{m0}^2}{\mu_0\omega_{pm}^2}||\mathbf{M}(t)||_0^2 \\
\le\ &\epsilon_0||\mathbf{E}(0)||_0^2 + \mu_0||\mathbf{H}(0)||_0^2 + \frac{1}{\epsilon_0\omega_{pe}^2}||\mathbf{J}(0)||_0^2 \\
&\quad + \frac{1}{\mu_0\omega_{pm}^2}||\mathbf{K}(0)||_0^2 + \frac{\omega_{e0}^2}{\epsilon_0\omega_{pe}^2}||\mathbf{P}(0)||_0^2 + \frac{\omega_{m0}^2}{\mu_0\omega_{pm}^2}||\mathbf{M}(0)||_0^2.
\end{aligned}$$

We can construct a Crank-Nicolson scheme for solving the system (3.164)–(3.169): For $k = 1, 2, \cdots, N$, find $\mathbf{E}_h^k \in \mathbf{V}_h^0, \mathbf{J}_h^k, \mathbf{P}_h^k \in \mathbf{V}_h, \mathbf{H}_h^k, \mathbf{K}_h^k, \mathbf{M}_h^k \in \mathbf{U}_h$ such that

$$\epsilon_0(\delta_\tau \mathbf{E}_h^k, \boldsymbol{\phi}_h) - (\overline{\mathbf{H}}_h^k, \nabla \times \boldsymbol{\phi}_h) + (\overline{\mathbf{J}}_h^k, \boldsymbol{\phi}_h) = 0, \tag{3.170}$$

$$\mu_0(\delta_\tau \mathbf{H}_h^k, \boldsymbol{\psi}_h) + (\nabla \times \overline{\mathbf{E}}_h^k, \boldsymbol{\psi}_h) + (\overline{\mathbf{K}}_h^k, \boldsymbol{\psi}_h) = 0, \tag{3.171}$$

$$\frac{1}{\epsilon_0\omega_{pe}^2}[(\delta_\tau \mathbf{J}_h^k, \boldsymbol{\phi}_{1h}) + \Gamma_e(\overline{\mathbf{J}}_h^k, \boldsymbol{\phi}_{1h}) + \omega_{e0}^2(\overline{\mathbf{P}}_h^k, \boldsymbol{\phi}_{1h})] = (\overline{\mathbf{E}}_h^k, \boldsymbol{\phi}_{1h}), \tag{3.172}$$

$$\frac{\omega_{e0}^2}{\epsilon_0\omega_{pe}^2}(\delta_\tau \mathbf{P}_h^k, \boldsymbol{\phi}_{2h}) - \frac{\omega_{e0}^2}{\epsilon_0\omega_{pe}^2}(\overline{\mathbf{J}}_h^k, \boldsymbol{\phi}_{2h}) = 0, \tag{3.173}$$

$$\frac{1}{\mu_0\omega_{pm}^2}[(\delta_\tau \mathbf{K}_h^k, \boldsymbol{\psi}_{1h}) + \Gamma_m(\overline{\mathbf{K}}_h^k, \boldsymbol{\psi}_{1h}) + \omega_{m0}^2(\overline{\mathbf{M}}_h^k, \boldsymbol{\psi}_{1h})] = (\overline{\mathbf{H}}_h^k, \boldsymbol{\psi}_{1h}), \tag{3.174}$$

$$\frac{\omega_{m0}^2}{\mu_0\omega_{pm}^2}(\delta_\tau \mathbf{M}_h^k, \boldsymbol{\psi}_{2h}) - \frac{\omega_{m0}^2}{\mu_0\omega_{pm}^2}(\overline{\mathbf{K}}_h^k, \boldsymbol{\psi}_{2h}) = 0, \tag{3.175}$$

hold true for any $\boldsymbol{\phi}_h \in \mathbf{V}_h^0, \boldsymbol{\psi}_h, \boldsymbol{\psi}_{1h}, \boldsymbol{\psi}_{2h} \in \mathbf{U}_h, \boldsymbol{\phi}_{1h}, \boldsymbol{\phi}_{2h} \in \mathbf{V}_h$, and are subject to the initial approximations

$$\begin{aligned}
&\mathbf{E}_h^0(\mathbf{x}) = \Pi_h \mathbf{E}_0(\mathbf{x}), \ \ \mathbf{J}_h^0(\mathbf{x}) = \Pi_h \mathbf{J}_0(\mathbf{x}), \ \ \mathbf{P}_h^0(\mathbf{x}) = \Pi_h \mathbf{P}_0(\mathbf{x}), \\
&\mathbf{H}_h^0(\mathbf{x}) = P_h \mathbf{H}_0(\mathbf{x}), \ \ \mathbf{K}_h^0(\mathbf{x}) = P_h \mathbf{K}_0(\mathbf{x}), \ \ \mathbf{M}_h^0(\mathbf{x}) = P_h \mathbf{M}_0(\mathbf{x}).
\end{aligned}$$

Choosing $\boldsymbol{\phi}_h = \overline{\mathbf{E}}_h^k, \boldsymbol{\psi}_h = \overline{\mathbf{H}}_h^k, \boldsymbol{\phi}_{1h} = \overline{\mathbf{J}}_h^k, \boldsymbol{\phi}_{2h} = \overline{\mathbf{P}}_h^k, \boldsymbol{\psi}_{1h} = \overline{\mathbf{K}}_h^k, \boldsymbol{\psi}_{2h} = \overline{\mathbf{M}}_h^k$ in (3.170)–(3.175), respectively, and adding the resultants together, we can obtain the following discrete stability in exactly the same form as in the continuous case.

Lemma 3.26. *For any $k \geq 1$, we have*

$$\begin{aligned}
&\epsilon_0||\mathbf{E}_h^k||_0^2 + \mu_0||\mathbf{H}_h^k||_0^2 + \frac{1}{\epsilon_0\omega_{pe}^2}||\mathbf{J}_h^k||_0^2 \\
&\quad + \frac{1}{\mu_0\omega_{pm}^2}||\mathbf{K}_h^k||_0^2 + \frac{\omega_{e0}^2}{\epsilon_0\omega_{pe}^2}||\mathbf{P}_h^k||_0^2 + \frac{\omega_{m0}^2}{\mu_0\omega_{pm}^2}||\mathbf{M}_h^k||_0^2 \\
\leq\ &\epsilon_0||\mathbf{E}_h^0||_0^2 + \mu_0||\mathbf{H}_h^0||_0^2 + \frac{1}{\epsilon_0\omega_{pe}^2}||\mathbf{J}_h^0||_0^2 \\
&\quad + \frac{1}{\mu_0\omega_{pm}^2}||\mathbf{K}_h^0||_0^2 + \frac{\omega_{e0}^2}{\epsilon_0\omega_{pe}^2}||\mathbf{P}_h^0||_0^2 + \frac{\omega_{m0}^2}{\mu_0\omega_{pm}^2}||\mathbf{M}_h^0||_0^2.
\end{aligned}$$

We want to show that practical implementation of (3.170)–(3.175) is actually not that scary. Solving (3.173), we obtain

$$\mathbf{P}_h^k = \mathbf{P}_h^{k-1} + \frac{\tau}{2}(\mathbf{J}_h^k + \mathbf{J}_h^{k-1}). \tag{3.176}$$

Substituting (3.176) into (3.172) and simplifying the result, we have

$$\beta\mathbf{J}_h^k = (1 - \frac{\tau\Gamma_e}{2} - \frac{\tau^2\omega_{e0}^2}{4})\mathbf{J}_h^{k-1} - \tau\omega_{e0}^2\mathbf{P}_h^{k-1} + \frac{\tau\epsilon_0\omega_{pe}^2}{2}(\mathbf{E}_h^k + \mathbf{E}_h^{k-1}), \tag{3.177}$$

where we denote $\beta = 1 + \frac{\tau\Gamma_e}{2} + \frac{\tau^2\omega_{e0}^2}{4}$.

Then substituting (3.177) into (3.170) and simplifying the result, we have

$$\begin{aligned}
&\epsilon_0(1 + \frac{\tau^2\omega_{pe}^2}{4\beta})(\mathbf{E}_h^k, \boldsymbol{\phi}_h) - \frac{\tau}{2}(\mathbf{H}_h^k, \nabla\times\boldsymbol{\phi}_h) = \epsilon_0(1 - \frac{\tau^2\omega_{pe}^2}{4\beta})(\mathbf{E}_h^{k-1}, \boldsymbol{\phi}_h) \\
&+\frac{\tau}{2}(\mathbf{H}_h^{k-1}, \nabla\times\boldsymbol{\phi}_h) - \frac{1}{\beta}(\tau\mathbf{J}_h^{k-1} - \frac{\tau^2\omega_{e0}^2}{2}\mathbf{P}_h^{k-1}, \boldsymbol{\phi}_h).
\end{aligned} \tag{3.178}$$

Similarly, from (3.173) to (3.175), we can obtain

$$\mathbf{M}_h^k = \mathbf{M}_h^{k-1} + \frac{\tau}{2}(\mathbf{K}_h^k + \mathbf{K}_h^{k-1}), \tag{3.179}$$

$$\tilde{\beta}\mathbf{K}_h^k = (1 - \frac{\tau\Gamma_m}{2} - \frac{\tau^2\omega_{m0}^2}{4})\mathbf{K}_h^{k-1} - \tau\omega_{m0}^2\mathbf{M}_h^{k-1} + \frac{\tau\mu_0\omega_{me}^2}{2}(\mathbf{H}_h^k + \mathbf{H}_h^{k-1}), \tag{3.180}$$

$$\mu_0(1+\frac{\tau^2\omega_{me}^2}{4\tilde{\beta}})(\mathbf{H}_h^k,\boldsymbol{\psi}_h)+\frac{\tau}{2}(\nabla\times\mathbf{E}_h^k,\boldsymbol{\psi}_h)=\mu_0(1-\frac{\tau^2\omega_{me}^2}{4\tilde{\beta}})(\mathbf{H}_h^{k-1},\boldsymbol{\psi}_h)$$
$$-\frac{\tau}{2}(\nabla\times\mathbf{E}_h^{k-1},\boldsymbol{\psi}_h)-\frac{1}{\tilde{\beta}}(\tau\mathbf{K}_h^{k-1}-\frac{\tau^2\omega_{m0}^2}{2}\mathbf{M}_h^{k-1},\boldsymbol{\psi}_h). \tag{3.181}$$

Hence, at each time step, we first solve a system formed by (3.178) and (3.181) for $\mathbf{E}_h^k$ and $\mathbf{H}_h^k$; then use (3.177) and (3.180) to update $\mathbf{J}_h^k$ and $\mathbf{K}_h^k$; finally, use (3.176) and (3.179) to update $\mathbf{P}_h^k$ and $\mathbf{M}_h^k$.

With proper regularity assumption, we can similarly prove the following optimal error estimate:

$$||\mathbf{E}^n-\mathbf{E}_h^n||_0+||\mathbf{P}^n-\mathbf{P}_h^n||_0+||\mathbf{H}^n-\mathbf{H}_h^n||_0+||\mathbf{M}^n-\mathbf{M}_h^n||_0$$
$$+||\mathbf{J}^n-\mathbf{J}_h^n||_0+||\mathbf{K}^n-\mathbf{K}_h^n||_0\le C(h^l+\tau^2). \tag{3.182}$$

Finally, we like to mention that leap-frog type schemes can be constructed for solving the system (3.164)–(3.169). For example, one leap-frog scheme is given as following: For $k\ge 0$, find $\mathbf{J}_h^{k+\frac{1}{2}},\mathbf{P}_h^{k+1}\in\mathbf{V}_h,\mathbf{E}_h^k\in\mathbf{V}_h^0,\mathbf{K}_h^{k+\frac{1}{2}},\mathbf{M}_h^{k+1},\mathbf{H}_h^{k+\frac{3}{2}}\in\mathbf{U}_h$ such that

$$\frac{1}{\epsilon_0\omega_{pe}^2}\frac{\mathbf{J}_h^{k+\frac{1}{2}}-\mathbf{J}_h^{k-\frac{1}{2}}}{\tau}+\frac{\Gamma_e}{2\epsilon_0\omega_{pe}^2}(\mathbf{J}_h^{k+\frac{1}{2}}+\mathbf{J}_h^{k-\frac{1}{2}})=\mathbf{E}_h^k-\frac{\omega_{e0}^2}{\epsilon_0\omega_{pe}^2}\mathbf{P}_h^k,$$
$$\frac{\omega_{e0}^2}{\epsilon_0\omega_{pe}^2}\frac{\mathbf{P}_h^{k+1}-\mathbf{P}_h^k}{\tau}=\frac{\omega_{e0}^2}{\epsilon_0\omega_{pe}^2}\mathbf{J}_h^{k+\frac{1}{2}},$$
$$\epsilon_0(\frac{\mathbf{E}_h^{k+1}-\mathbf{E}_h^k}{\tau},\boldsymbol{\phi}_h)-(\mathbf{H}_h^{k+\frac{1}{2}},\nabla\times\boldsymbol{\phi}_h)+(\mathbf{J}_h^{k+\frac{1}{2}},\boldsymbol{\phi}_h)=0,$$
$$\frac{1}{\mu_0\omega_{pm}^2}\frac{\mathbf{K}_h^{k+\frac{1}{2}}-\mathbf{K}_h^{k-\frac{1}{2}}}{\tau}+\frac{\Gamma_m}{2\mu_0\omega_{pm}^2}(\mathbf{K}_h^{k+\frac{1}{2}}+\mathbf{K}_h^{k-\frac{1}{2}})$$
$$=\frac{1}{2}(\mathbf{H}_h^{k+\frac{1}{2}}+\mathbf{H}_h^{k-\frac{1}{2}})-\frac{\omega_{m0}^2}{\mu_0\omega_{pm}^2}\mathbf{M}_h^k,$$
$$\frac{\omega_{m0}^2}{\mu_0\omega_{pm}^2}\frac{\mathbf{M}_h^{k+1}-\mathbf{M}_h^k}{\tau}=\frac{\omega_{m0}^2}{\mu_0\omega_{pm}^2}\mathbf{K}_h^{k+\frac{1}{2}},$$
$$\mu_0(\frac{\mathbf{H}_h^{k+\frac{3}{2}}-\mathbf{H}_h^{k+\frac{1}{2}}}{\tau},\boldsymbol{\psi}_h)+(\nabla\times\mathbf{E}_h^{k+1},\boldsymbol{\psi}_h)+(\mathbf{K}_h^{k+\frac{1}{2}},\boldsymbol{\psi}_h)=0,$$

hold true for any $\boldsymbol{\phi}_h\in\mathbf{V}_h^0$ and $\boldsymbol{\psi}_h\in\mathbf{U}_h$. Readers are encouraged to carry out the stability and error analysis by following the technique developed in Sect. 3.5 for the Drude model.

3.7 Extensions to the Drude-Lorentz Model

3.7.1 The Well-Posedness

Recall from Chap. 1 that the governing equations for the Drude-Lorentz model are:

$$\epsilon_0 \frac{\partial \mathbf{E}}{\partial t} = \nabla \times \mathbf{H} - \mathbf{J}, \tag{3.183}$$

$$\mu_0 \frac{\partial \mathbf{H}}{\partial t} = -\nabla \times \mathbf{E} - \mathbf{K}, \tag{3.184}$$

$$\frac{1}{\mu_0 \omega_0^2 F} \frac{\partial \mathbf{K}}{\partial t} + \frac{\gamma}{\mu_0 \omega_0^2 F} \mathbf{K} + \frac{1}{\mu_0 F} \mathbf{M} = \mathbf{H}, \tag{3.185}$$

$$\frac{1}{\mu_0 F} \frac{\partial \mathbf{M}}{\partial t} = \frac{1}{\mu_0 F} \mathbf{K}, \tag{3.186}$$

$$\frac{1}{\epsilon_0 \omega_p^2} \frac{\partial \mathbf{J}}{\partial t} + \frac{\nu}{\epsilon_0 \omega_p^2} \mathbf{J} = \mathbf{E}. \tag{3.187}$$

To complete the problem, we assume that the perfect conducting boundary condition (3.59) is imposed, and the initial conditions

$$\mathbf{E}(\mathbf{x}, 0) = \mathbf{E}_0(\mathbf{x}), \quad \mathbf{H}(\mathbf{x}, 0) = \mathbf{H}_0(\mathbf{x}), \tag{3.188}$$

$$\mathbf{K}(\mathbf{x}, 0) = \mathbf{K}_0(\mathbf{x}), \quad \mathbf{M}(\mathbf{x}, 0) = \mathbf{M}_0(\mathbf{x}), \quad \mathbf{J}(\mathbf{x}, 0) = \mathbf{J}_0(\mathbf{x}), \tag{3.189}$$

hold true. Here $\mathbf{E}_0, \mathbf{H}_0, \mathbf{K}_0, \mathbf{M}_0$ and $\mathbf{J}_0$ are some given functions.

First, we have the following stability for our model problem (3.183)–(3.187).

Lemma 3.27. *For the solution* $(\mathbf{E}, \mathbf{H}, \mathbf{K}, \mathbf{M}, \mathbf{J})$ *of problem (3.183)–(3.187) subject to boundary condition (3.59) and initial conditions (3.188) and (3.189), the following stability holds true:*

$$\begin{aligned} &\epsilon_0 ||\mathbf{E}(t)||_0^2 + \mu_0 ||\mathbf{H}(t)||_0^2 + \frac{1}{\mu_0 \omega_0^2 F} ||\mathbf{K}(t)||_0^2 + \frac{1}{\mu_0 F} ||\mathbf{M}(t)||_0^2 + \frac{1}{\epsilon_0 \omega_p^2} ||\mathbf{J}(t)||_0^2 \\ &\leq \epsilon_0 ||\mathbf{E}_0||_0^2 + \mu_0 ||\mathbf{H}_0||_0^2 + \frac{1}{\mu_0 \omega_0^2 F} ||\mathbf{K}_0||_0^2 + \frac{1}{\mu_0 F} ||\mathbf{M}_0||_0^2 + \frac{1}{\epsilon_0 \omega_p^2} ||\mathbf{J}_0||_0^2. \end{aligned} \tag{3.190}$$

Proof. Note that the problem (3.183)–(3.187) can be rewritten as

$$\frac{\partial}{\partial t} \mathscr{A} \mathbf{u}(t) = (\mathscr{B} + \mathscr{C}) \mathbf{u}(t), \tag{3.191}$$

where vector $\mathbf{u}(t) = (\mathbf{E}, \mathbf{H}, \mathbf{K}, \mathbf{M}, \mathbf{J})'$, matrices

$$\mathscr{A} = \text{diag}(\epsilon_0 I_3, \mu_0 I_3, \frac{1}{\mu_0 \omega_0^2 F} I_3, \frac{1}{\mu_0 F} I_3, \frac{1}{\epsilon_0 \omega_p^2} I_3),$$

$$\mathscr{C} = \text{diag}(0_3, 0_3, -\frac{\gamma}{\mu_0 \omega_0^2 F} I_3, 0_3, -\frac{\nu}{\epsilon_0 \omega_p^2} I_3),$$

and

$$\mathscr{B} = \begin{pmatrix} 0_3 & \nabla\times & 0_3 & 0_3 & -I_3 \\ -\nabla\times & 0_3 & -I_3 & 0_3 & 0_3 \\ 0_3 & I_3 & 0_3 & -\frac{1}{\mu_0 F} I_3 & 0_3 \\ 0_3 & 0_3 & \frac{1}{\mu_0 F} I_3 & 0_3 & 0_3 \\ I_3 & 0_3 & 0_3 & 0_3 & 0_3 \end{pmatrix}.$$

Here I_3 denotes a 3×3 identity matrix, and 0_3 denotes a 3×3 zero matrix.

To prove the stability, left-multiplying (3.191) by $\mathbf{u}'$, then integrating over Ω, and using the property $\mathbf{u}'\mathscr{B}\mathbf{u} = 0$, we obtain

$$\frac{d}{dt}(\mathbf{u}'\mathscr{A}\mathbf{u}) = \mathbf{u}'\mathscr{C}\mathbf{u} \le 0,$$

integrating which with respect to t leads to the stability (3.190). □

Remark 3.4. The stability (3.190) can be proved directly as we did previously for the Drude model and Lorentz model. Multiplying (3.183)–(3.187) by $\mathbf{E}, \mathbf{H}, \mathbf{K}, \mathbf{M}, \mathbf{J}$ and integrating over Ω, then adding the resultants together, we obtain

$$\frac{1}{2}\frac{d}{dt}[\epsilon_0 ||\mathbf{E}(t)||_0^2 + \mu_0 ||\mathbf{H}(t)||_0^2 + \frac{1}{\mu_0 \omega_0^2 F} ||\mathbf{K}(t)||_0^2 + \frac{1}{\mu_0 F} ||\mathbf{M}(t)||_0^2$$

$$+ \frac{1}{\epsilon_0 \omega_p^2} ||\mathbf{J}(t)||_0^2]$$

$$+ \frac{\nu}{\epsilon_0 \omega_p^2} ||\mathbf{J}(t)||_0^2 + \frac{\gamma}{\mu_0 \omega_0^2 F} ||\mathbf{K}(t)||_0^2 = 0,$$

integrating which from 0 to t leads to (3.190). Rewriting the last three governing equations (3.185)–(3.187) in this way leads to this elegant proof.

Now, let us prove the existence for the model problem (3.183)–(3.187).

Theorem 3.15. *The problem (3.183)–(3.187) has a unique solution* $(\mathbf{E}, \mathbf{H})$ *in* $H(curl; \Omega) \oplus H(curl; \Omega)$.

Proof. Though the technique developed in Theorems 3.8 and 3.13 can still be used here, we apply a different technique developed in [122].

From ordinary differential equation theory, it is easy to see that the solutions of (1.29) and (1.30) with zero initial conditions can be expressed as

$$\mathbf{P}(\mathbf{x},t) = \frac{\epsilon_0\omega_p^2}{\nu}\int_0^t (1-e^{-\nu(t-s)})\mathbf{E}(\mathbf{x},s)ds, \tag{3.192}$$

and

$$\mathbf{M}(\mathbf{x},t) = \mu_0 F\omega_0^2\int_0^t g(t-s)\mathbf{H}(\mathbf{x},s)ds, \tag{3.193}$$

respectively. Here the kernel $g(t) = \frac{1}{\alpha}e^{-\frac{\gamma}{2}t}\sin(\alpha t)$, where $\alpha = \sqrt{\omega_0^2 - (\frac{\gamma}{2})^2}$.

Using the definition $\mathbf{J} = \frac{\partial \mathbf{P}}{\partial t}$ and $\mathbf{K} = \frac{\partial \mathbf{M}}{\partial t}$ introduced in Chap. 1, then substituting (3.192) and (3.193) into (3.183) and (3.184), respectively, we can rewrite (3.183) and (3.184) as:

$$\frac{d}{dt}(A\mathscr{E} + K * \mathscr{E}) = L\mathscr{E} + \mathscr{F}, \tag{3.194}$$

where we denote $\mathscr{E} = (\mathbf{E},\mathbf{H})'$, $*$ for the convolution product, $\mathscr{F}$ for source terms obtained by transforming a problem with non-zero initial conditions to a problem with zero initial conditions. Moreover

$$A = \begin{pmatrix} \epsilon_0 I_3 & 0_3 \\ 0_3 & \mu_0 I_3 \end{pmatrix}, \quad K = \begin{pmatrix} \epsilon_1 I_3 & 0_3 \\ 0_3 & \mu_1 I_3 \end{pmatrix}, \quad L = \begin{pmatrix} 0_3 & \nabla\times \\ -\nabla\times & 0_3 \end{pmatrix},$$

where $\epsilon_1 = \frac{\epsilon_0\omega_p^2}{\nu}(u(t) - e^{-\nu t})$, $\mu_1 = \mu_0 F\omega_0^2 g(t)$, and $u(t)$ denotes the unit step function.

Note that problem (3.194) is a special case of Problem I of [122], whose existence and uniqueness is guaranteed by Theorem 3.1 of [122]. □

3.7.2 Two Numerical Schemes

In this section, we present two fully-discrete schemes developed in [195] for solving the problem (3.183)–(3.187).

First, let us start with a Crank-Nicolson type scheme: For $k = 1, 2, \cdots, N$, find $\mathbf{E}_h^k \in \mathbf{V}_h^0, \mathbf{J}_h^k \in \mathbf{V}_h, \mathbf{H}_h^k, \mathbf{K}_h^k, \mathbf{M}_h^k \in \mathbf{U}_h$ such that

$$\epsilon_0(\delta_\tau \mathbf{E}_h^k, \boldsymbol{\phi}_h) - (\overline{\mathbf{H}}_h^k, \nabla\times\boldsymbol{\phi}_h) + (\overline{\mathbf{J}}_h^k, \boldsymbol{\phi}_h) = 0, \tag{3.195}$$

$$\mu_0(\delta_\tau \mathbf{H}_h^k, \boldsymbol{\psi}_h) + (\nabla\times\overline{\mathbf{E}}_h^k, \boldsymbol{\psi}_h) + (\overline{\mathbf{K}}_h^k, \boldsymbol{\psi}_h) = 0, \tag{3.196}$$

$$\frac{1}{\mu_0\omega_0^2 F}(\delta_\tau \mathbf{K}_h^k, \tilde{\boldsymbol{\psi}}_{1h}) + \frac{\gamma}{\mu_0\omega_0^2 F}(\overline{\mathbf{K}}_h^k, \tilde{\boldsymbol{\psi}}_{1h}) + \frac{1}{\mu_0 F}(\overline{\mathbf{M}}_h^k, \tilde{\boldsymbol{\psi}}_{1h}) = (\overline{\mathbf{H}}_h^k, \tilde{\boldsymbol{\psi}}_{1h}),$$

$$\frac{1}{\mu_0 F}(\delta_\tau \mathbf{M}_h^k, \tilde{\boldsymbol{\psi}}_{2h}) = \frac{1}{\mu_0 F}(\overline{\mathbf{K}}_h^k, \tilde{\boldsymbol{\psi}}_{2h}), \tag{3.197}$$

$$\frac{1}{\epsilon_0\omega_p^2}(\delta_\tau \mathbf{J}_h^k, \tilde{\boldsymbol{\phi}}_h) + \frac{\nu}{\epsilon_0\omega_p^2}(\overline{\mathbf{J}}_h^k, \tilde{\boldsymbol{\phi}}_h) = (\overline{\mathbf{E}}_h^k, \tilde{\boldsymbol{\phi}}_h), \tag{3.198}$$

hold true for any $\boldsymbol{\phi}_h \in \mathbf{V}_h^0, \boldsymbol{\psi}_h, \tilde{\boldsymbol{\psi}}_{1h}, \tilde{\boldsymbol{\psi}}_{2h} \in \mathbf{U}_h, \tilde{\boldsymbol{\phi}}_h \in \mathbf{V}_h$, and are subject to the initial approximations

$$\mathbf{E}_h^0(\mathbf{x}) = \Pi_h \mathbf{E}_0(\mathbf{x}), \quad \mathbf{H}_h^0(\mathbf{x}) = P_h \mathbf{H}_0(\mathbf{x}),$$
$$\mathbf{K}_h^0(\mathbf{x}) = P_h \mathbf{K}_0(\mathbf{x}), \ \mathbf{M}_h^0(\mathbf{x}) = P_h \mathbf{M}_0(\mathbf{x}), \ \mathbf{J}_h^0(\mathbf{x}) = \Pi_h \mathbf{J}_0(\mathbf{x}).$$

Below we show that the scheme (3.195)–(3.198) satisfies a discrete stability, which has exactly the same form as the continuous case proved in Lemma 3.27.

Lemma 3.28. *For any $k \geq 1$, we have*

$$\epsilon_0 ||\mathbf{E}_h^k||_0^2 + \mu_0 ||\mathbf{H}_h^k||_0^2 + \frac{1}{\epsilon_0 \omega_p^2} ||\mathbf{J}_h^k||_0^2 + \frac{1}{\mu_0 \omega_0^2 F} ||\mathbf{K}_h^k||_0^2 + \frac{1}{\mu_0 F} ||\mathbf{M}_h^k||_0^2$$
$$\leq \epsilon_0 ||\mathbf{E}_h^0||_0^2 + \mu_0 ||\mathbf{H}_h^0||_0^2 + \frac{1}{\epsilon_0 \omega_p^2} ||\mathbf{J}_h^0||_0^2 + \frac{1}{\mu_0 \omega_0^2 F} ||\mathbf{K}_h^0||_0^2 + \frac{1}{\mu_0 F} ||\mathbf{M}_h^0||_0^2.$$

Proof. Choosing $\boldsymbol{\phi}_h = \tau \overline{\mathbf{E}}_h^k, \boldsymbol{\psi}_h = \tau \overline{\mathbf{H}}_h^k, \tilde{\boldsymbol{\psi}}_{1h} = \tau \overline{\mathbf{K}}_h^k, \tilde{\boldsymbol{\psi}}_{2h} = \tau \overline{\mathbf{M}}_h^k, \tilde{\boldsymbol{\phi}}_h = \tau \overline{\mathbf{J}}_h^k$ in (3.195)–(3.198), respectively, then adding the resultants together, we have

$$\frac{\epsilon_0}{2}(||\mathbf{E}_h^k||_0^2 - ||\mathbf{E}_h^{k-1}||_0^2) + \frac{\mu_0}{2}(||\mathbf{H}_h^k||_0^2 - ||\mathbf{H}_h^{k-1}||_0^2) + \frac{1}{2\epsilon_0 \omega_p^2}(||\mathbf{J}_h^k||_0^2 - ||\mathbf{J}_h^{k-1}||_0^2)$$
$$+ \frac{\tau \nu}{\epsilon_0 \omega_p^2} ||\overline{\mathbf{J}}_h^k||_0^2 + \frac{1}{2\mu_0 \omega_0^2 F}(||\mathbf{K}_h^k||_0^2 - ||\mathbf{K}_h^{k-1}||_0^2)$$
$$+ \frac{\tau \gamma}{\mu_0 \omega_0^2 F} ||\overline{\mathbf{K}}_h^k||_0^2 + \frac{1}{2\mu_0 F}(||\mathbf{M}_h^k||_0^2 - ||\mathbf{M}_h^{k-1}||_0^2) = 0,$$

which easily concludes the proof. □

For the Crank-Nicolson scheme (3.195)–(3.198), the following optimal error estimate can be proved similarly to Theorem 3.10. Details can be found in the original paper [195].

Theorem 3.16. *Let $(\mathbf{E}^m, \mathbf{H}^m, \mathbf{K}^m, \mathbf{M}^m, \mathbf{J}^m)$ and $(\mathbf{E}_h^m, \mathbf{H}_h^m, \mathbf{K}_h^m, \mathbf{M}_h^m, \mathbf{J}_h^m)$ be the analytic and numerical solutions of (3.183)–(3.187) and (3.195)–(3.198) at time t_m, respectively. Under proper regularity assumptions, there exists a constant $C > 0$ independent of mesh size h and time step τ, such that*

$$\sqrt{\epsilon_0} ||\mathbf{E}^n - \mathbf{E}_h^n||_0 + \sqrt{\mu_0} ||\mathbf{H}^n - \mathbf{H}_h^n||_0 + \frac{1}{\sqrt{\epsilon_0 \omega_p^2}} ||\mathbf{J}^n - \mathbf{J}_h^n||_0$$
$$+ \frac{1}{\sqrt{\mu_0 \omega_0^2 F}} ||\mathbf{K}^n - \mathbf{K}_h^n||_0 + \frac{1}{\sqrt{\mu_0 F}} ||\mathbf{M}^n - \mathbf{M}_h^n||_0 \leq C(h^l + \tau^2),$$

where $l \geq 1$ is the order of the basis functions in spaces $\mathbf{U}_h$ and $\mathbf{V}_h$.

Note that the Crank-Nicolson scheme (3.195)–(3.198) has a non-symmetric linear system of as many as 15 unknown functions (five 3-D unknown variables), which results a very large-scale system even for linear edge elements. Hence directly solving the coupled system is quite challenging. In this aspect the leap-frog scheme developed below (cf. (3.199)–(3.202)) is more practical and appealing, since one unknown variable is solved at each step. Of course, the leap-frog scheme has to obey the CFL time step constraint.

A leap-frog type scheme for solving (3.183)–(3.187) is proposed in [195]: For $k = 1, 2, \cdots$, find $\mathbf{E}_h^k \in \mathbf{V}_h^0, \mathbf{J}_h^{k+\frac{1}{2}} \in \mathbf{V}_h, \mathbf{H}_h^{k+\frac{1}{2}}, \mathbf{K}_h^k, \mathbf{M}_h^{k+\frac{1}{2}} \in \mathbf{U}_h$ such that

$$\epsilon_0(\frac{\mathbf{E}_h^k - \mathbf{E}_h^{k-1}}{\tau}, \boldsymbol{\phi}_h) - (\mathbf{H}_h^{k-\frac{1}{2}}, \nabla \times \boldsymbol{\phi}_h) + (\mathbf{J}_h^{k-\frac{1}{2}}, \boldsymbol{\phi}_h) = 0, \tag{3.199}$$

$$\mu_0(\frac{\mathbf{H}_h^{k+\frac{1}{2}} - \mathbf{H}_h^{k-\frac{1}{2}}}{\tau}, \boldsymbol{\psi}_h) + (\nabla \times \mathbf{E}_h^k, \boldsymbol{\psi}_h) + (\mathbf{K}_h^k, \boldsymbol{\psi}_h) = 0, \tag{3.200}$$

$$\begin{aligned}&\frac{1}{\mu_0\omega_0^2 F}(\frac{\mathbf{K}_h^k - \mathbf{K}_h^{k-1}}{\tau}, \tilde{\boldsymbol{\psi}}_{1h}) + \frac{\gamma}{\mu_0\omega_0^2 F}(\frac{\mathbf{K}_h^k + \mathbf{K}_h^{k-1}}{2}, \tilde{\boldsymbol{\psi}}_{1h}) + \frac{1}{\mu_0 F}(\mathbf{M}_h^{k-\frac{1}{2}}, \tilde{\boldsymbol{\psi}}_{1h})\\ &= (\mathbf{H}_h^{k-\frac{1}{2}}, \tilde{\boldsymbol{\psi}}_{1h}),\end{aligned}$$

$$\frac{1}{\mu_0 F}(\frac{\mathbf{M}_h^{k+\frac{1}{2}} - \mathbf{M}_h^{k-\frac{1}{2}}}{\tau}, \tilde{\boldsymbol{\psi}}_{2h}) = \frac{1}{\mu_0 F}(\mathbf{K}_h^k, \tilde{\boldsymbol{\psi}}_{2h}), \tag{3.201}$$

$$\frac{1}{\epsilon_0\omega_p^2}(\frac{\mathbf{J}_h^{k+\frac{1}{2}} - \mathbf{J}_h^{k-\frac{1}{2}}}{\tau}, \tilde{\boldsymbol{\phi}}_h) + \frac{\nu}{\epsilon_0\omega_p^2}(\frac{\mathbf{J}_h^{k+\frac{1}{2}} + \mathbf{J}_h^{k-\frac{1}{2}}}{2}, \tilde{\boldsymbol{\phi}}_h) = (\mathbf{E}_h^k, \tilde{\boldsymbol{\phi}}_h), \tag{3.202}$$

hold true for any $\boldsymbol{\phi}_h \in \mathbf{V}_h^0, \boldsymbol{\psi}_h, \tilde{\boldsymbol{\psi}}_{1h}, \tilde{\boldsymbol{\psi}}_{2h} \in \mathbf{U}_h, \tilde{\boldsymbol{\phi}}_h \in \mathbf{V}_h$, and are subject to the initial approximations

$$\mathbf{E}_h^0(\mathbf{x}) = \Pi_h \mathbf{E}_0(\mathbf{x}), \quad \mathbf{K}_h^0(\mathbf{x}) = P_h \mathbf{K}_0(\mathbf{x}), \tag{3.203}$$

$$\begin{aligned}\mathbf{H}_h^{\frac{1}{2}}(\mathbf{x}) &= P_h[\mathbf{H}_0(\mathbf{x}) - \frac{\tau}{2}\mu_0^{-1}(\nabla \times \mathbf{E}_0(\mathbf{x}) + \mathbf{K}_0(\mathbf{x}))],\\ \mathbf{M}_h^{\frac{1}{2}}(\mathbf{x}) &= P_h[\mathbf{M}_0(\mathbf{x}) + \frac{\tau}{2}\mathbf{K}_0(\mathbf{x})],\\ \mathbf{J}_h^{\frac{1}{2}}(\mathbf{x}) &= \Pi_h[\mathbf{J}_0(\mathbf{x}) + \frac{\tau}{2}(\epsilon_0\omega_p^2\mathbf{E}_0(\mathbf{x}) - \nu\mathbf{J}_0(\mathbf{x}))].\end{aligned} \tag{3.204}$$

The following discrete stability for the leap-frog scheme (3.199)–(3.202) can be proved similarly to Theorem 3.11.

Theorem 3.17. *Under the time step constraint*

$$\tau = \min\{\frac{1}{2\omega_0\sqrt{F}}, \frac{1}{2\omega_0}, \frac{1}{2\omega_p}, \frac{h}{2C_v C_{inv}}\}, \tag{3.205}$$

where C_v and C_{inv} are defined in Theorem 3.11. Then for any $k \geq 1$, we have

$$\epsilon_0||\mathbf{E}_h^k||_0^2 + \mu_0||\mathbf{H}_h^{k+\frac{1}{2}}||_0^2 + \frac{1}{\epsilon_0\omega_p^2}||\mathbf{J}_h^{k+\frac{1}{2}}||_0^2 + \frac{1}{\mu_0\omega_0^2 F}||\mathbf{K}_h^k||_0^2 + \frac{1}{\mu_0 F}||\mathbf{M}_h^{k+\frac{1}{2}}||_0^2$$
$$\leq C[\epsilon_0||\mathbf{E}_h^0||_0^2 + \mu_0||\mathbf{H}_h^{\frac{1}{2}}||_0^2 + \frac{1}{\epsilon_0\omega_p^2}||\mathbf{J}_h^{\frac{1}{2}}||_0^2 + \frac{1}{\mu_0\omega_0^2 F}||\mathbf{K}_h^0||_0^2 + \frac{1}{\mu_0 F}||\mathbf{M}_h^{\frac{1}{2}}||_0^2],$$

where $C > 1$ is independent of h and τ.

Similarly to Theorem 3.12, the following optimal error estimate can be proved.

Theorem 3.18. *Let $(\mathbf{E}^m, \mathbf{H}^{m+\frac{1}{2}}, \mathbf{K}^m, \mathbf{M}^{m+\frac{1}{2}}, \mathbf{J}^{m+\frac{1}{2}})$ and $(\mathbf{E}_h^m, \mathbf{H}_h^{m+\frac{1}{2}}, \mathbf{K}_h^m, \mathbf{M}_h^{m+\frac{1}{2}}, \mathbf{J}_h^{m+\frac{1}{2}})$ be the analytic and numerical solutions of (3.183)–(3.187) and (3.199)–(3.202), respectively. Under proper regularity assumptions, there exists a constant $C > 0$ independent of h and τ such that*

$$\sqrt{\epsilon_0}||\mathbf{E}^n - \mathbf{E}_h^n||_0 + \sqrt{\mu_0}||\mathbf{H}^{n+\frac{1}{2}} - \mathbf{H}_h^{n+\frac{1}{2}}||_0 + \frac{1}{\sqrt{\epsilon_0\omega_p^2}}||\mathbf{J}^{n+\frac{1}{2}} - \mathbf{J}_h^{n+\frac{1}{2}}||_0$$
$$+ \frac{1}{\sqrt{\mu_0\omega_0^2 F}}||\mathbf{K}^n - \mathbf{K}_h^n||_0 + \frac{1}{\sqrt{\mu_0 F}}||\mathbf{M}^{n+\frac{1}{2}} - \mathbf{M}_h^{n+\frac{1}{2}}||_0 \leq C(h^l + \tau^2),$$

where $l \geq 1$ is the order of the basis functions in spaces $\mathbf{U}_h$ and $\mathbf{V}_h$.

3.8 Bibliographical Remarks

In this chapter we presented some basic time-domain finite element (FETD) methods developed for Maxwell's equations in metamaterials. Some early works on FETD can be found in papers [79, 178] and references cited therein. Since 2000, in addition to our own work on FETD (e.g., [189, 190, 193, 194]), there has been a growing interest in developing FETD methods for Maxwell's equations in dispersive media [25, 160, 207, 251, 258, 290]. A nice list of literature on FETD methods for general complex media (including metamaterials) can be found in the review paper by Teixeira [277], which provides over 300 papers (though many are on FDTD methods) published by 2007. Another very recent and excellent review on FETD was written by Chen and Monk [68], which provides some numerical analysis on use of edge elements and certain A-stable schemes.

For more advanced finite element theory for Maxwell's equations, interested readers should consult some more theoretical papers such as [11, 13, 40, 41, 75, 145, 222] and the classic book by Monk [217].

Chapter 4
Discontinuous Galerkin Methods for Metamaterials

In this chapter, we introduce several discontinuous Galerkin (DG) methods for solving time-dependent Maxwell's equations in dispersive media and metamaterials. We first present a succint review of DG methods in Sect. 4.1. Then we present some DG methods for the cold plasma model in Sect. 4.2. Here the DG methods are developed for a second-order integro-differential vector wave equation. We then consider DG methods for the Drude model written in a system of first-order differential equations in Sect. 4.3. Finally, we extend the nodal DG methods developed by Hesthaven and Warburton (Nodal discontinuous Galerkin methods: algorithms, analysis, and applications. Springer, New York, 2008) to metamaterial Maxwell's equations in Sect. 4.4.

4.1 A Brief Overview of DG Methods

The discontinuous Galerkin method was originally introduced in 1973 by Reed and Hill for solving a neutron transport equation. In recent years the DG method gained more popularity in solving various differential equations due to its great flexibility in mesh construction, easily handling complex geometries or interfaces, and efficiency in parallel implementation. A detailed overview on the evolution of the DG methods from 1973 to 1999 is provided by Cockburn et al. [83]. More details and references on DG methods can be found in books [83, 99, 141, 247] and references therein.

In the past decade, there has been considerable interest in developing DG methods for Maxwell's equations in the free space [70, 84, 100, 120, 133, 140, 147, 168, 219, 238]. However, the study of DG method for Maxwell's equations in dispersive media (including metamaterial, a lossy dispersive composite material) are very limited. In 2004, a time-domain DG method was investigated in [207] for solving the first-order Maxwell's equations in dispersive media, but no error analysis was carried out. In 2009, a priori error estimate [151] and a posteriori error estimation [182] of the interior penalty DG method were obtained for Maxwell's

J. Li and Y. Huang, *Time-Domain Finite Element Methods for Maxwell's Equations in Metamaterials*, Springer Series in Computational Mathematics 43,
DOI 10.1007/978-3-642-33789-5_4,

equations in dispersive media. However, the error estimate obtained in [151] was optimal in the energy norm, but sub-optimal in the L^2-norm. Later, the error estimates were improved to be optimal in the L^2-norm for both a semi-discrete DG scheme and a fully explicit DG scheme [186]. In [155], a fully implicit DG method was developed for solving dispersive media models. This scheme is proved to be unconditionally stable and has optimal error estimates in both L^2 norm and DG energy norm. Very recently, some DG methods have been developed for dispersive [251, 290] and metamaterial [185] Maxwell's equations written as a system of first-order differential equations.

4.2 Discontinuous Galerkin Methods for Cold Plasma

4.2.1 The Modeling Equations

It is known that in reality all electromagnetic media show some dispersion, i.e., some physical parameters such as permittivity (and/or permeability) depends on the wavelength. Such media are often called dispersive media. In most applications, we are interested in linear dispersive media, which satisfy the relation (cf. (1.11)):

$$\mathbf{D}(\mathbf{x}, t) = \epsilon(\omega)\mathbf{E}, \quad \mathbf{B}(\mathbf{x}, t) = \mu(\omega)\mathbf{E},$$

and are often encountered in nature. For example, rock, soil, ice, snow, and plasma are dispersive media. Hence, transient simulation of electromagnetic wave propagation and scattering in dispersive media is important for a wide range of applications involving biological media, optical materials, artificial dielectrics, or earth media, where the host medium is frequency dispersive.

Since early 1990s, many FDTD methods have been developed for modeling electromagnetic propagation in isotropic cold plasma. Early references can be found in Chap. 9 of [276]. It is known that for an isotropic nonmagnetized cold electron plasma, the complete governing equations are:

$$\epsilon_0 \frac{\partial \mathbf{E}}{\partial t} = \nabla \times \mathbf{H} - \tilde{\mathbf{J}} \tag{4.1}$$

$$\mu_0 \frac{\partial \mathbf{H}}{\partial t} = -\nabla \times \mathbf{E} \tag{4.2}$$

$$\frac{\partial \tilde{\mathbf{J}}}{\partial t} + \nu \tilde{\mathbf{J}} = \epsilon_0 \omega_p^2 \mathbf{E} \tag{4.3}$$

where $\mathbf{E}$ is the electric field, $\mathbf{H}$ is the magnetic field, ϵ_0 is the permittivity of free space, μ_0 is the permeability of free space, $\tilde{\mathbf{J}}$ is the polarization current density, ω_p is the plasma frequency, $\nu \geq 0$ is the electron-neutron collision frequency. Solving (4.3) with the assumption that the initial electron velocity is 0, we obtain

$$\tilde{\mathbf{J}}(\mathbf{E}) \equiv \tilde{\mathbf{J}}(\mathbf{x}, t; \mathbf{E})$$
$$= \epsilon_0 \omega_p^2 e^{-\nu t} \int_0^t e^{\nu s} \mathbf{E}(\mathbf{x}, s) ds = \epsilon_0 \omega_p^2 \int_0^t e^{-\nu(t-s)} \mathbf{E}(\mathbf{x}, s) ds. \quad (4.4)$$

Taking derivative of (4.1) with respect to t, and eliminating $\mathbf{H}$ and $\tilde{\mathbf{J}}$ by using (4.2)–(4.4), we can reduce the modeling equations to the following integro-differential equation

$$\mathbf{E}_{tt} + \nabla \times (C_v^2 \nabla \times \mathbf{E}) + \omega_p^2 \mathbf{E} - \mathbf{J}(\mathbf{E}) = 0 \quad \text{in } \Omega \times I, \quad (4.5)$$

where the rescaled polarization current density $\mathbf{J}$ is represented as

$$\mathbf{J}(\mathbf{E}) = \nu \omega_p^2 \int_0^t e^{-\nu(t-s)} \mathbf{E}(\mathbf{x}, s) ds. \quad (4.6)$$

Recall that $C_v = \frac{1}{\sqrt{\epsilon_0 \mu_0}}$ denotes the wave speed in free space. Here $I = (0, T)$ is a finite time interval and Ω is a bounded Lipschitz polyhedron in R^3.

To make the problem complete, we assume that the boundary of Ω is a perfect conductor so that

$$\mathbf{n} \times \mathbf{E} = \mathbf{0} \quad \text{on} \quad \partial\Omega \times I, \quad (4.7)$$

and the initial conditions for (4.5) are given as

$$\mathbf{E}(\mathbf{x}, 0) = \mathbf{E}_0(\mathbf{x}) \quad \text{and} \quad \mathbf{E}_t(\mathbf{x}, 0) = \mathbf{E}_1(\mathbf{x}), \quad (4.8)$$

where $\mathbf{E}_0(\mathbf{x})$ and $\mathbf{E}_1(\mathbf{x})$ are some given functions.

Lemma 4.1. *There exists a unique solution* $\mathbf{E} \in H_0(curl; \Omega)$ *for (4.5).*

Proof. Taking the Laplace transformation of (4.5) and denoting $\hat{\mathbf{E}}(s)$ as the Laplace transformation of $\mathbf{E}(t)$, we have

$$s^2 \hat{\mathbf{E}} - s\mathbf{E}(0) - \mathbf{E}_t(0) + \nabla \times (C_v^2 \nabla \times \hat{\mathbf{E}}) + \omega_p^2 \hat{\mathbf{E}} - \nu \omega_p^2 \frac{1}{s + \nu} \hat{\mathbf{E}} = 0,$$

which can be rewritten as

$$s(s^2 + \nu s + \omega_p^2)\hat{\mathbf{E}} + (s + \nu)\nabla \times (C_v^2 \nabla \times \hat{\mathbf{E}}) = s(s + \nu)\mathbf{E}(0) + (s + \nu)\mathbf{E}_t(0). \quad (4.9)$$

The weak formulation of (4.9) can be formulated as: Find $\hat{\mathbf{E}} \in H_0(\text{curl}; \Omega)$ such that

$$s(s^2 + \nu s + \omega_p^2)(\hat{\mathbf{E}}, \phi) + (s + \nu)(C_v^2 \nabla \times \hat{\mathbf{E}}, \nabla \times \phi)$$
$$= (s(s + \nu)\mathbf{E}(0) + (s + \nu)\mathbf{E}_t(0), \phi), \quad (4.10)$$

holds true for any $\phi \in H_0(\text{curl}; \Omega)$. The existence of a unique weak solution $\hat{\mathbf{E}}$ is guaranteed by the Lax-Milgram lemma. Taking the inverse Laplace transformation of $\hat{\mathbf{E}}$ leads to the solution $\mathbf{E}$ for (4.5). □

4.2.2 A Semi-discrete Scheme

We consider a shape-regular mesh T_h that partitions the domain Ω into disjoint tetrahedral elements $\{K\}$, such that $\overline{\Omega} = \bigcup_{K \in T_h} K$. Furthermore, we denote the set of all interior faces by F_h^I, the set of all boundary faces by F_h^B, and the set of all faces by $F_h = F_h^I \cup F_h^B$. We want to remark that the optimal L^2-norm error estimate is based on a duality argument and inverse estimate, hence we need to assume that the mesh be quasi-uniform and the domain Ω be convex.

We assume that the finite element space is given by

$$\mathbf{V}_h = \{\mathbf{v} \in L^2(\Omega)^3 : \ \mathbf{v}|_K \in (P_l(K))^3, \ K \in T_h\}, \ \ l \geq 1, \tag{4.11}$$

where $P_l(K)$ denotes the space of polynomials of total degree at most l on K.

A semi-discrete DG scheme can be formed for (4.5): For any $t \in (0, T)$, find $\mathbf{E}^h(\cdot, t) \in \mathbf{V}_h$ such that

$$(\mathbf{E}_{tt}^h, \phi) + a_h(\mathbf{E}^h, \phi) + \omega_p^2(\mathbf{E}^h, \phi) - (\mathbf{J}(\mathbf{E}^h), \phi) = 0, \ \ \forall\, \phi \in \mathbf{V}_h, \tag{4.12}$$

subject to the initial conditions

$$\mathbf{E}^h|_{t=0} = \Pi_2 \mathbf{E}_0, \quad \mathbf{E}_t^h|_{t=0} = \Pi_2 \mathbf{E}_1, \tag{4.13}$$

where Π_2 denotes the standard L_2-projection onto $\mathbf{V}_h$. Moreover, the bilinear form a_h is defined on $\mathbf{V}_h \times \mathbf{V}_h$ as

$$\begin{aligned} a_h(\mathbf{u}, \mathbf{v}) = & \sum_{K \in T_h} \int_K C_v^2 \nabla \times \mathbf{u} \cdot \nabla \times \mathbf{v} d\mathbf{x} - \sum_{f \in F_h} \int_f [[\mathbf{u}]]_T \cdot \{\{C_v^2 \nabla \times \mathbf{v}\}\} dA \\ & - \sum_{f \in F_h} \int_f [[\mathbf{v}]]_T \cdot \{\{C_v^2 \nabla \times \mathbf{u}\}\} dA + \sum_{f \in F_h} \int_f a [[\mathbf{u}]]_T \cdot [[\mathbf{v}]]_T dA. \end{aligned}$$

Here $[[\mathbf{v}]]_T$ and $\{\{\mathbf{v}\}\}$ are the standard notation for the tangential jumps and averages of $\mathbf{v}$ across an interior face $f = \partial K^+ \cap \partial K^-$ between two neighboring elements K^+ and K^-:

$$[[\mathbf{v}]]_T = \mathbf{n}^+ \times \mathbf{v}^+ + \mathbf{n}^- \times \mathbf{v}^-, \quad \{\{\mathbf{v}\}\} = (\mathbf{v}^+ + \mathbf{v}^-)/2, \tag{4.14}$$

where $\mathbf{v}^{\pm}$ denote the traces of $\mathbf{v}$ from within $K^{\pm}$, and $\mathbf{n}^{\pm}$ denote the unit outward normal vectors on the boundaries $\partial K^{\pm}$, respectively. While on a boundary face $f = \partial K \cap \partial \Omega$, we define $[[\mathbf{v}]]_T = \mathbf{n} \times \mathbf{v}$ and $\{\{\mathbf{v}\}\} = \mathbf{v}$. Finally, a is a penalty function, which is defined on each face $f \in F_h$ as:

$$a|_f = \gamma c_v^2 \hbar^{-1},$$

where $\hbar|_f = \min\{h_{K^+}, h_{K^-}\}$ for an interior face $f = \partial K^+ \cap \partial K^-$, and $\hbar|_f = h_K$ for a boundary face $f = \partial K \cap \partial \Omega$. The penalty parameter γ is a positive constant and has to be chosen sufficiently large in order to guarantee the coercivity of $\tilde{a}_h(\cdot, \cdot)$ defined below.

Furthermore, we denote the space $\mathbf{V}(h) = H_0(\text{curl}; \Omega) + \mathbf{V}_h$ and define the semi-norm

$$|\mathbf{v}|_h^2 = \sum_{K \in F_h} ||C_v \nabla \times \mathbf{v}||_{0,K}^2 + \sum_{f \in F_h} ||a^{1/2}[[\mathbf{v}]]_T||_{0,f}^2,$$

and the DG energy norm by

$$||\mathbf{v}||_h^2 = ||\omega_p \mathbf{v}||_{0,\Omega}^2 + |\mathbf{v}|_h^2.$$

In order to carry out the error analysis, we introduce an auxiliary bilinear form $\tilde{a}_h$ on $\mathbf{V}(h) \times \mathbf{V}(h)$ defined as [134]

$$\begin{aligned}\tilde{a}_h(\mathbf{u}, \mathbf{v}) = &\sum_{K \in T_h} \int_K C_v^2 \nabla \times \mathbf{u} \cdot \nabla \times \mathbf{v} d\mathbf{x} - \sum_{f \in F_h} \int_f [[\mathbf{u}]]_T \cdot \{\{C_v^2 \Pi_2(\nabla \times \mathbf{v})\}\} dA \\ &- \sum_{f \in F_h} \int_f [[\mathbf{v}]]_T \cdot \{\{C_v^2 \Pi_2(\nabla \times \mathbf{u})\}\} dA + \sum_{f \in F_h} \int_f a[[\mathbf{u}]]_T \cdot [[\mathbf{v}]]_T dA.\end{aligned}$$

Note that $\tilde{a}_h$ equals a_h on $\mathbf{V}_h \times \mathbf{V}_h$ and is well defined on $H_0(\text{curl}; \Omega) \times H_0(\text{curl}; \Omega)$.

It is shown that $\tilde{a}_h$ is both continuous and coersive:

Lemma 4.2 ([133, Lemma 5]). *For γ larger than a positive constant γ_{min}, independent of the local mesh sizes, we have*

$$|\tilde{a}_h(\mathbf{u}, \mathbf{v})| \leq C_{cont} |\mathbf{u}|_h |\mathbf{v}|_h, \quad \tilde{a}_h(\mathbf{v}, \mathbf{v}) \geq C_{coer} |\mathbf{v}|_h^2, \quad \mathbf{u}, \mathbf{v} \in \mathbf{V}(h),$$

where $C_{cont} = \sqrt{2}$ and $C_{coer} = \frac{1}{2}$.

For an element K and any $\mathbf{u} \in (P_l(K))^3$, we have the standard inverse estimate

$$||\nabla \times \mathbf{u}||_{0,K} \leq C h_K^{-1} ||\mathbf{u}||_{0,K},$$

and the trace estimate

$$||\mathbf{u}||_{0,\partial K} \leq C h_K^{-\frac{1}{2}} ||\mathbf{u}||_{0,K},$$

which, along with Lemma 4.2, yields the following lemma.

Lemma 4.3. *For a quasi-uniform mesh T_h, there holds*

$$|\tilde{a}_h(\mathbf{u},\mathbf{u})| \le C_b h^{-2}||\mathbf{u}||_0^2, \quad \mathbf{u} \in \mathbf{V}_h,$$

where the constant $C_b > 0$ depends on the quasi-uniformity constant of the mesh and polynomial degree l, but is independent of the mesh size h.

First, we can prove that (4.12) has a unique solution.

Lemma 4.4. *There exists a unique solution $\mathbf{E}^h$ for the discrete model (4.12).*

Proof. Choosing $\phi = \mathbf{E}_t^h$ in (4.12), we obtain

$$\frac{1}{2}\frac{d}{dt}(||\mathbf{E}_t^h||_0^2 + a_h(\mathbf{E}^h,\mathbf{E}^h) + \omega_p^2||\mathbf{E}^h||_0^2) - (\mathbf{J}(\mathbf{E}^h),\mathbf{E}_t^h) = 0,$$

integrating which, we have

$$\begin{aligned}&||\mathbf{E}_t^h(t)||_0^2 + a_h(\mathbf{E}^h(t),\mathbf{E}^h(t)) + \omega_p^2||\mathbf{E}^h(t)||_0^2\\ &= ||\mathbf{E}_t^h(0)||_0^2 + a_h(\mathbf{E}^h(0),\mathbf{E}^h(0)) + \omega_p^2||\mathbf{E}^h(0)||_0^2 + 2\int_0^t(\mathbf{J}(\mathbf{E}^h),\mathbf{E}_t^h)dt. \quad (4.15)\end{aligned}$$

Using the Cauchy-Schwarz inequality and the definition of $\mathbf{J}(\mathbf{E})$, we have

$$\begin{aligned}2\int_0^t(\mathbf{J}(\mathbf{E}^h),\mathbf{E}_t^h)dt &\le \int_0^t||\mathbf{J}(\mathbf{E}^h(t))||_0^2dt + \int_0^t||\mathbf{E}_t^h(t)||_0^2dt\\ &\le \int_0^t\frac{\nu\omega_p^4}{2}\int_0^t||\mathbf{E}^h(s)||_0^2dsdt + \int_0^t||\mathbf{E}_t^h(t)||_0^2dt\\ &\le \frac{\nu\omega_p^4 t}{2}\int_0^t||\mathbf{E}^h(t)||_0^2dt + \int_0^t||\mathbf{E}_t^h(t)||_0^2dt.\end{aligned}$$

Substituting the above estimate into (4.15), then using Lemma 4.2 and the discrete Gronwall inequality, we have the following stability

$$||\mathbf{E}_t^h(t)||_0^2 + |\mathbf{E}^h(t)|_h^2 + ||\mathbf{E}^h(t)||_0^2 \le ||\mathbf{E}_t^h(0)||_0^2 + |\mathbf{E}^h(0)|_h^2 + ||\mathbf{E}^h(0)||_0^2,$$

which implies the uniqueness of solution for (4.12). Since (4.12) is a finite dimensional linear system, the uniqueness of solution gives the existence immediately. □

The following optimal error estimate for (4.12) is proved in [186].

Theorem 4.1. *Let $\mathbf{E}$ and $\mathbf{E}^h$ be the solutions of (4.5) and (4.12), respectively. Then under the following regularity assumptions*

$$\mathbf{E},\mathbf{E}_t \in L^\infty(0,T;(H^{\alpha+\sigma_E}(\Omega))^3),\ \ \nabla\times\mathbf{E},\nabla\times\mathbf{E}_t \in L^\infty(0,T;(H^\alpha(\Omega))^3),\forall\,\alpha > \frac{1}{2},$$

there holds

$$||\mathbf{E} - \mathbf{E}^h||_{L^\infty(0,T;(L^2(\Omega))^3)} \le C h^{\min(\alpha,l)+\sigma_E},$$

where $l \ge 1$ *is the degree of the polynomial function in the finite element space (4.11),* $\sigma_E \in (\frac{1}{2}, 1]$ *is related to the regularity of the Laplacian in polyhedra (*$\sigma_E = 1$ *when* Ω *is convex), and the constant* $C > 0$ *is independent of* h.

Note that when $\mathbf{E}$ is smooth enough on a convex domain, Theorem 4.1 gives the optimal error estimate in the L^2-norm:

$$||\mathbf{E} - \mathbf{E}^h||_{L^\infty(0,T;(L^2(\Omega))^3)} \le C h^{l+1}.$$

4.2.3 A Fully Explicit Scheme

To define a fully discrete scheme, we divide the time interval $(0, T)$ into N uniform subintervals by points $0 = t_0 < t_1 < \cdots < t_N = T$, where $t_k = k\tau$.

A fully explicit scheme can be formulated for (4.5): For any $1 \le n \le N-1$, find $\mathbf{E}_h^{n+1} \in \mathbf{V}_h$ such that

$$(\delta_\tau^2 \mathbf{E}_h^n, \mathbf{v}) + a_h(\mathbf{E}_h^n, \mathbf{v}) + \omega_p^2(\mathbf{E}_h^n, \mathbf{v}) - (\mathbf{J}_h^n, \mathbf{v}) = 0, \quad \forall \mathbf{v} \in \mathbf{V}_h, \tag{4.16}$$

subject to the initial approximation

$$\mathbf{E}_h^0 = \Pi_2 \mathbf{E}_0, \quad \mathbf{E}_h^1 = \Pi_2(\mathbf{E}_0 + \tau \mathbf{E}_1 + \frac{\tau^2}{2}\mathbf{E}_{tt}(0)), \tag{4.17}$$

where $\delta_\tau^2 \mathbf{E}_h^n = (\mathbf{E}_h^{n+1} - 2\mathbf{E}_h^n + \mathbf{E}_h^{n-1})/\tau^2$. Furthermore, $\mathbf{E}_{tt}(0) = -[\nabla \times (C_v^2 \nabla \times \mathbf{E}_0) + \omega_p^2 \mathbf{E}_0]$ is obtained by setting $t = 0$ in the governing equation (4.5), and $\mathbf{J}_h^n$ is obtained from the following recursive formula

$$\mathbf{J}_h^0 = 0, \quad \mathbf{J}_h^n = e^{-\nu\tau}\mathbf{J}_h^{n-1} + \frac{\nu\omega_p^2}{2}\tau(e^{-\nu\tau}\mathbf{E}_h^{n-1} + \mathbf{E}_h^n), \quad n \ge 1. \tag{4.18}$$

Theorem 4.2. *Let* $\mathbf{E}$ *and* $\mathbf{E}_h^n$ *be the solutions of the problem (4.5) and the finite element scheme (4.16)–(4.18) at time* t *and* t_n*, respectively. Under the CFL condition*

$$\tau < \frac{2h}{\sqrt{C_b + \omega_p^2 h^2}}, \tag{4.19}$$

where C_b *is the constant of Lemma 4.3. Furthermore, we assume that*

$$\mathbf{E}, \mathbf{E}_t \in L^\infty(0,T;(H^{\alpha+\sigma_E}(\Omega))^3), \quad \nabla \times \mathbf{E}, \nabla \times \mathbf{E}_t \in L^\infty(0,T;(H^\alpha(\Omega))^3),$$
$$\mathbf{E}_t, \mathbf{E}_{t^2}, \mathbf{E}_{t^3} \in L^\infty(0,T;(L^2(\Omega))^3), \quad \mathbf{E}_{t^4} \in L^2(0,T;(L^2(\Omega))^3).$$

Then there is a constant $C > 0$, independent of both the time step τ and mesh size h, such that

$$\max_{1\le n\le N} ||\mathbf{E}_h^n - \mathbf{E}^n||_0 \le C(\tau^2 + h^{\min(\alpha,l)+\sigma_E}), \quad l \ge 1,$$

where σ_E has the same meaning as in Theorem 4.1.

Interested readers can find the detailed proof in the original paper [186]. For smooth solutions on convex domain, we have the optimal L^2 error estimate:

$$\max_{1\le n\le N} ||\mathbf{E}_h^n - \mathbf{E}^n||_0 \le C(\tau^2 + h^{l+1}), \quad l \ge 1.$$

4.2.4 A Fully Implicit Scheme

An implicit scheme for (4.5) can be constructed as follows: For any $k \ge 1$, find $\mathbf{E}_h^{k+1} \in \mathbf{V}_h$ such that

$$(\delta_\tau^2 \mathbf{E}_h^k, v) + a_h(\overline{\mathbf{E}}_h^k v) + \omega_p^2(\overline{\mathbf{E}}_h^k, v) - (\mathbf{J}_h^k, v) = 0, \qquad \forall\, v \in \mathbf{V}_h, \tag{4.20}$$

subject to the same initial approximation $\mathbf{E}_h^0$ and $\mathbf{E}_h^1$ as (4.17), and the same recursive definition $\mathbf{J}_h^k$ as (4.18). Here we use the averaging operator $\overline{\mathbf{E}}_h^k = (\mathbf{E}_h^{k+1} + \mathbf{E}_h^{k-1})/2$.

Lemma 4.5. *For the $\mathbf{J}_h^k$ defined in (4.18), we have*

$$|\mathbf{J}_h^k| \le C\tau \sum_{j=0}^{k} |\mathbf{E}_h^j|, \quad \forall\, k \ge 1,$$

and

$$||\mathbf{J}_h^k||_0^2 \le CT\tau \sum_{j=0}^{k} ||\mathbf{E}_h^j||_0^2, \quad \forall\, k \ge 1.$$

Proof. Denote $a = e^{-\nu\tau}, b = \frac{\nu\omega_p^2}{2}\tau$. Then we can rewrite (4.18) as:

$$\mathbf{J}_h^k = a\mathbf{J}_h^{k-1} + ab\mathbf{E}_h^{k-1} + b\mathbf{E}_h^k,$$

from which we obtain

$$\begin{aligned}\mathbf{J}_h^k &= a(a\mathbf{J}_h^{k-2} + ab\mathbf{E}_h^{k-2} + b\mathbf{E}_h^{k-1}) + ab\mathbf{E}_h^{k-1} + b\mathbf{E}_h^k \\ &= a^2\mathbf{J}_h^{k-2} + a^2b\mathbf{E}_h^{k-2} + 2ab\mathbf{E}_h^{k-1} + b\mathbf{E}_h^k \\ &= a^2(a\mathbf{J}_h^{k-3} + ab\mathbf{E}_h^{k-3} + b\mathbf{E}_h^{k-2}) + a^2b\mathbf{E}_h^{k-2} + 2ab\mathbf{E}_h^{k-1} + b\mathbf{E}_h^k\end{aligned}$$

$$= a^3\mathbf{J}_h^{k-3} + a^3 b\mathbf{E}_h^{k-3} + 2a^2 b\mathbf{E}_h^{k-2} + 2ab\mathbf{E}_h^{k-1} + b\mathbf{E}_h^k$$
$$= \cdots$$
$$= a^k\mathbf{J}_h^0 + a^k b\mathbf{E}_h^0 + 2a^{k-1}b\mathbf{E}_h^1 + \cdots + 2a^2 b\mathbf{E}_h^{k-2} + 2ab\mathbf{E}_h^{k-1} + b\mathbf{E}_h^k. \quad (4.21)$$

Using $\mathbf{J}_h^0 = 0$ and the definitions of a and b in (4.21), we have

$$|\mathbf{J}_h^k| \le C\tau \sum_{j=0}^{k} |\mathbf{E}_h^j|,$$

which leads to

$$||\mathbf{J}_h^k||_0^2 \le C\tau^2 (\sum_{j=0}^{k} 1^2)(\sum_{j=0}^{k} ||\mathbf{E}_h^j||_0^2) \le CT\tau \sum_{j=0}^{k} ||\mathbf{E}_h^j||_0^2,$$

which concludes the proof. □

Using Lemma 4.5, we can prove the following unconditional stability for the scheme (4.20).

Theorem 4.3. *Denote the backward difference* $\partial_\tau u^k = (u^k - u^{k-1})/\tau$. *Then for the solution of scheme (4.20), we have*

$$||\partial_\tau \mathbf{E}_h^n||_0^2 + ||\mathbf{E}_h^n||_h^2 \le C(||\mathbf{E}_h^1||_h^2 + ||\mathbf{E}_h^0||_h^2 + ||\partial_\tau \mathbf{E}_h^1||_0^2), \quad \forall n \ge 2.$$

Proof. Choosing $v = \mathbf{E}_h^{k+1} - \mathbf{E}_h^{k-1} = \tau(\partial_\tau \mathbf{E}_h^{k+1} + \partial_\tau \mathbf{E}_h^k)$ in (4.20), we obtain

$$||\partial_\tau \mathbf{E}_h^{k+1}||_0^2 - ||\partial_\tau \mathbf{E}_h^k||_0^2 + \frac{1}{2}(\widetilde{a}_h(\mathbf{E}_h^{k+1}, \mathbf{E}_h^{k+1}) - \widetilde{a}_h(\mathbf{E}_h^{k-1}, \mathbf{E}_h^{k-1}))$$
$$+ \frac{\omega_p^2}{2}(||\mathbf{E}_h^{k+1}||_0^2 - ||\mathbf{E}_h^{k-1}||_0^2) = \tau(\mathbf{J}_h^k, \partial_\tau \mathbf{E}_h^{k+1} + \partial_\tau \mathbf{E}_h^k). \quad (4.22)$$

Summing up (4.22) from $k = 1$ to $k = n-1$ $(2 \le n \le M)$, we have

$$||\partial_\tau \mathbf{E}_h^n||_0^2 - ||\partial_\tau \mathbf{E}_h^1||_0^2 + \frac{1}{2}(\widetilde{a}_h(\mathbf{E}_h^n, \mathbf{E}_h^n) + \widetilde{a}_h(\mathbf{E}_h^{n-1}, \mathbf{E}_h^{n-1}) - \widetilde{a}_h(\mathbf{E}_h^1, \mathbf{E}_h^1) - \widetilde{a}_h(\mathbf{E}_h^0, \mathbf{E}_h^0))$$
$$+ \frac{\omega_p^2}{2}(||\mathbf{E}_h^n||_0^2 + ||\mathbf{E}_h^{n-1}||_0^2 - ||\mathbf{E}_h^1||_0^2 - ||\mathbf{E}_h^0||_0^2)$$
$$= \tau \sum_{k=1}^{n-1} (\mathbf{J}_h^k, \partial_\tau \mathbf{E}_h^{k+1} + \partial_\tau \mathbf{E}_h^k). \quad (4.23)$$

By Lemma 4.5, we obtain

$$\begin{aligned}
&\tau \sum_{k=1}^{n-1} (\mathbf{J}_h^k, \partial_\tau \mathbf{E}_h^{k+1} + \partial_\tau \mathbf{E}_h^k) \\
&\leq \tau \sum_{k=1}^{n-1} \left(\frac{1}{2\delta_1} ||\mathbf{J}_h^k||_0^2 + \frac{\delta_1}{2} ||\partial_\tau \mathbf{E}_h^{k+1} + \partial_\tau \mathbf{E}_h^k||_0^2\right) \\
&\leq \frac{\tau}{2\delta_1} \sum_{k=1}^{n-1} \left(CT\tau \sum_{j=0}^{k} ||\mathbf{E}_h^j||_0^2\right) + \delta_1 \tau \sum_{k=1}^{n-1} \left(||\partial_\tau \mathbf{E}_h^{k+1}||_0^2 + ||\partial_\tau \mathbf{E}_h^k||_0^2\right) \\
&\leq CT^2 \tau \sum_{k=0}^{n-1} ||\mathbf{E}_h^k||_0^2 + 2\delta_1 \tau \sum_{k=1}^{n-1} ||\partial_\tau \mathbf{E}_h^k||_0^2 + \delta_1 \tau ||\partial_\tau \mathbf{E}_h^n||_0^2. \qquad (4.24)
\end{aligned}$$

Substituting (4.24) into (4.23), choosing δ_1 small enough, then using the discrete Gronwall inequality, we obtain

$$||\partial_\tau \mathbf{E}_h^n||_0^2 + ||\mathbf{E}_h^n||_h^2 \leq C(||\mathbf{E}_h^1||_h^2 + ||\mathbf{E}_h^0||_h^2 + ||\partial_\tau \mathbf{E}_h^1||_0^2),$$

which concludes the proof. □

The following optimal error estimate is proved in [155].

Theorem 4.4. *Let* $\mathbf{E}$ *and* $\mathbf{E}_h^k$ *be the solutions of the problem (4.5) and the finite element scheme (4.20) at the time* t *and* t_k*, respectively. Under the regularity assumptions:*

$$\begin{aligned}
&\mathbf{E}, \nabla \times \mathbf{E} \in L^\infty(0, T; (H^{\alpha+\sigma_E}(\Omega))^3), \quad \nabla \times \mathbf{E}_{tt} \in L^2(0, T; (H^\alpha(\Omega))^3), \\
&\mathbf{E}_t, \mathbf{E}_{tt}, \nabla \times \nabla \times \mathbf{E}_{tt} \in L^2(0, T; (L^2(\Omega))^3), \\
&\mathbf{E}_{t^3} \in L^\infty(0, \tau; H(curl; \Omega)), \ \nabla \times \mathbf{E}_{t^3} \in L^\infty(0, \tau; (H^\alpha(\Omega))^3),
\end{aligned}$$

there is a constant $C > 0$*, independent of time step* τ *and mesh size* h*, such that*

$$\max_{1 \leq n \leq M} ||\mathbf{E}_h^n - \mathbf{E}^n||_0 \leq C(\tau^2 + h^{\min(\alpha, l)+\sigma_E}),$$

and

$$\max_{1 \leq n \leq M} ||\mathbf{E}_h^n - \mathbf{E}^n||_h \leq C(\tau^2 + h^{\min(\alpha, l)}),$$

where $l \geq 1$ *is the degree of the polynomial function in the finite element space (4.11). Hence, on a convex domain* Ω*, if the solution* $\mathbf{E}$ *has enough regularity, we have the optimal error estimates*

$$\max_{1 \leq n \leq M} ||\mathbf{E}_h^n - \mathbf{E}^n||_0 \leq C(\tau^2 + h^{l+1}), \quad \max_{1 \leq n \leq M} |\mathbf{E}_h^n - \mathbf{E}^n|_h \leq C(\tau^2 + h^l).$$

4.3 Discontinuous Galerkin Methods for the Drude Model

Taking the product of Eqs. (1.18)–(1.21) by test functions $\mathbf{u}, \mathbf{v}, \boldsymbol{\phi}, \boldsymbol{\psi}$ and integrating by parts over any element $T_i \in T_h$, we have

$$\epsilon_0 \int_{T_i} \frac{\partial \mathbf{E}}{\partial t} \cdot \mathbf{u} - \int_{T_i} \mathbf{H} \cdot \nabla \times \mathbf{u} - \int_{\partial T_i} \mathbf{n}_i \times \mathbf{H} \cdot \mathbf{u} + \int_{T_i} \mathbf{J} \cdot \mathbf{u} = 0, \tag{4.25}$$

$$\mu_0 \int_{T_i} \frac{\partial \mathbf{H}}{\partial t} \cdot \mathbf{v} + \int_{T_i} \mathbf{E} \cdot \nabla \times \mathbf{v} + \int_{\partial T_i} \mathbf{n}_i \times \mathbf{E} \cdot \mathbf{v} + \int_{T_i} \mathbf{K} \cdot \mathbf{v} = 0, \tag{4.26}$$

$$\frac{1}{\epsilon_0 \omega_{pe}^2} \int_{T_i} \frac{\partial \mathbf{J}}{\partial t} \cdot \phi + \frac{\Gamma_e}{\epsilon_0 \omega_{pe}^2} \int_{T_i} \mathbf{J} \cdot \phi = \int_{T_i} \mathbf{E} \cdot \phi, \tag{4.27}$$

$$\frac{1}{\mu_0 \omega_{pm}^2} \int_{T_i} \frac{\partial \mathbf{K}}{\partial t} \cdot \psi + \frac{\Gamma_m}{\mu_0 \omega_{pm}^2} \int_{T_i} \mathbf{K} \cdot \psi = \int_{T_i} \mathbf{H} \cdot \psi. \tag{4.28}$$

Let us look at the semi-discrete solution $\mathbf{E}_h, \mathbf{H}_h, \mathbf{J}_h, \mathbf{K}_h \in C^1(0, T; \mathbf{V}_h)$ as a solution of the following weak formulation: For any $\mathbf{u}_h, \mathbf{v}_h, \phi_h, \psi_h \in \mathbf{V}_h$, and any element $T_i \in T_h$,

$$\epsilon_0 \int_{T_i} \frac{\partial \mathbf{E}_h}{\partial t} \cdot \mathbf{u}_h - \int_{T_i} \mathbf{H}_h \cdot \nabla \times \mathbf{u}_h - \sum_{K \in \nu_i} \int_{a_{ik}} \mathbf{u}_h \cdot \mathbf{n}_{ik} \times \{\{\mathbf{H}_h\}\}_{ik} + \int_{T_i} \mathbf{J}_h \cdot \mathbf{u}_h = 0, \tag{4.29}$$

$$\mu_0 \int_{T_i} \frac{\partial \mathbf{H}_h}{\partial t} \cdot \mathbf{v}_h + \int_{T_i} \mathbf{E}_h \cdot \nabla \times \mathbf{v}_h + \sum_{K \in \nu_i} \int_{a_{ik}} \mathbf{v}_h \cdot \mathbf{n}_{ik} \times \{\{\mathbf{E}_h\}\}_{ik} + \int_{T_i} \mathbf{K}_h \cdot \mathbf{v}_h = 0, \tag{4.30}$$

$$\frac{1}{\epsilon_0 \omega_{pe}^2} \int_{T_i} \frac{\partial \mathbf{J}_h}{\partial t} \cdot \phi_h + \frac{\Gamma_e}{\epsilon_0 \omega_{pe}^2} \int_{T_i} \mathbf{J}_h \cdot \phi_h = \int_{T_i} \mathbf{E}_h \cdot \phi_h, \tag{4.31}$$

$$\frac{1}{\mu_0 \omega_{pm}^2} \int_{T_i} \frac{\partial \mathbf{K}_h}{\partial t} \cdot \psi_h + \frac{\Gamma_m}{\mu_0 \omega_{pm}^2} \int_{T_i} \mathbf{K}_h \cdot \psi_h = \int_{T_i} \mathbf{H}_h \cdot \psi_h, \tag{4.32}$$

hold true and are subject to the initial conditions:

$$\mathbf{E}_h(0) = \Pi_2 \mathbf{E}_0, \ \mathbf{H}_h(0) = \Pi_2 \mathbf{H}_0, \ \mathbf{J}_h(0) = \Pi_2 \mathbf{J}_0, \ \mathbf{K}_h(0) = \Pi_2 \mathbf{K}_0, \tag{4.33}$$

where Π_2 denotes the standard L^2-projection onto $\mathbf{V}_h$. Recall that $\mathbf{E}_0, \mathbf{H}_0, \mathbf{J}_0$ and $\mathbf{K}_0$ are the given initial condition functions. Here we denote V_i for the set of indices of all neighboring elements of T_i and a_{ik} for the internal face $a_{ik} = T_i \cap T_k$.

Denote the semi-discrete energy $\mathscr{E}_h$:

$$\mathscr{E}_h(t) = \frac{1}{2}(\epsilon_0 ||\mathbf{E}_h(t)||_0^2 + \mu_0 ||\mathbf{H}_h(t)||_0^2 + \frac{1}{\epsilon_0 \omega_{pe}^2} ||\mathbf{J}_h(t)||_0^2 + \frac{1}{\mu_0 \omega_{pm}^2} ||\mathbf{K}_h(t)||_0^2), \tag{4.34}$$

and a bilinear form $\mathscr{B}_i$:

$$\mathscr{B}_i(\mathbf{E},\mathbf{H}) = -\int_{T_i} \mathbf{H}_i \cdot \nabla \times \mathbf{E}_i - \sum_{K\in v_i} \int_{a_{ik}} \mathbf{E}_h \cdot \mathbf{n}_{ik} \times \{\{\mathbf{H}_h\}\}_{ik} + \int_{T_i} \mathbf{E}_i \cdot \nabla \times \mathbf{H}_i + \sum_{K\in v_i} \int_{a_{ik}} \mathbf{H}_h \cdot \mathbf{n}_{ik} \times \{\{\mathbf{E}_h\}\}_{ik}. \tag{4.35}$$

Theorem 4.5. *The energy $\mathscr{E}_h$ is decreasing in time, i.e., $\mathscr{E}_h(t) \le \mathscr{E}_h(0)$.*

Proof. Choosing $\mathbf{u}_h = \mathbf{E}_h, \mathbf{v}_h = \mathbf{H}_h, \phi_h = \mathbf{J}_h, \psi_h = \mathbf{K}_h$ in (4.29)–(4.32) and adding the results together over all element $T_i \in T_h$, we obtain

$$\frac{d}{dt}\mathscr{E}_h(t) + \frac{\Gamma_e}{\epsilon_0\omega_{pe}^2}||\mathbf{J}_h(t)||_0^2 + \frac{\Gamma_m}{\mu_0\omega_{pm}^2}||\mathbf{K}_h(t)||_0^2 + \sum_i \mathscr{B}_i(\mathbf{E},\mathbf{H}) = 0. \tag{4.36}$$

By the definition of $\mathscr{B}_i$ and integration by parts, we have

$$\begin{aligned}\mathscr{B}_i(\mathbf{E},\mathbf{H}) &= \sum_{K\in v_i} \int_{a_{ik}} \mathbf{E}_i \cdot \mathbf{n}_{ik} \times \mathbf{H}_i \\ &\quad + \sum_{K\in v_i} \int_{a_{ik}} \mathbf{E}_i \cdot \{\{\mathbf{H}_h\}\}_{ik} \times \mathbf{n}_{ik} - \sum_{K\in v_i} \int_{a_{ik}} \mathbf{H}_i \cdot \{\{\mathbf{E}_h\}\}_{ik} \times \mathbf{n}_{ik} \\ &= \sum_{K\in v_i} \int_{a_{ik}} [-\mathbf{E}_i \times \mathbf{H}_i + \mathbf{E}_i \times \frac{\mathbf{H}_i + \mathbf{H}_k}{2} - \mathbf{H}_i \times \frac{\mathbf{E}_i + \mathbf{E}_k}{2}] \cdot \mathbf{n}_{ik} \\ &= \frac{1}{2} \sum_{K\in v_i} \int_{a_{ik}} (\mathbf{E}_i \times \mathbf{H}_k + \mathbf{E}_k \times \mathbf{H}_i) \cdot \mathbf{n}_{ik}. \end{aligned} \tag{4.37}$$

From (4.37), we obtain $\sum_i \mathscr{B}_i(\mathbf{E},\mathbf{H}) = 0$, which, along with (4.36), concludes the proof. □

For the semi-discrete scheme (4.29)–(4.32), we have the following convergence result.

Theorem 4.6. *If $\mathbf{E},\mathbf{H},\mathbf{J},\mathbf{K} \in C^0([0,T];(H^{s+1}(\Omega))^3)$ for $s \ge 0$, then there exists a constant $C > 0$ independent of h such that*

$$\begin{aligned}&\max_{t\in[0,T]} (||\mathbf{E}-\mathbf{E}_h||_0 + ||\mathbf{H}-\mathbf{H}_h||_0 + ||\mathbf{J}-\mathbf{J}_h||_0 + ||\mathbf{K}-\mathbf{K}_h||_0) \\ &\le C h^{\min(s,k)} ||(\mathbf{E},\mathbf{H},\mathbf{J},\mathbf{K})||_{C^0([0,T];(H^{s+1}(\Omega))^3)}. \end{aligned} \tag{4.38}$$

Proof. Let us introduce the notation $\tilde{\mathbf{W}}_h = \Pi_2(\mathbf{W}) - \mathbf{W}_h$ and $\overline{\mathbf{W}}_h = \Pi_2(\mathbf{W}) - \mathbf{W}$ for $\mathbf{W} = \mathbf{E},\mathbf{H},\mathbf{J},\mathbf{K}$.

Subtracting (4.29)–(4.32) from (4.25)–(4.28), we have the error equations:

$$(i)\ \epsilon_0 \int_{T_i} \frac{\partial \tilde{\mathbf{E}}_h}{\partial t} \cdot \mathbf{u}_h - \int_{T_i} \tilde{\mathbf{H}}_h \cdot \nabla \times \mathbf{u}_h - \sum_{K \in \nu_i} \int_{a_{ik}} \mathbf{u}_h \cdot \mathbf{n}_{ik} \times \{\{\tilde{\mathbf{H}}_h\}\}_{ik} + \int_{T_i} \tilde{\mathbf{J}}_h \cdot \mathbf{u}_h$$
$$= \epsilon_0 \int_{T_i} \frac{\partial \overline{\mathbf{E}}_h}{\partial t} \cdot \mathbf{u}_h - \int_{T_i} \overline{\mathbf{H}}_h \cdot \nabla \times \mathbf{u}_h - \sum_{K \in \nu_i} \int_{a_{ik}} \mathbf{u}_h \cdot \mathbf{n}_{ik} \times \{\{\overline{\mathbf{H}}_h\}\}_{ik} + \int_{T_i} \overline{\mathbf{J}}_h \cdot \mathbf{u}_h, \tag{4.39}$$

$$(ii)\ \mu_0 \int_{T_i} \frac{\partial \tilde{\mathbf{H}}_h}{\partial t} \cdot \mathbf{v}_h + \int_{T_i} \tilde{\mathbf{E}}_h \cdot \nabla \times \mathbf{v}_h + \sum_{K \in \nu_i} \int_{a_{ik}} \mathbf{v}_h \cdot \mathbf{n}_{ik} \times \{\{\tilde{\mathbf{E}}_h\}\}_{ik} + \int_{T_i} \tilde{\mathbf{K}}_h \cdot \mathbf{v}_h$$
$$= \mu_0 \int_{T_i} \frac{\partial \overline{\mathbf{H}}_h}{\partial t} \cdot \mathbf{v}_h + \int_{T_i} \overline{\mathbf{E}}_h \cdot \nabla \times \mathbf{v}_h + \sum_{K \in \nu_i} \int_{a_{ik}} \mathbf{v}_h \cdot \mathbf{n}_{ik} \times \{\{\overline{\mathbf{E}}_h\}\}_{ik} + \int_{T_i} \overline{\mathbf{K}}_h \cdot \mathbf{v}_h, \tag{4.40}$$

$$(iii)\ \frac{1}{\epsilon_0 \omega_{pe}^2} \int_{T_i} \frac{\partial \tilde{\mathbf{J}}_h}{\partial t} \cdot \phi_h + \frac{\Gamma_e}{\epsilon_0 \omega_{pe}^2} \int_{T_i} \tilde{\mathbf{J}}_h \cdot \phi_h - \int_{T_i} \tilde{\mathbf{E}}_h \cdot \phi_h$$
$$= \frac{1}{\epsilon_0 \omega_{pe}^2} \int_{T_i} \frac{\partial \overline{\mathbf{J}}_h}{\partial t} \cdot \phi_h + \frac{\Gamma_e}{\epsilon_0 \omega_{pe}^2} \int_{T_i} \overline{\mathbf{J}}_h \cdot \phi_h - \int_{T_i} \overline{\mathbf{E}}_h \cdot \phi_h, \tag{4.41}$$

$$(iv)\ \frac{1}{\mu_0 \omega_{pm}^2} \int_{T_i} \frac{\partial \tilde{\mathbf{K}}_h}{\partial t} \cdot \psi_h + \frac{\Gamma_m}{\mu_0 \omega_{pm}^2} \int_{T_i} \tilde{\mathbf{K}}_h \cdot \psi_h - \int_{T_i} \tilde{\mathbf{H}}_h \cdot \psi_h$$
$$= \frac{1}{\mu_0 \omega_{pm}^2} \int_{T_i} \frac{\partial \overline{\mathbf{K}}_h}{\partial t} \cdot \psi_h + \frac{\Gamma_m}{\mu_0 \omega_{pm}^2} \int_{T_i} \overline{\mathbf{K}}_h \cdot \psi_h - \int_{T_i} \overline{\mathbf{H}}_h \cdot \psi_h. \tag{4.42}$$

Choosing $\mathbf{u}_h = \tilde{\mathbf{E}}_h, \mathbf{v}_h = \tilde{\mathbf{H}}_h, \phi_h = \tilde{\mathbf{J}}_h, \psi_h = \tilde{\mathbf{K}}_h$ in (4.39)–(4.42), summing up the results for all elements T_i of T_h, then using the projection property and the energy definition (4.34), we have

$$\frac{d}{dt}\tilde{\mathscr{E}}_h + \frac{\Gamma_e}{\epsilon_0 \omega_{pe}^2} ||\tilde{\mathbf{J}}_h||_0^2 + \frac{\Gamma_m}{\mu_0 \omega_{pm}^2} ||\tilde{\mathbf{K}}_h||_0^2 + \sum_i \mathscr{B}_i(\tilde{\mathbf{E}}, \tilde{\mathbf{H}})$$
$$= \sum_i \sum_{K \in \nu_i} [\int_{a_{ik}} \tilde{\mathbf{E}}_h \cdot \mathbf{n}_{ik} \times \{\{\overline{\mathbf{H}}_h\}\}_{ik} + \int_{a_{ik}} \tilde{\mathbf{H}}_h \cdot \mathbf{n}_{ik} \times \{\{\overline{\mathbf{E}}_h\}\}_{ik}]$$
$$\leq \sum_i [||\tilde{\mathbf{E}}_h||_{0,\partial T_i} ||\overline{\mathbf{H}}_h||_{0,\partial T_i} + ||\tilde{\mathbf{H}}_h||_{0,\partial T_i} ||\overline{\mathbf{E}}_h||_{0,\partial T_i}]$$
$$\leq \sum_i [C h_{T_i}^{-\frac{1}{2}} ||\tilde{\mathbf{E}}_h||_{0,T_i} C h_{T_i}^{\min(s,k)+\frac{1}{2}} ||\mathbf{H}||_{s+1,T_i} + C h_{T_i}^{-\frac{1}{2}} ||\tilde{\mathbf{H}}_h||_{0,T_i} C h_{T_i}^{\min(s,k)+\frac{1}{2}} ||\mathbf{E}||_{s+1,T_i}],$$

where in the last step we used the standard inverse inequality and interpolation error estimate.

The proof is completed by using the fact $\sum_i \mathscr{B}_i(\tilde{\mathbf{E}}, \tilde{\mathbf{H}}) = 0$ and the Gronwall inequality. □

Similar to those fully-discrete schemes developed in Chap. 3, we can construct a simple leap-frog scheme as follows: find $\mathbf{E}_h^{n+1}, \mathbf{H}_h^{n+\frac{3}{2}}, \mathbf{J}_h^{n+\frac{3}{2}}, \mathbf{K}_h^{n+1} \in \mathbf{V}_h$ such that for any $\mathbf{u}_h, \mathbf{v}_h, \phi_h, \psi_h \in \mathbf{V}_h$, and any element $T_i \in T_h$,

$$
\begin{aligned}
&\epsilon_0 \int_{T_i} \frac{\mathbf{E}_i^{n+1} - \mathbf{E}_i^n}{\tau} \cdot \mathbf{u}_h - \int_{T_i} \mathbf{H}_i^{n+\frac{1}{2}} \cdot \nabla \times \mathbf{u}_h \\
&\quad - \sum_{K \in v_i} \int_{a_{ik}} \mathbf{u}_h \cdot \mathbf{n}_{ik} \times \{\{\mathbf{H}_h^{n+\frac{1}{2}}\}\}_{ik} + \int_{T_i} \mathbf{J}_i^{n+\frac{1}{2}} \cdot \mathbf{u}_h = 0, \\
&\mu_0 \int_{T_i} \frac{\mathbf{H}_h^{n+\frac{3}{2}} - \mathbf{H}_h^{n+\frac{1}{2}}}{\tau} \cdot \mathbf{v}_h + \int_{T_i} \mathbf{E}_i^{n+1} \cdot \nabla \times \mathbf{v}_h \\
&\quad + \sum_{K \in v_i} \int_{a_{ik}} \mathbf{v}_h \cdot \mathbf{n}_{ik} \times \{\{\mathbf{E}_h^{n+1}\}\}_{ik} + \int_{T_i} \mathbf{K}_i^{n+1} \cdot \mathbf{v}_h = 0, \\
&\frac{1}{\epsilon_0 \omega_{pe}^2} \int_{T_i} \frac{\mathbf{J}_i^{n+\frac{3}{2}} - \mathbf{J}_i^{n+\frac{1}{2}}}{\tau} \cdot \phi_h + \frac{\Gamma_e}{\epsilon_0 \omega_{pe}^2} \int_{T_i} \frac{\mathbf{J}_i^{n+\frac{3}{2}} + \mathbf{J}_i^{n+\frac{1}{2}}}{2} \cdot \phi_h = \int_{T_i} \mathbf{E}_i^{n+1} \cdot \phi_h, \\
&\frac{1}{\mu_0 \omega_{pm}^2} \int_{T_i} \frac{\mathbf{K}_i^{n+1} - \mathbf{K}_i^n}{\tau} \cdot \psi_h + \frac{\Gamma_m}{\mu_0 \omega_{pm}^2} \int_{T_i} \frac{\mathbf{K}_i^{n+1} + \mathbf{K}_i^n}{2} \cdot \psi_h = \int_{T_i} \mathbf{H}_i^{n+\frac{1}{2}} \cdot \psi_h,
\end{aligned}
$$

subject to the initial conditions (4.33). Stability and convergence analysis can be carried out for this scheme. We leave the details to interested readers.

4.4 Nodal Discontinuous Galerkin Methods for the Drude Model

In this section, we extend the nodal discontinuous Galerkin methods developed by Hesthaven and Warburton [141] for general conservation laws to solve Maxwell's equations when metamaterials are involved. The package *nudg* developed in [141] provides a very good template for solving many common partial differential equations such as elliptic problems, Euler equations, Maxwell's equations and Navier-Stokes equations. Here we provide detailed MATLAB source codes to show readers how to modify the package *nudg* to solve the metamaterial Maxwell's equations. The contents of this section are mainly derived from Li [185].

4.4.1 The Algorithm

To simplify the presentation, we first non-dimensionalize the Drude model equations (1.18)–(1.21). Let us introduce the vacuum speed of light C_v, the vacuum impedance Z_0:

$$C_v = \frac{1}{\sqrt{\epsilon_0 \mu_0}} \approx 3 \times 10^8 \text{ m/s}, \quad Z_0 = \sqrt{\mu_0/\epsilon_0} \approx 120\pi \text{ ohms},$$

and unit-free variables

$$\tilde{t} = \frac{C_v t}{L}, \quad \tilde{x} = \frac{x}{L},$$
$$\tilde{\Gamma}_e = \frac{\Gamma_e L}{C_v}, \quad \tilde{\omega}_{pe} = \frac{\omega_{pe} L}{C_v}, \quad \tilde{\Gamma}_m = \frac{\Gamma_m L}{C_v}, \quad \tilde{\omega}_{pm} = \frac{\omega_{pm} L}{C_v},$$
$$\tilde{\mathbf{E}} = \frac{\mathbf{E}}{Z_0 H_0}, \quad \tilde{\mathbf{H}} = \frac{\mathbf{H}}{H_0}, \quad \tilde{\mathbf{J}} = \frac{L\mathbf{J}}{H_0}, \quad \tilde{\mathbf{K}} = \frac{L\mathbf{K}}{Z_0 H_0},$$

where H_0 is a unit magnetic field strength, and L is a reference length (typically the wavelength of one interested object).

It is not difficult to check that the equations (1.18)–(1.21) can be written as

$$\frac{\partial \tilde{\mathbf{E}}}{\partial \tilde{t}} = \nabla \times \tilde{\mathbf{H}} - \tilde{\mathbf{J}}, \tag{4.43}$$
$$\frac{\partial \tilde{\mathbf{H}}}{\partial \tilde{t}} = -\nabla \times \tilde{\mathbf{E}} - \tilde{\mathbf{K}}, \tag{4.44}$$
$$\frac{\partial \tilde{\mathbf{J}}}{\partial \tilde{t}} + \tilde{\Gamma}_e \tilde{\mathbf{J}} = \tilde{\omega}_e^2 \tilde{\mathbf{E}}, \tag{4.45}$$
$$\frac{\partial \tilde{\mathbf{K}}}{\partial \tilde{t}} + \tilde{\Gamma}_m \tilde{\mathbf{K}} = \tilde{\omega}_m^2 \tilde{\mathbf{H}}, \tag{4.46}$$

which have the same form as the original governing equations (1.18)–(1.21) if we set $\epsilon_0 = \mu_0 = 1$ in (1.18)–(1.21).

In the rest of this section, our discussion is based on the non-dimensionalized form (4.43)–(4.46) by dropping all those tildes and adding fixed sources $\mathbf{f}$ and $\mathbf{g}$ to (4.43) and (4.44), i.e.,

$$\frac{\partial \mathbf{E}}{\partial t} = \nabla \times \mathbf{H} - \mathbf{J} + \mathbf{f}, \tag{4.47}$$
$$\frac{\partial \mathbf{H}}{\partial \tilde{t}} = -\nabla \times \mathbf{E} - \mathbf{K} + \mathbf{g}, \tag{4.48}$$

$$\frac{\partial \mathbf{J}}{\partial t} + \Gamma_e \mathbf{J} = \omega_e^2 \mathbf{E}, \tag{4.49}$$

$$\frac{\partial \mathbf{K}}{\partial t} + \Gamma_m \mathbf{K} = \omega_m^2 \mathbf{H}, \tag{4.50}$$

Using the same idea as [140], we can rewrite (4.47) and (4.48) in the conservation form

$$\frac{\partial \mathbf{q}}{\partial t} + \nabla \cdot \psi(\mathbf{q}) = \mathbf{S}, \tag{4.51}$$

where we denote

$$\mathbf{q} = \begin{bmatrix} \mathbf{E} \\ \mathbf{H} \end{bmatrix}, \quad \mathbf{S} \equiv \begin{bmatrix} S_E \\ S_H \end{bmatrix} = \begin{bmatrix} -\mathbf{J} + \mathbf{f} \\ -\mathbf{K} + \mathbf{g} \end{bmatrix}, \quad F_i(\mathbf{q}) = \begin{bmatrix} -\mathbf{e}_i \times \mathbf{H} \\ \mathbf{e}_i \times \mathbf{E} \end{bmatrix},$$

and $\psi(\mathbf{q}) = [F_1(\mathbf{q}), F_2(\mathbf{q}), F_3(\mathbf{q})]^T$. Here $\mathbf{e}_i$ are the three Cartesian unit vectors.

We assume that the domain Ω is decomposed into tetrahedral (or triangular in 2-D) elements Ω_k, and the numerical solution $\mathbf{q}_N$ is represented as

$$\mathbf{q}_N(\mathbf{x}, t) = \sum_{j=1}^{N_n} \mathbf{q}_j(\mathbf{x}_j, t) L_j(\mathbf{x}) = \sum_{j=1}^{N_n} \mathbf{q}_j(t) L_j(\mathbf{x}), \tag{4.52}$$

where $L_j(\mathbf{x})$ is the multivariate Lagrange interpolation polynomial of degree n. Here $N_n = \frac{1}{6}(n+1)(n+2)(n+3)$ in 3-D; while $N_n = \frac{1}{2}(n+1)(n+2)$ in 2-D.

Multiplying (4.51) by a test function $L_i(\mathbf{x})$ and integrating over each element Ω_k, we obtain

$$\int_{\Omega_k} \left(\frac{\partial \mathbf{q}_N}{\partial t} + \nabla \cdot \psi(\mathbf{q}_N) - \mathbf{S}_N\right) L_i(\mathbf{x}) dx = \int_{\partial \Omega_k} \hat{\mathbf{n}} \cdot (\psi(\mathbf{q}_N) - \psi_N^*) L_i(\mathbf{x}) dx, \tag{4.53}$$

where $\hat{\mathbf{n}}$ is an outward normal unit vector of $\partial \Omega_k$, and ψ_N^* is a numerical flux. For the Maxwell's equations, we usually choose the upwind flux [140]

$$\hat{\mathbf{n}} \cdot (\psi(\mathbf{q}_N) - \psi_N^*) = \begin{cases} \frac{1}{2}\hat{\mathbf{n}} \times ([\mathbf{H}_N] - \hat{\mathbf{n}} \times [\mathbf{E}_N]) \\ \frac{1}{2}\hat{\mathbf{n}} \times (-\hat{\mathbf{n}} \times [\mathbf{H}_N] - [\mathbf{E}_N]) \end{cases},$$

where $[\mathbf{E}_N] = \mathbf{E}_N^+ - \mathbf{E}_N^-$, and $[\mathbf{H}_N] = \mathbf{H}_N^+ - \mathbf{H}_N^-$. Here superscripts '+' and '−' refer to field values from the neighboring element and the local element, respectively.

Substituting (4.52) into (4.53), we obtain the elementwise equations for the electric field components

$$\sum_{j=0}^{N} \left(M_{ij} \frac{d\mathbf{E}_j}{dt} - S_{ij} \times \mathbf{H}_j - M_{ij} \mathbf{S}_{E,j}\right) = \frac{1}{2} \sum_l F_{il} \cdot \hat{\mathbf{n}}_l \times ([\mathbf{H}_l] - \hat{\mathbf{n}}_l \times [\mathbf{E}_l]), \tag{4.54}$$

and for the magnetic field components

$$\sum_{j=0}^{N}(M_{ij}\frac{d\mathbf{H}_j}{dt}+S_{ij}\times\mathbf{E}_j-M_{ij}\mathbf{S}_{H,j})=\frac{1}{2}\sum_l F_{il}\cdot\hat{\mathbf{n}}_l\times(-\hat{\mathbf{n}}_l\times[\mathbf{H}_l]-[\mathbf{E}_l]),\quad(4.55)$$

where

$$M_{ij}=(L_i(\mathbf{x}),L_j(\mathbf{x}))_{\Omega_k},\quad S_{ij}=(L_i(\mathbf{x}),\nabla L_j(\mathbf{x}))_{\Omega_k}$$

represent the local mass and stiffness matrices, respectively. Furthermore,

$$F_{il}=(L_i(\mathbf{x}),L_l(\mathbf{x}))_{\partial\Omega_k}$$

represents the face-based mass matrix.

We can rewrite (4.54) and (4.55) in a fully explicit form, while the constitutive equations (4.49) and (4.50) keep the same form. In summary, we have the following semi-discrete discontinuous Galerkin scheme:

$$\frac{d\mathbf{E}_N}{dt}=M^{-1}S\times\mathbf{H}_N-\mathbf{J}_N+\mathbf{f}_N+\frac{1}{2}M^{-1}F\Big(\hat{\mathbf{n}}\times([\mathbf{H}_N]-\hat{\mathbf{n}}\times[\mathbf{E}_N])\Big)|_{\partial\Omega_k},\quad(4.56)$$

$$\frac{d\mathbf{H}_N}{dt}=-M^{-1}S\times\mathbf{E}_N-\mathbf{K}_N+\mathbf{g}_N-\frac{1}{2}M^{-1}F\Big(\hat{\mathbf{n}}\times(\hat{\mathbf{n}}\times[\mathbf{H}_N]+[\mathbf{E}_N])\Big)|_{\partial\Omega_k},$$

$$\frac{d\mathbf{J}_N}{dt}=\omega_e^2\mathbf{E}_N-\Gamma_e\mathbf{J}_N,\quad(4.57)$$

$$\frac{d\mathbf{K}_N}{dt}=\omega_m^2\mathbf{H}_N-\Gamma_m\mathbf{K}_N.\quad(4.58)$$

The system (4.56)–(4.58) can be solved by various methods used for ordinary differential equations. Below we adopt the classic low-storage five-stage fourth-order explicit Runge-Kutta method [141, Sect. 3.4].

4.4.2 MATLAB Codes and Numerical Results

We implement the above algorithm using the package *nudg* provided by Hesthaven and Warburton [141]. Considering that the 3-D case is quite similar to the 2-D case (though computational time in 3-D is much longer), here we only consider the 2-D transverse magnetic mode with respect to z (TM_z: no magnetic field in z-direction) metamaterial model:

$$\frac{\partial H_x}{\partial t}=-\frac{\partial E_z}{\partial y}-K_x+g_x\quad(4.59)$$

$$\frac{\partial H_y}{\partial t}=\frac{\partial E_z}{\partial x}-K_y+g_y\quad(4.60)$$

$$\frac{\partial E_z}{\partial t} = \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} - J_z + f \tag{4.61}$$

$$\frac{\partial J_z}{\partial t} = \omega_e^2 E_z - \Gamma_e J_z \tag{4.62}$$

$$\frac{\partial K_x}{\partial t} = \omega_m^2 H_x - \Gamma_m K_x \tag{4.63}$$

$$\frac{\partial K_y}{\partial t} = \omega_m^2 H_y - \Gamma_m K_y \tag{4.64}$$

where the subscripts 'x, y' and 'z' denote the corresponding components.

For the metamaterial model (4.59)–(4.64), we only need to provide three main MATLAB functions: *Meta2DDriver.m*, *MetaRHS2D.m*, and *Meta2D.m*. The rest supporting functions are provided by the package *nudg* of Hesthaven and Warburton [141].

To check the convergence rate, we construct the following exact solutions for the 2-D TM model (assuming that $\Gamma_m = \Gamma_e = \omega_m = \omega_e = 1$) on domain $\Omega = (0,1)^2$:

$$\mathbf{H} \equiv \begin{pmatrix} H_x \\ H_y \end{pmatrix} = \begin{pmatrix} \sin(\omega\pi x)\cos(\omega\pi y)\exp(-t) \\ -\cos(\omega\pi x)\sin(\omega\pi y)\exp(-t) \end{pmatrix},$$

$$E_z = \sin(\omega\pi x)\sin(\omega\pi y)\exp(-t).$$

The corresponding magnetic and electric currents are

$$\mathbf{K} \equiv \begin{pmatrix} K_x \\ K_y \end{pmatrix} = \begin{pmatrix} t\sin(\omega\pi x)\cos(\omega\pi y)\exp(-t) \\ -t\cos(\omega\pi x)\sin(\omega\pi y)\exp(-t) \end{pmatrix},$$

and

$$J_z = t\sin(\omega\pi x)\sin(\omega\pi y)\exp(-t),$$

respectively. The corresponding source term

$$f = (t - 1 - 2\omega\pi)\sin(\omega\pi x)\sin(\omega\pi y)\exp(-t),$$

while $\mathbf{g} = (g_x, g_y)'$ is given by

$$g_x = (\omega\pi - 1 + t)\sin(\omega\pi x)\cos(\omega\pi y)\exp(-t),$$
$$g_y = (1 - \omega\pi - t)\cos(\omega\pi x)\sin(\omega\pi y)\exp(-t),$$

Notice that E_z satisfies the boundary condition $E_z = 0$ on $\partial\Omega$.

The function *MetaRHS2D.m* is used to evaluate the right-hand-side flux in the 2-D TM form. Its detailed implementation is shown below:

```
function [rhsHx, rhsHy, rhsEz] = MetaRHS2D(Hx,Hy,Ez)
% Purpose: Evaluate RHS flux in 2D Maxwell TM form
```

```
Globals2D;

% Define field differences at faces
dHx = zeros(Nfp*Nfaces,K); dHx(:) = Hx(vmapM)-Hx(vmapP);
dHy = zeros(Nfp*Nfaces,K); dHy(:) = Hy(vmapM)-Hy(vmapP);
dEz = zeros(Nfp*Nfaces,K); dEz(:) = Ez(vmapM)-Ez(vmapP);

% Impose reflective boundary conditions (Ez+ = -Ez-)
dHx(mapB) = 0; dHy(mapB) = 0; dEz(mapB) = 2*Ez(vmapB);

% upwind flux (alpha = 1.0); central flux (alpha = 0.0);
alpha = 1.0;
ndotdH =  nx.*dHx+ny.*dHy;
fluxHx =  ny.*dEz + alpha*(ndotdH.*nx-dHx);
fluxHy = -nx.*dEz + alpha*(ndotdH.*ny-dHy);
fluxEz = -nx.*dHy + ny.*dHx - alpha*dEz;

% local derivatives of fields
[Ezx,Ezy] = Grad2D(Ez);
[CuHx,CuHy,CuHz] = Curl2D(Hx,Hy,[]);

% compute right hand sides of the PDE's
rhsHx = -Ezy  + LIFT*(Fscale.*fluxHx)/2.0;
rhsHy = Ezx  + LIFT*(Fscale.*fluxHy)/2.0;
rhsEz =  CuHz + LIFT*(Fscale.*fluxEz)/2.0;

return;
```

The function *Meta2D.m* is used to perform the time-marching using a classic low-storage five-stage fourth-order explicit Runge-Kutta method. The code *Meta2D.m* is shown below:

```
function [Hx,Hy,Ez,Kx,Ky,Jz,time] = ...
         Meta2D(Hx,Hy,Ez,Kx,Ky,Jz,x,y,FinalT)

% Purpose: Integrate TM-mode Maxwell equations until
%  FinalT starting with initial conditions Hx,Hy,Ez

Globals2D;
time = 0;

% Runge-Kutta residual storage
resHx = zeros(Np,K);
resHy = zeros(Np,K);
resEz = zeros(Np,K);
resKx = zeros(Np,K);
resKy = zeros(Np,K);
resJz = zeros(Np,K);
```

```
dt = 1e-6; omepi=4*pi;
istep=0;
% outer time step loop
while (time<FinalT)

  if(time+dt>FinalT), dt = FinalT-time; end
  f=feval(@fun_f21,x,y,time,omepi);
  gx=feval(@fun_gx21,x,y,time,omepi);
  gy=feval(@fun_gy21,x,y,time,omepi);
  for INTRK = 1:5
       rhsKx = -Kx+Hx; rhsKy = -Ky+Hy; rhsJz = -Jz+Ez;
       resKx = rk4a(INTRK)*resKx+dt*rhsKx;
       resKy = rk4a(INTRK)*resKy+dt*rhsKy;
       resJz = rk4a(INTRK)*resJz+dt*rhsJz;

      % compute RHS of TM-mode Maxwell equations
      [rhsHx, rhsHy, rhsEz] = MetaRHS2D(Hx,Hy,Ez);
      rhsHx = rhsHx-Kx+gx;
      rhsHy = rhsHy-Ky+gy;
      rhsEz = rhsEz-Jz+f;
      % initiate and increment Runge-Kutta residuals
      resHx = rk4a(INTRK)*resHx + dt*rhsHx;
      resHy = rk4a(INTRK)*resHy + dt*rhsHy;
      resEz = rk4a(INTRK)*resEz + dt*rhsEz;

      % update fields
      Hx = Hx+rk4b(INTRK)*resHx;
      Hy = Hy+rk4b(INTRK)*resHy;
      Ez = Ez+rk4b(INTRK)*resEz;
      Kx = Kx+rk4b(INTRK)*resKx;
      Ky = Ky+rk4b(INTRK)*resKy;
      Jz = Jz+rk4b(INTRK)*resJz;
   end;
   % Increment time
   time = time+dt;
   istep = istep + 1;
   disp('step, time ='), istep, time
end
return
```

The function *Meta2DDriver.m* is the driver script. The detailed implementation for our example is shown below:

```
% Driver script for 2D metamaterial equations

Globals2D;
```

```
% Polynomial order used for approximation
N = 1;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Generate a uniform triangular grid

nelex=20;     % number of elements in x-direction
nx = nelex+1;   % number of points in x direction
Nv = nx*nx;       % total number of grid points
K = 2*nelex*nelex;  % total number of elements
no2xy = genrecxygrid(0,1,0,1,nx,nx)';
VX = no2xy(1,:); VY=no2xy(2,:);
EToV = delaunay(VX,VY);

% Reorder elements to ensure counterclockwise order
ax = VX(EToV(:,1)); ay = VY(EToV(:,1));
bx = VX(EToV(:,2)); by = VY(EToV(:,2));
cx = VX(EToV(:,3)); cy = VY(EToV(:,3));

D = (ax-cx).*(by-cy)-(bx-cx).*(ay-cy);
i = find(D<0);
EToV(i,:) = EToV(i,[1 3 2]);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Initialize solver and construct grid and metric

StartUp2D;

% Set initial conditions
omepi = 4*pi;   % omega*pi always together
Hx = sin(omepi*x).*cos(omepi*y);
Hy = -cos(omepi*x).*sin(omepi*y);
Ez = sin(omepi*x).*sin(omepi*y);
Kx = zeros(Np,K); Ky=zeros(Np,K); Jz=zeros(Np,K);

% Solve Problem
FinalT = 1e2*1.0e-6;
% measure elapsed time.
tic
[Hx,Hy,Ez,Kx,Ky,Jz,time] ...
      = Meta2D(Hx,Hy,Ez,Kx,Ky,Jz,x,y,FinalT);
toc

exactHx=sin(omepi*x).*cos(omepi*y)*exp(-FinalT);
exactHy=-cos(omepi*x).*sin(omepi*y)*exp(-FinalT);
exactEz=sin(omepi*x).*sin(omepi*y)*exp(-FinalT);
exactKx=FinalT*sin(omepi*x).*cos(omepi*y)*exp(-FinalT);
exactKy=-FinalT*cos(omepi*x).*sin(omepi*y)*exp(-FinalT);
exactJz=FinalT*sin(omepi*x).*sin(omepi*y)*exp(-FinalT);

errorHx = max(max(abs(Hx-exactHx))),
errorHy = max(max(abs(Hy-exactHy))),
errorEz = max(max(abs(Ez-exactEz))),
errorKx = max(max(abs(Kx-exactKx))),
```

```
errorKy = max(max(abs(Ky-exactKy))),
errorJz = max(max(abs(Jz-exactJz))),

figure(1)
quiver(x,y,Hx,Hy);
title('numerical magnetic field');
tri = delaunay(x,y);
figure(2)
quiver(x,y,exactHx,exactHy);
title('analytic magnetic field');

figure(3)
trisurf(tri,x,y,Ez);
title('Numerical electric field');
figure(4)
trisurf(tri,x,y,Ez-exactEz);
title('Pointwise error of electric field');

figure(5)
trisurf(tri,x,y,Jz);
title('Numerical induced electric current');
figure(6)
trisurf(tri,x,y,Jz-exactJz);
title('Pointwise error of induced electric current');
```

Of course, to solve our example, we need three supporting MATLAB functions *fun_f21.m, fun_gx21.m, fun_gy21.m* to evaluate functions f, g_x and g_y, respectively. Also we need a mesh generator function *genrecxygrid.m*.

The code *fun_f21.m* is shown below:

```
function val=fun_f21(x,y,t,omepi)

val =(t-1-2*omepi)*exp(-t)*sin(omepi*x).*sin(omepi*y);
```

The code *fun_gx21.m* is shown below:

```
function val=fun_gx21(x,y,t,omepi)

val =(omepi+t-1)*exp(-t)*sin(omepi*x).*cos(omepi*y);
```

The code *fun_gy21.m* is shown below:

```
function val=fun_gy21(x,y,t,omepi)

val =(1-omepi-t)*exp(-t)*cos(omepi*x).*sin(omepi*y);
```

The code *genrecxygrid.m* is shown below:

```
% generate a square grid of points on the xy-plane
% Inputs:
%    Domain [xlow,xhigh]x[ylow,yhigh]
%    xn, yn: number of points in the x- and y-directions.

function [xy] = genrecxygrid(xlow,xhigh,ylow,yhigh,xn,yn)
```

Table 4.1 The L^∞ errors with $\tau = 10^{-6}$, $Nr = 1$ at 100 time step

Errors	$h = 1/10$	$h = 1/20$	$h = 1/40$	$h = 1/80$	$h = 1/160$
Hx	0.0013	7.3433e−004	3.8435e−004	1.9291e−004	9.5417e−005
Hy	0.0013	7.3433e−004	3.8435e−004	1.9291e−004	9.5417e−005
Ez	0.0021	0.0011	5.7637e−004	2.8954e−004	1.4261e−004
Kx	6.4776e−008	3.6787e−008	1.9283e−008	9.7061e−009	4.8216e−009
Ky	6.4776e−008	3.6787e−008	1.9283e−008	9.7061e−009	4.8216e−009
Jz	1.0302e−007	5.4335e−008	2.8892e−008	1.4562e−008	7.2149e−009

Table 4.2 The L^∞ errors with $\tau = 10^{-6}$, $Nr = 2$ at 100 time step

Errors	$h = 1/5$	$h = 1/10$	$h = 1/20$	$h = 1/40$	$h = 1/80$
Hx	0.0018	5.9136e−004	1.5830e−004	4.0508e−005	1.0099e−005
Hy	0.0018	5.9136e−004	1.5830e−004	4.0508e−005	1.0100e−005
Ez	0.0022	7.1938e−004	1.9800e−004	5.0975e−005	1.2856e−005
Kx	8.9076e−008	2.9593e−008	7.9259e−009	2.0324e−009	5.0834e−010
Ky	8.9076e−008	2.9593e−008	7.9259e−009	2.0324e−009	5.0834e−010
Jz	1.1244e−007	3.6006e−008	9.9139e−009	2.5500e−009	6.4321e−010

```
xorig=[linspace(xlow,xhigh,xn),linspace(ylow,yhigh,yn)];
n  = xn*yn;
xy = zeros(n,2);     % x,y coordinates of all points
pt = 1;
for j = 1:yn
   for i = 1:xn
      xy(pt,:)=[xorig(i)  xorig(xn+j)];
      pt=pt+1;
   end
end

return
```

With these MATLAB functions, we can solve this example on uniformly refined meshes with various time step sizes τ and different orders Nr of polynomial basis functions. Exemplary results are shown in Tables 4.1 and 4.2, which justify the following convergence result:

$$\max_{m\geq 1}(||\mathbf{H}^m - \mathbf{H}_h^m||_{L^\infty(\Omega)} + ||\mathbf{E}^m - \mathbf{E}_h^m||_{L^\infty(\Omega)} + ||\mathbf{J}^m - \mathbf{J}_h^m||_{L^\infty(\Omega)} + ||\mathbf{K}^m - \mathbf{K}_h^m||_{L^\infty(\Omega)}) \leq C h^{Nr}.$$

Exemplary solutions for E_z and the corresponding pointwise errors obtained with $Nr = 2, \tau = 10^{-6}$ at 100 time steps are presented in Fig. 4.1. More numerical results using the package *nudg* of [141] can be found in Li [185].

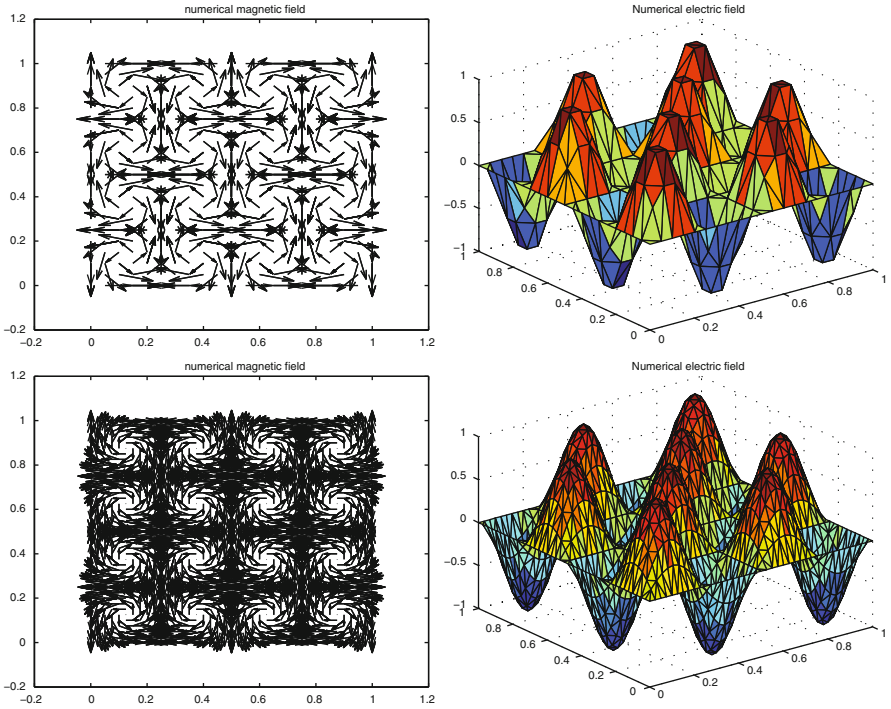

Fig. 4.1 Results obtained with $Nr = 2, \tau = 10^{-6}$ at 100 time steps. *Top row* (with $h = 1/10$): magnetic field H (*Left*) and electric field E (*Right*); *bottom row* (with $h = 1/20$): magnetic field H (*Left*) and electric field E (*Right*)

Chapter 5
Superconvergence Analysis for Metamaterials

In this chapter, we first give a quick review of superconvergence analysis in Sect. 5.1. Then we carry out the superclose analysis for 3-D metamaterial Maxwell's equations represented by the Drude model. The analysis for a semi-discrete scheme is presented in Sect. 5.2, which is followed by the analysis for two fully-discrete schemes in Sect. 5.3. In Sect. 5.4, a superconvergence result in the discrete l_2 norm is proved. Finally, the superconvergence analysis is extended to the 2-D case in Sect. 5.5.

5.1 A Brief Overview of Superconvergence Analysis

In finite element methods, when the underlying differential equations have smooth solutions and the differential equations are solved on very structured meshes such as rectangular grids or strongly regular triangular grids, we often see that the obtained convergence rates have higher order than the theoretical approximation results suggested. Such a phenomenon is called superconvergence. Study of the superconvergence phenomenon started in the early 1970s, and many interesting results have been obtained for problems described by elliptic equations [22,23,128], parabolic equations [292], the second-order wave equations, and porous media flows [113]. Detailed superconvergence analysis can be found in classic books [67, 201, 289]. A detailed bibliography on superconvergence by 1996 can be found in a review paper by Krizek and Neittaanmaki [170].

Compared to those widely studied equations, there are not many superconvergence results existing for Maxwell's equations. In 1994, Monk [215] obtained the first superconvergence result for Maxwell's equations in vacuum. Later, Brandts [50] presented another superconvergence analysis for 2-D Maxwell's equations in vacuum. Also Lin and his collaborators [199, 200, 202] systematically obtained many global superconvergence results using the so-called Lin's Integral Identity technique [203, 204, 308] developed in the early 1990s. More details on Lin's Integral Identity technique can be found in books [201, 297]. In 2008, Lin

J. Li and Y. Huang, *Time-Domain Finite Element Methods for Maxwell's Equations in Metamaterials*, Springer Series in Computational Mathematics 43,
DOI 10.1007/978-3-642-33789-5_5, © Springer-Verlag Berlin Heidelberg 2013

and Li [198] extended the superconvergence result for vacuum to three popular dispersive media models. Some superconvergence work has been recently carried out for metamaterial models [153, 156]. In this chapter, we will present detailed superconvergence analysis for both semi-discrete and fully-discrete schemes on cubic and rectangular meshes.

5.2 Superclose Analysis for a Semi-discrete Scheme

To simplify the presentation, we consider the non-dimensionalized Drude model equations

$$\frac{\partial \mathbf{E}}{\partial t} = \nabla \times \mathbf{H} - \mathbf{J}, \tag{5.1}$$

$$\frac{\partial \mathbf{H}}{\partial t} = -\nabla \times \mathbf{E} - \mathbf{K}, \tag{5.2}$$

$$\frac{\partial \mathbf{J}}{\partial t} + \Gamma_e \mathbf{J} = \omega_e^2 \mathbf{E}, \tag{5.3}$$

$$\frac{\partial \mathbf{K}}{\partial t} + \Gamma_m \mathbf{K} = \omega_m^2 \mathbf{H}, \tag{5.4}$$

subject to the perfect conducting boundary condition (3.59) and the initial conditions (3.60) and (3.61). Derivation of (5.1)–(5.4) can be found in Sect. 4.4. Here for clarity, all tildes are dropped.

For superconvergence analysis, we assume that the domain Ω is a rectangular cuboid, which is partitioned by a family of regular cubic meshes T_h with maximum mesh size h. Recall that the Raviart-Thomas-Nédélec cubic elements (the pair of divergence and curl conforming elements) are defined as (cf. Chap. 3):

$$\mathbf{U}_h = \{\psi_h \in H(\mathrm{div}; \Omega) : \ \psi_h|_K \in Q_{k,k-1,k-1} \times Q_{k-1,k,k-1} \times Q_{k-1,k-1,k}, \ \forall\, K \in T_h\},$$

$$\mathbf{V}_h = \{\phi_h \in H(\mathrm{curl}; \Omega) : \ \phi_h|_K \in Q_{k-1,k,k} \times Q_{k,k-1,k} \times Q_{k,k,k-1}, \ \forall\, K \in T_h\}.$$

Furthermore, we need the so-called Nédélec interpolation operator Π_h, which has been defined in Chap. 3.

The superclose analysis depends on the following two fundamental results.

Lemma 5.1 ([202, Lemma 3.1]). *On any cubic element K, for any $\mathbf{E} \in (H^{k+2}(K))^3$, we have*

$$\int_K \nabla \times (\mathbf{E} - \Pi_h \mathbf{E}) \cdot \psi_h dx dy dz = O(h^{k+1}) ||\mathbf{E}||_{k+2,K} ||\psi_h||_{0,K}, \quad \forall\, \psi_h|_K \in \mathbf{U}_h(K).$$

Lemma 5.2 ([202, Lemma 3.2]). *On any cubic element K, for any $\mathbf{E} \in (H^{k+1}(K))^3$, we have*

$$\int_K (\mathbf{E} - \Pi_h \mathbf{E}) \cdot \phi_h dx dy dz = O(h^{k+1}) ||\mathbf{E}||_{k+1,K} ||\phi_h||_{0,K}, \quad \forall\, \phi_h|_K \in \mathbf{V}_h(K).$$

Though Lemmas 5.1 and 5.2 were stated for the whole domain Ω in [202], the proofs of [202] actually show that the results hold true element-wisely.

A corresponding weak formulation for the system (5.1)–(5.4) is: For any $t \in (0, T]$, find the solutions $\mathbf{E} \in H_0(\text{curl}; \Omega)$, $\mathbf{J} \in H(\text{curl}; \Omega)$, $\mathbf{H}$ and $\mathbf{K} \in (L^2(\Omega))^3$ such that

$$(\mathbf{E}_t, \phi) - (\mathbf{H}, \nabla \times \phi) + (\mathbf{J}, \phi) = 0, \quad \forall\, \phi \in H_0(\text{curl}; \Omega), \tag{5.5}$$

$$(\mathbf{H}_t, \psi) + (\nabla \times \mathbf{E}, \psi) + (\mathbf{K}, \psi) = 0, \quad \forall\, \psi \in (L^2(\Omega))^3, \tag{5.6}$$

$$(\mathbf{J}_t, \tilde{\phi}) + \Gamma_e(\mathbf{J}, \tilde{\phi}) - \omega_e^2(\mathbf{E}, \tilde{\phi}) = 0, \quad \forall\, \tilde{\phi} \in H(\text{curl}; \Omega), \tag{5.7}$$

$$(\mathbf{K}_t, \tilde{\psi}) + \Gamma_m(\mathbf{K}, \tilde{\psi}) - \omega_m^2(\mathbf{H}, \tilde{\psi}) = 0, \quad \forall\, \tilde{\psi} \in (L^2(\Omega))^3, \tag{5.8}$$

subject to the initial conditions (3.60) and (3.61), i.e.,

$$\mathbf{E}(\mathbf{x}, 0) = \mathbf{E}_0(\mathbf{x}), \quad \mathbf{H}(\mathbf{x}, 0) = \mathbf{H}_0(\mathbf{x}),$$
$$\mathbf{J}(\mathbf{x}, 0) = \mathbf{J}_0(\mathbf{x}), \quad \mathbf{K}(\mathbf{x}, 0) = \mathbf{K}_0(\mathbf{x}).$$

Now a semi-discrete mixed method can be constructed for solving (5.5)–(5.8): For any $t \in (0, T]$, find the solutions $\mathbf{E}^h \in \mathbf{V}_h^0, \mathbf{J}^h \in \mathbf{V}_h, \mathbf{H}^h, \mathbf{K}^h \in \mathbf{U}_h$ such that

$$(\mathbf{E}_t^h, \phi_h) - (\mathbf{H}^h, \nabla \times \phi_h) + (\mathbf{J}^h, \phi_h) = 0, \quad \forall\, \phi_h \in \mathbf{V}_h^0, \tag{5.9}$$

$$(\mathbf{H}_t^h, \psi_h) + (\nabla \times \mathbf{E}^h, \psi_h) + (\mathbf{K}^h, \psi_h) = 0, \quad \forall\, \psi_h \in \mathbf{U}_h, \tag{5.10}$$

$$(\mathbf{J}_t^h, \tilde{\phi}_h) + \Gamma_e(\mathbf{J}^h, \tilde{\phi}_h) - \omega_e^2(\mathbf{E}^h, \tilde{\phi}_h) = 0, \quad \forall\, \tilde{\phi}_h \in \mathbf{V}_h, \tag{5.11}$$

$$(\mathbf{K}_t^h, \tilde{\psi}_h) + \Gamma_m(\mathbf{K}^h, \tilde{\psi}_h) - \omega_m^2(\mathbf{H}^h, \tilde{\psi}_h) = 0, \quad \forall\, \tilde{\psi}_h \in \mathbf{U}_h, \tag{5.12}$$

with the initial approximations

$$\mathbf{E}_h^0(\mathbf{x}) = \Pi_h \mathbf{E}_0(\mathbf{x}), \quad \mathbf{H}_h^0(\mathbf{x}) = P_h \mathbf{H}_0(\mathbf{x}), \tag{5.13}$$

$$\mathbf{J}_h^0(\mathbf{x}) = \Pi_h \mathbf{J}_0(\mathbf{x}), \quad \mathbf{K}_h^0(\mathbf{x}) = P_h \mathbf{K}_0(\mathbf{x}). \tag{5.14}$$

Recall that P_h denotes the standard L^2 projection operator onto space $\mathbf{U}_h$, and $\mathbf{V}_h^0 = \{\mathbf{v}_h \in \mathbf{V}_h : \mathbf{v}_h \times \mathbf{n} = \mathbf{0} \text{ on } \partial\Omega\}$.

For this scheme, we have the following superclose result.

Theorem 5.1. *Let* $(\mathbf{E}, \mathbf{H}, \mathbf{J}, \mathbf{K})$ *and* $(\mathbf{E}^h, \mathbf{H}^h, \mathbf{J}^h, \mathbf{K}^h)$ *be the analytic and finite element solutions of (5.5)–(5.8) and (5.9)–(5.12) at time* $t \in (0, T]$, *respectively. Under the regularity assumptions*

$$\mathbf{E}_t, \mathbf{J}_t, \mathbf{J} \in L^\infty(0, T; (H^{k+1}(\Omega))^3), \quad \mathbf{E} \in L^\infty(0, T; (H^{k+2}(\Omega))^3),$$

there exists a constant $C > 0$ *independent of* h *but linearly dependent on* T *such that*

$$
\begin{aligned}
&||\Pi_h\mathbf{E}-\mathbf{E}^h||_{L^\infty(0,T;(L^2(\Omega))^3)}+||P_h\mathbf{H}-\mathbf{H}^h||_{L^\infty(0,T;(L^2(\Omega))^3)}\\
&+\frac{1}{\omega_e}||\Pi_h\mathbf{J}-\mathbf{J}^h||_{L^\infty(0,T;(L^2(\Omega))^3)}+\frac{1}{\omega_m}||P_h\mathbf{K}-\mathbf{K}^h||_{L^\infty(0,T;(L^2(\Omega))^3)}\\
\le\ & Ch^{k+1}(||\mathbf{E}_t||_{L^\infty(0,T;(H^{k+1}(\Omega))^3)}+||\mathbf{J}||_{L^\infty(0,T;(H^{k+1}(\Omega))^3)}\\
&+||\mathbf{E}||_{L^\infty(0,T;(H^{k+2}(\Omega))^3)}+||\mathbf{J}_t||_{L^\infty(0,T;(H^{k+1}(\Omega))^3)}),
\end{aligned}
$$

where $k \ge 1$ is the order of the basis functions in spaces $\mathbf{U}_h$ and $\mathbf{V}_h$.

Proof. Denote $\xi = \Pi_h\mathbf{E}-\mathbf{E}^h, \eta = P_h\mathbf{H}-\mathbf{H}^h, \tilde{\xi} = \Pi_h\mathbf{J}-\mathbf{J}^h, \tilde{\eta} = P_h\mathbf{K}-\mathbf{K}^h$. Choosing $\phi = \phi_h = \xi$ in (5.5) and (5.9), $\psi = \psi_h = \eta$ in (5.6) and (5.10), $\tilde{\phi} = \tilde{\phi}_h = \tilde{\xi}$ in (5.7) and (5.11), $\tilde{\psi} = \tilde{\psi}_h = \tilde{\eta}$ in (5.8) and (5.12), respectively, and rearranging the resultants, we obtain the error equations

$$
\begin{aligned}
(i)\quad & (\xi_t,\xi)-(\eta,\nabla\times\xi)+(\tilde{\xi},\xi)\\
&=((\Pi_h\mathbf{E}-\mathbf{E})_t,\xi)-(P_h\mathbf{H}-\mathbf{H},\nabla\times\xi)+(\Pi_h\mathbf{J}-\mathbf{J},\xi),\\
(ii)\quad & (\eta_t,\eta)+(\nabla\times\xi,\eta)+(\tilde{\eta},\eta)\\
&=((P_h\mathbf{H}-\mathbf{H})_t,\eta)+(\nabla\times(\Pi_h\mathbf{E}-\mathbf{E}),\eta)+(P_h\mathbf{K}-\mathbf{K},\eta),\\
(iii)\quad & (\tilde{\xi}_t,\tilde{\xi})+\Gamma_e(\tilde{\xi},\tilde{\xi})-\omega_e^2(\xi,\tilde{\xi})\\
&=((\Pi_h\mathbf{J}-\mathbf{J})_t,\tilde{\xi})+\Gamma_e(\Pi_h\mathbf{J}-\mathbf{J},\tilde{\xi})-\omega_e^2(\Pi_h\mathbf{E}-\mathbf{E},\tilde{\xi}),\\
(iv)\quad & (\tilde{\eta}_t,\tilde{\eta})+\Gamma_m(\tilde{\eta},\tilde{\eta})-\omega_m^2(\eta,\tilde{\eta})\\
&=((P_h\mathbf{K}-\mathbf{K})_t,\tilde{\eta})+\Gamma_m(P_h\mathbf{K}-\mathbf{K},\tilde{\eta})-\omega_m^2(P_h\mathbf{H}-\mathbf{H},\tilde{\eta}).
\end{aligned}
$$

Dividing the last two equations by ω_e^2 and ω_m^2, respectively, then adding the above four equations together, we obtain

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}(||\xi||_0^2+||\eta||_0^2+\frac{1}{\omega_e^2}||\tilde{\xi}||_0^2+\frac{1}{\omega_m^2}||\tilde{\eta}||_0^2)+\frac{\Gamma_e}{\omega_e^2}||\tilde{\xi}||_0^2+\frac{\Gamma_m}{\omega_m^2}||\tilde{\eta}||_0^2\\
=\ & ((\Pi_h\mathbf{E}-\mathbf{E})_t,\xi)+(\Pi_h\mathbf{J}-\mathbf{J},\xi)+(\nabla\times(\Pi_h\mathbf{E}-\mathbf{E}),\eta)\\
&+\frac{1}{\omega_e^2}((\Pi_h\mathbf{J}-\mathbf{J})_t,\tilde{\xi})+\frac{\Gamma_e}{\omega_e^2}(\Pi_h\mathbf{J}-\mathbf{J},\tilde{\xi})-(\Pi_h\mathbf{E}-\mathbf{E},\tilde{\xi}),
\end{aligned}
\tag{5.15}
$$

where we used the L^2-projection property in the above derivation.

Using Lemmas 5.1 and 5.2 and the Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
&((\Pi_h\mathbf{E}-\mathbf{E})_t,\xi)\le Ch^{k+1}||\mathbf{E}_t||_{L^\infty(0,T;\mathbf{H}^{k+1}(\Omega))}||\xi||_{L^\infty(0,T;\mathbf{L}^2(\Omega))},\\
&(\Pi_h\mathbf{J}-\mathbf{J},\xi)\le Ch^{k+1}||\mathbf{J}||_{L^\infty(0,T;\mathbf{H}^{k+1}(\Omega))}||\xi||_{L^\infty(0,T;\mathbf{L}^2(\Omega))},\\
&(\nabla\times(\Pi_h\mathbf{E}-\mathbf{E}),\eta)\le Ch^{k+1}||\mathbf{E}||_{L^\infty(0,T;\mathbf{H}^{k+2}(\Omega))}||\eta||_{L^\infty(0,T;\mathbf{L}^2(\Omega))},
\end{aligned}
$$

$$\frac{1}{\omega_e^2}((\Pi_h\mathbf{J}-\mathbf{J})_t,\tilde{\xi}) \le \frac{1}{\omega_e^2}\cdot Ch^{k+1}||\mathbf{J}_t||_{L^\infty(0,T;\mathbf{H}^{k+1}(\Omega))}||\tilde{\xi}||_{L^\infty(0,T;\mathbf{L}^2(\Omega))},$$

$$\frac{\Gamma_e}{\omega_e^2}(\Pi_h\mathbf{J}-\mathbf{J},\tilde{\xi}) \le \frac{\Gamma_e}{\omega_e^2}\cdot Ch^{k+1}||\mathbf{J}||_{L^\infty(0,T;\mathbf{H}^{k+1}(\Omega))}||\tilde{\xi}||_{L^\infty(0,T;\mathbf{L}^2(\Omega))},$$

$$(\Pi_h\mathbf{E}-\mathbf{E},\tilde{\xi}) \le Ch^{k+1}||\mathbf{E}||_{L^\infty(0,T;\mathbf{H}^{k+1}(\Omega))}||\tilde{\xi}||_{L^\infty(0,T;\mathbf{L}^2(\Omega))}.$$

Substituting the above estimates into (5.15), and using Gronwall inequality, we obtain

$$\begin{aligned}
&||\xi||^2_{L^\infty(0,T;\mathbf{L}^2(\Omega))} + ||\eta||^2_{L^\infty(0,T;\mathbf{L}^2(\Omega))} + \frac{1}{\omega_e^2}||\tilde{\xi}||^2_{L^\infty(0,T;\mathbf{L}^2(\Omega))} + \frac{1}{\omega_m^2}||\tilde{\eta}||^2_{L^\infty(0,T;\mathbf{L}^2(\Omega))}\\
\le\ & Ch^{2(k+1)}(||\mathbf{E}_t||^2_{L^\infty(0,T;\mathbf{H}^{k+1}(\Omega))} + ||\mathbf{J}||^2_{L^\infty(0,T;\mathbf{H}^{k+1}(\Omega))}\\
&+||\mathbf{E}||^2_{L^\infty(0,T;\mathbf{H}^{k+2}(\Omega))} + ||\mathbf{J}_t||^2_{L^\infty(0,T;\mathbf{H}^{k+1}(\Omega))}),
\end{aligned}$$

where the constant C is linearly dependent on T^2. The proof is completed by using the triangle inequality, and the interpolation error (3.78) and projection error (3.79). □

5.3 Superclose Analysis for Fully-Discrete Schemes

Now we can formulate the Crank-Nicolson mixed finite element scheme for solving (5.5)–(5.8): For $m = 1, 2, \cdots, M$, find $\mathbf{E}_h^m \in \mathbf{V}_h^0, \mathbf{J}_h^m \in \mathbf{V}_h, \mathbf{H}_h^m, \mathbf{K}_h^m \in \mathbf{U}_h$ such that

$$(\delta_\tau\mathbf{E}_h^m,\phi_h) - (\overline{\mathbf{H}}_h^m,\nabla\times\phi_h) + (\overline{\mathbf{J}}_h^m,\phi_h) = 0, \quad \forall\ \phi_h \in \mathbf{V}_h^0, \tag{5.16}$$

$$(\delta_\tau\mathbf{H}_h^m,\psi_h) + (\nabla\times\overline{\mathbf{E}}_h^m,\psi_h) + (\overline{\mathbf{K}}_h^m,\psi_h) = 0, \quad \forall\ \psi_h \in \mathbf{U}_h, \tag{5.17}$$

$$(\delta_\tau\mathbf{J}_h^m,\tilde{\phi}_h) + \Gamma_e(\overline{\mathbf{J}}_h^m,\tilde{\phi}_h) - \omega_e^2(\overline{\mathbf{E}}_h^m,\tilde{\phi}_h) = 0, \quad \forall\ \tilde{\phi}_h \in \mathbf{V}_h, \tag{5.18}$$

$$(\delta_\tau\mathbf{K}_h^m,\tilde{\psi}_h) + \Gamma_m(\overline{\mathbf{K}}_h^m,\tilde{\psi}_h) - \omega_m^2(\overline{\mathbf{H}}_h^m,\tilde{\psi}_h) = 0, \quad \forall\ \tilde{\psi}_h \in \mathbf{U}_h, \tag{5.19}$$

subject to the initial approximations (5.13) and (5.14). As before, we denote

$$\delta_\tau\mathbf{E}_h^m = (\mathbf{E}_h^m - \mathbf{E}_h^{m-1})/\tau, \quad \overline{\mathbf{H}}_h^m = \frac{1}{2}(\mathbf{H}_h^m + \mathbf{H}_h^{m-1}).$$

For this fully-discrete scheme, we have the following superclose result.

Theorem 5.2. *Let* $(\mathbf{E}^m,\mathbf{H}^m,\mathbf{J}^m,\mathbf{K}^m)$ *and* $(\mathbf{E}_h^m,\mathbf{H}_h^m,\mathbf{J}_h^m,\mathbf{K}_h^m)$ *be the analytic and finite element solutions of (5.5)–(5.8) and (5.16)–(5.19) at time* t_m*, respectively. Under the regularity assumptions*

$$\mathbf{E}_t, \mathbf{J}_t, \mathbf{J} \in L^\infty(0,T;(H^{k+1}(\Omega))^3), \quad \mathbf{E} \in L^\infty(0,T;(H^{k+2}(\Omega))^3),$$
$$\mathbf{E}_{tt}, \mathbf{H}_{tt}, \mathbf{J}_{tt}, \mathbf{K}_{tt}, \nabla\times\mathbf{E}_{tt}, \nabla\times\mathbf{H}_{tt} \in L^\infty(0,T;(L^2(\Omega))^3),$$

there exists a constant $C > 0$, independent of h but linearly dependent on T such that

$$\begin{aligned}
&\max_{1\le m\le M}(||\Pi_h\mathbf{E}^m - \mathbf{E}_h^m||_0 + ||P_h\mathbf{H}^m - \mathbf{H}_h^m||_0 + ||\Pi_h\mathbf{J}^m - \mathbf{J}_h^m||_0 + ||P_h\mathbf{K}^m - \mathbf{K}_h^m||_0)\\
&\le Ch^{k+1}(||\mathbf{E}_t||_{L^\infty(0,T;(H^{k+1}(\Omega))^3)} + ||\mathbf{J}||_{L^\infty(0,T;(H^{k+1}(\Omega))^3)}\\
&\quad + ||\mathbf{E}||_{L^\infty(0,T;(H^{k+2}(\Omega))^3)} + ||\mathbf{J}_t||_{L^\infty(0,T;(H^{k+1}(\Omega))^3)})\\
&\quad + C\tau^2(||\nabla\times\mathbf{H}_{tt}||_{L^\infty(0,T;(L^2(\Omega))^3)} + ||\mathbf{J}_{tt}||_{L^\infty(0,T;(L^2(\Omega))^3)} + ||\nabla\times\mathbf{E}_{tt}||_{L^\infty(0,T;(L^2(\Omega))^3)}\\
&\quad + ||\mathbf{K}_{tt}||_{L^\infty(0,T;(L^2(\Omega))^3)} + ||\mathbf{E}_{tt}||_{L^\infty(0,T;(L^2(\Omega))^3)} + ||\mathbf{H}_{tt}||_{L^\infty(0,T;(L^2(\Omega))^3)}),
\end{aligned}$$

where $k \ge 1$ is the order of the basis functions in spaces $\mathbf{U}_h$ and $\mathbf{V}_h$.

Proof. Integrating (5.5)–(5.8) in time over $I_m = [t_{m-1}, t_m]$ and dividing all by τ, we have

$$(\delta_\tau\mathbf{E}^m, \phi) - (\frac{1}{\tau}\int_{I_m}\mathbf{H}(s)ds, \nabla\times\phi) + (\frac{1}{\tau}\int_{I_m}\mathbf{J}(s)ds, \phi) = 0, \tag{5.20}$$

$$(\delta_\tau\mathbf{H}^m, \psi) + (\nabla\times\frac{1}{\tau}\int_{I_m}\mathbf{E}(s)ds, \psi) + (\frac{1}{\tau}\int_{I_m}\mathbf{K}(s)ds, \psi) = 0, \tag{5.21}$$

$$(\delta_\tau\mathbf{J}^m, \tilde{\phi}) + \Gamma_e(\frac{1}{\tau}\int_{I_m}\mathbf{J}(s)ds, \tilde{\phi}) - \omega_e^2(\frac{1}{\tau}\int_{I_m}\mathbf{E}(s)ds, \tilde{\phi}) = 0, \tag{5.22}$$

$$(\delta_\tau\mathbf{K}^m, \tilde{\psi}) + \Gamma_m(\frac{1}{\tau}\int_{I_m}\mathbf{K}(s)ds, \tilde{\psi}) - \omega_m^2(\frac{1}{\tau}\int_{I_m}\mathbf{H}(s)ds, \tilde{\psi}) = 0. \tag{5.23}$$

Denote $\xi_h^m = \Pi_h\mathbf{E}^m - \mathbf{E}_h^m$, $\eta_h^m = P_h\mathbf{H}^m - \mathbf{H}_h^m$, $\tilde{\xi}_h^m = \Pi_h\mathbf{J}^m - \mathbf{J}_h^m$, $\tilde{\eta}_h^m = P_h\mathbf{K}^m - \mathbf{K}_h^m$. Subtracting (5.16)–(5.19) from (5.20)–(5.23) with $\phi = \phi_h$, $\psi = \psi_h$, $\tilde{\phi} = \tilde{\phi}_h$, and $\tilde{\psi} = \tilde{\psi}_h$, we can obtain the error equations

$$\begin{aligned}
(i)\quad &(\delta_\tau\xi_h^m, \phi_h) - (\overline{\eta}_h^m, \nabla\times\phi_h) + (\overline{\tilde{\xi}}_h^m, \phi_h) = (\delta_\tau(\Pi_h\mathbf{E}^m - \mathbf{E}^m), \phi_h)\\
&-(P_h\overline{\mathbf{H}}^m - \frac{1}{\tau}\int_{I_m}\mathbf{H}(s)ds, \nabla\times\phi_h) + (\Pi_h\overline{\mathbf{J}}^m - \frac{1}{\tau}\int_{I_m}\mathbf{J}(s)ds, \phi_h),\\
(ii)\quad &(\delta_\tau\eta_h^m, \psi_h) + (\nabla\times\overline{\tilde{\xi}}_h^m, \psi_h) + (\overline{\tilde{\eta}}_h^m, \psi_h) = (\delta_\tau(P_h\mathbf{H}^m - \mathbf{H}^m), \psi_h)\\
&+(\nabla\times(\Pi_h\overline{\mathbf{E}}^m - \frac{1}{\tau}\int_{I_m}\mathbf{E}(s)ds), \psi_h) + (P_h\overline{\mathbf{K}}_h^m - \frac{1}{\tau}\int_{I_m}\mathbf{K}(s)ds, \psi_h),
\end{aligned}$$

$$
\begin{aligned}
(iii)\ \ &(\delta_\tau\tilde{\xi}_h^m,\tilde{\phi}_h)+\Gamma_e(\overline{\tilde{\xi}}_h^m,\tilde{\phi}_h)-\omega_e^2(\overline{\tilde{\xi}}_h^m,\tilde{\phi}_h)=(\delta_\tau(\Pi_h\mathbf{J}^m-\mathbf{J}^m),\tilde{\phi}_h)\\
&+\Gamma_e(\Pi_h\overline{\mathbf{J}}^m-\frac{1}{\tau}\int_{I_m}\mathbf{J}(s)ds,\tilde{\phi}_h)-\omega_e^2(\Pi_h\overline{\mathbf{E}}_h^m-\frac{1}{\tau}\int_{I_m}\mathbf{E}(s)ds,\tilde{\phi}_h),\\
(iv)\ \ &(\delta_\tau\tilde{\eta}_h^m,\tilde{\psi}_h)+\Gamma_m(\overline{\tilde{\eta}}_h^m,\tilde{\psi}_h)-\omega_m^2(\overline{\tilde{\eta}}_h^m,\tilde{\psi}_h)=(\delta_\tau(P_h\mathbf{K}^m-\mathbf{K}^m),\tilde{\psi}_h)\\
&+\Gamma_m(P_h\overline{\mathbf{K}}^m-\frac{1}{\tau}\int_{I_m}\mathbf{K}(s)ds,\tilde{\psi}_h)-\omega_m^2(P_h\overline{\mathbf{H}}_h^m-\frac{1}{\tau}\int_{I_m}\mathbf{H}(s)ds,\tilde{\psi}_h).
\end{aligned}
$$

Choosing $\phi_h=\tau\overline{\xi}_h^m,\psi_h=\tau\overline{\eta}_h^m,\tilde{\phi}_h=\tau\overline{\tilde{\xi}}_h^m,\tilde{\psi}_h=\tau\overline{\tilde{\eta}}_h^m$ in the above error equations, dividing the last two equations by ω_e^2 and ω_m^2, adding the resultants together, and using the property of operator P_h, we obtain

$$
\begin{aligned}
&\frac{1}{2}[||\xi_h^m||_0^2-||\xi_h^{m-1}||_0^2+||\eta_h^m||_0^2-||\eta_h^{m-1}||_0^2\\
&+\frac{1}{\omega_e^2}(||\tilde{\xi}_h^m||_0^2-||\tilde{\xi}_h^{m-1}||_0^2)+\frac{1}{\omega_m^2}(||\tilde{\eta}_h^m||_0^2-||\tilde{\eta}_h^{m-1}||_0^2)]\\
\le\ &\tau(\delta_\tau(\Pi_h\mathbf{E}^m-\mathbf{E}^m),\overline{\tilde{\xi}}_h^m)-\tau(\overline{\mathbf{H}}^m-\frac{1}{\tau}\int_{I_m}\mathbf{H}(s)ds,\nabla\times\overline{\tilde{\xi}}_h^m)\\
&+\tau(\Pi_h\overline{\mathbf{J}}^m-\frac{1}{\tau}\int_{I_m}\mathbf{J}(s)ds,\overline{\tilde{\xi}}_h^m)+\tau(\nabla\times(\Pi_h\overline{\mathbf{E}}^m-\frac{1}{\tau}\int_{I_m}\mathbf{E}(s)ds),\overline{\eta}_h^m)\\
&+\tau(\overline{\mathbf{K}}_h^m-\frac{1}{\tau}\int_{I_m}\mathbf{K}(s)ds,\overline{\eta}_h^m)+\frac{\tau}{\omega_e^2}(\delta_\tau(\Pi_h\mathbf{J}^m-\mathbf{J}^m),\overline{\tilde{\xi}}_h^m)\\
&+\frac{\tau\Gamma_e}{\omega_e^2}(\Pi_h\overline{\mathbf{J}}^m-\frac{1}{\tau}\int_{I_m}\mathbf{J}(s)ds,\overline{\tilde{\xi}}_h^m)-\tau(\Pi_h\overline{\mathbf{E}}_h^m-\frac{1}{\tau}\int_{I_m}\mathbf{E}(s)ds,\overline{\tilde{\xi}}_h^m)\\
&+\frac{\tau\Gamma_m}{\omega_m^2}(\overline{\mathbf{K}}^m-\frac{1}{\tau}\int_{I_m}\mathbf{K}(s)ds,\overline{\tilde{\eta}}_h^m)-\tau(\overline{\mathbf{H}}_h^m-\frac{1}{\tau}\int_{I_m}\mathbf{H}(s)ds,\overline{\tilde{\eta}}_h^m)\\
=\ &\sum_{i=1}^{10}Err_i.
\end{aligned}
\tag{5.24}
$$

After careful estimating of $Err_i, i=1,\cdots,10$ (details can be found in the original paper [153]), and substituting them into (5.24), summing up the result from $m=1$ to any $n\le M$ with the fact that

$$\xi_h^0=\eta_h^0=\tilde{\xi}_h^0=\tilde{\eta}_h^0=0,$$

then using the arithmetic-geometric mean inequality and taking the maximum with respect to n, we obtain

$$
\begin{aligned}
&||\xi_h||^2_{l^\infty(L^2)} + ||\eta_h||^2_{l^\infty(L^2)} + ||\tilde{\xi}_h||^2_{l^\infty(L^2)} + ||\tilde{\eta}_h||^2_{l^\infty(L^2)} \\
\le\ & Ch^{2(k+1)}(||\mathbf{E}_t||^2_{L^\infty(0,T;(H^{k+1}(\Omega))^3)} + ||\mathbf{J}||^2_{L^\infty(0,T;(H^{k+1}(\Omega))^3)} \\
&+||\mathbf{E}||^2_{L^\infty(0,T;(H^{k+2}(\Omega))^3)} + ||\mathbf{J}_t||^2_{L^\infty(0,T;(H^{k+1}(\Omega))^3)}) \\
&+C\tau^4(||\nabla\times\mathbf{H}_{tt}||^2_{L^\infty(0,T;(L^2(\Omega))^3)} + ||\mathbf{J}_{tt}||^2_{L^\infty(0,T;(L^2(\Omega))^3)} + ||\nabla\times\mathbf{E}_{tt}||^2_{L^\infty(0,T;(L^2(\Omega))^3)} \\
&+||\mathbf{K}_{tt}||^2_{L^\infty(0,T;(L^2(\Omega))^3)} + ||\mathbf{E}_{tt}||^2_{L^\infty(0,T;(L^2(\Omega))^3)} + ||\mathbf{H}_{tt}||^2_{L^\infty(0,T;(L^2(\Omega))^3)}),
\end{aligned}
$$

which concludes the proof. Note that C linearly depends on T^2. □

Remark 5.1. Similarly, we can formulate a leap-frog mixed finite element scheme for solving (5.5)–(5.8): Given initial approximations $\mathbf{E}_h^0, \mathbf{K}_h^0, \mathbf{H}_h^{\frac{1}{2}}, \mathbf{J}_h^{\frac{1}{2}}$, for $m \ge 1$, find $\mathbf{E}_h^m \in \mathbf{V}_h^0, \mathbf{J}_h^{m+\frac{1}{2}} \in \mathbf{V}_h, \mathbf{H}_h^{m+\frac{1}{2}}, \mathbf{K}_h^m \in \mathbf{U}_h$ such that

$$
\begin{aligned}
&(\frac{\mathbf{E}_h^m - \mathbf{E}_h^{m-1}}{\tau}, \phi_h) - (\mathbf{H}_h^{m-\frac{1}{2}}, \nabla\times\phi_h) + (\mathbf{J}_h^{m-\frac{1}{2}}, \phi_h) = 0, \\
&(\frac{\mathbf{H}_h^{m+\frac{1}{2}} - \mathbf{H}_h^{m-\frac{1}{2}}}{\tau}, \psi_h) + (\nabla\times\mathbf{E}_h^m, \psi_h) + (\mathbf{K}_h^m, \psi_h) = 0, \\
&(\frac{\mathbf{J}_h^{m+\frac{1}{2}} - \mathbf{J}_h^{m-\frac{1}{2}}}{\tau}, \tilde{\phi}_h) + \Gamma_e(\frac{1}{2}(\mathbf{J}_h^{m+\frac{1}{2}} + \mathbf{J}_h^{m-\frac{1}{2}}), \tilde{\phi}_h) - \omega_e^2(\mathbf{E}_h^m, \tilde{\phi}_h) = 0, \\
&(\frac{\mathbf{K}_h^m - \mathbf{K}_h^{m-1}}{\tau}, \tilde{\psi}_h) + \Gamma_m(\frac{1}{2}(\mathbf{K}_h^m + \mathbf{K}_h^{m-1}), \tilde{\psi}_h) - \omega_m^2(\mathbf{H}_h^{m-\frac{1}{2}}, \tilde{\psi}_h) = 0,
\end{aligned}
$$

hold true for test functions $\phi_h \in \mathbf{V}_h^0, \psi_h \in \mathbf{U}_h, \tilde{\phi}_h \in \mathbf{V}_h, \tilde{\psi}_h \in \mathbf{U}_h$. Combining the above proof techniques with those developed for the leap-frog scheme [183], we can obtain the following superclose result:

$$
\begin{aligned}
&\max_{1\le m}(||\Pi_h\mathbf{E}^m - \mathbf{E}_h^m||_0 + ||P_h\mathbf{H}^{m+\frac{1}{2}} - \mathbf{H}_h^{m+\frac{1}{2}}||_0 \\
&\quad +||\Pi_h\mathbf{J}^{m+\frac{1}{2}} - \mathbf{J}_h^{m+\frac{1}{2}}||_0 + ||P_h\mathbf{K}^m - \mathbf{K}_h^m||_0) \le C(\tau^2 + h^{k+1}).
\end{aligned}
$$

5.4 Superconvergence in the Discrete l_2 Norm

In this section, we first prove a superconvergence interpolation result obtained at element centers for the lowest order cubic edge element (i.e., $k = 1$ in spaces $\mathbf{U}_h$ and $\mathbf{V}_h$). Then we use that to obtain a global superconvergence result in the discrete l_2 norm.

Lemma 5.3. *Let $K = [x_c - h_x, x_c + h_x] \times [y_c - h_y, y_c + h_y] \times [z_c - h_z, z_c + h_z]$ be an arbitrary cubic element with maximum length h. Then for any $\mathbf{u} \in \mathbf{W}^{2,\infty}(K)$ and its corresponding Nédélec interpolation $\Pi_K^c \mathbf{u} \in Q_{0,1,1} \times Q_{1,0,1} \times Q_{1,1,0}$, we have*

$$\left(\mathbf{u} - \Pi_K^c \mathbf{u}\right)(x_c, y_c, z_c) \leq C h^2. \tag{5.25}$$

Proof. Note that the lowest order $H(curl)$ interpolation $\Pi_K^c \mathbf{u}$ can be written explicitly as (cf. Example 3.5)

$$\Pi_K^c \mathbf{u}(x, y, z) = (\Pi_K^c u_1, \Pi_K^c u_2, \Pi_K^c u_3) = \sum_{i=1}^{12} \left(\int_{l_i} \mathbf{u} \cdot \tau_i dl \right) \mathbf{N}_i(x, y, z), \tag{5.26}$$

where l_i are the 12 edges of the element, and τ_i represent the unit tangent vector along l_i (cf. Fig. 5.1), and $\mathbf{N}_i \in Q_{0,1,1} \times Q_{1,0,1} \times Q_{1,1,0}$ are the basis functions.

The first component of $\Pi_K^c \mathbf{u}$ is

$$\begin{aligned}
(\Pi_K^c u)_1 = & \left(\int_{l_1} u_1(x, y_c - h_y, z_c - h_z) dl \right) \cdot \frac{1}{8 h_x h_y h_z} (y_c + h_y - y)(z_c + h_z - z) \\
& + \left(\int_{l_2} u_1(x, y_c + h_y, z_c - h_z) dl \right) \cdot \frac{1}{8 h_x h_y h_z} (y + h_y - y_c)(z_c + h_z - z) \\
& + \left(\int_{l_3} u_1(x, y_c - h_y, z_c + h_z) dl \right) \cdot \frac{1}{8 h_x h_y h_z} (y_c + h_y - y)(z + h_z - z_c) \\
& + \left(\int_{l_4} u_1(x, y_c + h_y, z_c + h_z) dl \right) \cdot \frac{1}{8 h_x h_y h_z} (y + h_y - y_c)(z + h_z - z_c),
\end{aligned}$$

from which we see that the value at the element center is

$$\begin{aligned}
(\Pi_K^c u)_1(x_c, y_c, z_c) = & \frac{1}{8 h_x} \left(\int_{l_1} u_1(x, y_c - h_y, z_c - h_z) dl + \int_{l_2} u_1(x, y_c + h_y, z_c - h_z) dl \right) \\
& + \frac{1}{8 h_x} \left(\int_{l_3} u_1(x, y_c - h_y, z_c + h_z) dl + \int_{l_4} u_1(x, y_c + h_y, z_c + h_z) dl \right).
\end{aligned}$$

By Taylor expansion at x_c and the fact that $\int_{x_c - h_x}^{x_c + h_x} (x - x_c) dx = 0$, we easily have

$$\begin{aligned}
& \int_{l_1} u_1(x, y_c - h_y, z_c - h_z) \\
= & \int_{l_1} [u_1(x_c, y_c - h_y, z_c - h_z) + O(h_x^2) \partial_{xx} u_1(x_*, y_c - h_y, z_c - h_z)] dl \\
= & \, 2 h_x u_1(x_c, y_c - h_y, z_c - h_z) + 2 h_x O(h_x^2) \partial_{xx} u_1(x_*, y_c - h_y, z_c - h_z),
\end{aligned}$$

where x_* is some number between x and x_c.

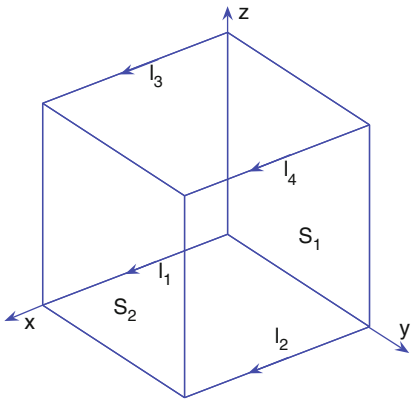

Fig. 5.1 The exemplary cubic edge element

Similar estimates can be obtained for other line integrals. Hence we have

$$(\Pi_K^c u)_1(x_c, y_c, z_c) = \frac{1}{4}[u_1(x_c, y_c - h_y, z_c - h_z) + u_1(x_c, y_c + h_y, z_c - h_z) + u_1(x_c, y_c - h_y, z_c + h_z) + u_1(x_c, y_c + h_y, z_c + h_z)] + O(h_x^2).$$

Using Taylor expansion at (x_c, y_c, z_c) again, we can easily see that

$$(\Pi_K^c u)_1(x_c, y_c, z_c) = u_1(x_c, y_c, z_c) + O(h_x^2 + h_y^2 + h_z^2).$$

By symmetry, the same estimates can be proved for the second and third components of $\Pi_K^c \mathbf{u}$. □

With the above estimates, we can now obtain a superconvergence result in the discrete l_2 norm, which is one-order higher compared to the optimal error estimate obtained in the continuous L_2 norm.

Theorem 5.3. *Let* $\mathbf{x}_c^K = (x_c, y_c, z_c)$ *be the center of a cubic element* $K = [x_c - h_x, x_c + h_x] \times [y_c - h_y, y_c + h_y] \times [z_c - h_z, z_c + h_z]$, $(\mathbf{E}^m, \mathbf{H}^m)$ *and* $(\mathbf{E}_h^m, \mathbf{H}_h^m)$ *be the analytical and numerical solutions of (5.1)–(5.4) and (5.16)–(5.19), respectively. Under the assumptions of Theorem 5.2 (with* $m = 1$*) and Lemma 5.3, we have*

$$\max_{1 \le m \le M} \left(||\mathbf{E}^m - \mathbf{E}_h^m||_{l^2} + ||\mathbf{H}^m - \mathbf{H}_h^m||_{l^2} \right) \le C(\tau^2 + h^2),$$

where we denote $||u||_{l_2} = \left(\sum_e |u(\mathbf{x}_c^K)|^2 \cdot |K| \right)^{\frac{1}{2}}$, *and* $|K|$ *for the volume of element* K.

Proof. Note that any $u_h \in Q_{0,1,1}$ can be written as $(c_1 + c_2 y)(c_3 + c_4 z)$, which satisfies the identity

$$u_h(\mathbf{x}_c^K) = (c_1 + c_2 y_c)(c_3 + c_4 z_c) = \frac{1}{|K|} \int_K u_h dx dy dz. \tag{5.27}$$

For the lowest order edge element space $\mathbf{V}_h$, both $\Pi_h\mathbf{E}^m$ (simplified notation of $\Pi_h^c\mathbf{E}^m$) and $\mathbf{E}_h^m \in Q_{0,1,1} \times Q_{1,0,1} \times Q_{1,1,0}$, hence applying (5.27) to the first component of $\Pi_h\mathbf{E}^m - \mathbf{E}_h^m$, we have

$$(\Pi_h\mathbf{E}^m - \mathbf{E}_h^m)_1(\mathbf{x}_c^K) = \frac{1}{|K|}\int_K (\Pi_K^c\mathbf{E}^m - \mathbf{E}_h^m)_1 dxdydz.$$

Using the Cauchy-Schwarz inequality and Theorem 5.2 with $k = 1$, we obtain

$$\begin{aligned}\sum_{K\in T^h} |(\Pi_h\mathbf{E}^k - \mathbf{E}_h^k)_1(\mathbf{x}_c^K)|^2 \cdot |K| &= \sum_{K\in T^h} \frac{1}{|K|}\left(\int_K (\Pi_K^c\mathbf{E}^m - \mathbf{E}_h^m)_1 dxdydz\right)^2 \\ &\le \int_\Omega (\Pi_h\mathbf{E}^m - \mathbf{E}_h^m)_1^2 dxdydz \le C(\tau^4 + h^4).\end{aligned}$$

Same estimates can be proved for the other two components. Then by the triangle inequality and Lemma 5.3, we have

$$||\mathbf{E}^m - \mathbf{E}_h^m||_{l_2} \le ||\mathbf{E}^m - \Pi_h\mathbf{E}^m||_{l_2} + ||\Pi_h\mathbf{E}^m - \mathbf{E}_h^m||_{l_2} \le C(\tau^2 + h^2). \quad (5.28)$$

The estimate $||\mathbf{H}^m - \mathbf{H}_h^m||_{l_2} \le C(\tau^2 + h^2)$ can be proved similarly (cf. [156]) □

5.5 Extensions to 2-D Superconvergence Analysis

In this section, we want to prove similar superclose results for the 2-D Maxwell's equations. Note that in some sense the 2-D case is more complicated than the 3-D case, since in the 2-D Maxwell's equations, one field is a 2-D vector, while the other field becomes a scalar. Without loss of generality, here we assume that the electrical field $\mathbf{E}$ is a vector, while the magnetic field H is a scalar. To make the extension clearly, we define the 2-D vector and scalar curl operators:

$$\nabla \times H = (\frac{\partial H}{\partial y}, -\frac{\partial H}{\partial x})', \quad \nabla \times \mathbf{E} = \frac{\partial E_2}{\partial x} - \frac{\partial E_1}{\partial y}, \quad \forall\ \mathbf{E} \equiv (E_1, E_2)'. \quad (5.29)$$

5.5.1 Superconvergence on Rectangular Edge Elements

For a 2-D domain Ω, we partition it by a family of regular rectangular meshes T_h with maximum mesh size h. The corresponding Raviart-Thomas-Nédélec rectangular elements are:

$$\begin{aligned}U_h &= \{\psi_h \in L^2(\Omega) :\ \psi_h|_K \in Q_{k-1,k-1},\ \forall\ K \in T_h\},\\ \mathbf{V}_h &= \{\phi_h \in H(\mathrm{curl};\Omega) :\ \phi_h|_K \in Q_{k-1,k} \times Q_{k,k-1},\ \forall\ K \in T_h\},\end{aligned}$$

for any $k \geq 1$. Recall that $Q_{i,j}$ denotes the space of polynomials whose degrees are less than or equal to i, j in variables x, y, respectively. It is easy to see that $\nabla \times \mathbf{V}_h \subset U_h$ still holds.

In the 2-D case, the Nédélec operator $\Pi_h \mathbf{E} \in \mathbf{V}_h$ is defined as:

$$\int_{l_i} (\mathbf{E} - \Pi_h \mathbf{E}) \cdot \tau_i q dl = 0, \quad \forall\, q \in P_{k-1}(l_i),\ i = 1, \cdots, 4, \tag{5.30}$$

$$\int_K (\mathbf{E} - \Pi_h \mathbf{E}) \cdot \mathbf{q} dx dy = 0, \quad \forall\, \mathbf{q} \in Q_{k-1,k-2} \times Q_{k-2,k-1}, \tag{5.31}$$

where l_i denotes the i-th edge of an element K, and τ_i is the unit tangent vector along the edge l_i. When $k = 1$ (the lowest-order rectangular edge element), $\Pi_h \mathbf{E}$ is defined by (5.30) only.

The 2-D superclose analysis depends on the following fundamental results.

Lemma 5.4. *For any $\mathbf{u} \in H(curl; K)$ and $q \in Q_{k-1,k-1}(K), k \geq 1$, we have*

$$\int_K \nabla \times (\mathbf{u} - \Pi_h \mathbf{u}) \cdot q dx dy = 0.$$

Proof. The proof follows from the Stokes' formula

$$\int_K \nabla \times (\mathbf{u} - \Pi_h \mathbf{u}) \cdot q dx dy = \int_{\partial K} (\mathbf{u} - \Pi_h \mathbf{u}) \cdot \tau q dl + \int_K (\mathbf{u} - \Pi_h \mathbf{u}) \cdot (\nabla \times q) dx dy$$

and the property (5.30) and (5.31) for the operator Π_h. □

Let P_h be the L^2-projection operator onto the space U_h. By the property $\nabla \times \mathbf{V}_h \subset U_h$, we immediately have

Lemma 5.5. *For any $w \in L^2(K)$ and $\phi_h|_K \in Q_{k-1,k} \times Q_{k,k-1}, k \geq 1$, we have*

$$\int_K (w - P_h w) \cdot \nabla \times \phi_h dx dy = 0.$$

Lemma 5.6. *Let $K = [x_c - h_x, x_c + h_x] \times [y_c - h_y, y_c + h_y]$ be an arbitrary rectangular element. Then for any $\mathbf{u} \in H(curl; K)$ and $\phi_h|_K \in Q_{k-1,k} \times Q_{k,k-1}, k \geq 1$, we have*

$$\int_K (u_1 - (\Pi_h \mathbf{u})_1) \phi_1 dx dy = O(h_y^{k+1}) ||\partial_y^{k+1} u_1||_{0,K} ||\phi_1||_{0,K}, \tag{5.32}$$

$$\int_K (u_2 - (\Pi_h \mathbf{u})_2) \phi_2 dx dy = O(h_x^{k+1}) ||\partial_x^{k+1} u_2||_{0,K} ||\phi_2||_{0,K}, \tag{5.33}$$

where u_1, u_2 and ϕ_1, ϕ_2 are the two components of $\mathbf{u}$ *and* ϕ_h*, respectively. Hence, we have*

$$\int_K (\mathbf{u} - \Pi_h \mathbf{u}) \cdot \phi_h dxdy = O(h^{k+1}) ||\mathbf{u}||_{k+1,K} ||\phi_h||_{0,K}.$$

Proof. Since

$$\int_K (\mathbf{u} - \Pi_h \mathbf{u}) \cdot \phi_h dxdy = \int_K (u_1 - (\Pi_h \mathbf{u})_1)\phi_1 dxdy + \int_K (u_2 - (\Pi_h \mathbf{u})_2)\phi_2 dxdy,$$

we just need to consider the first inner product. For simplicity, below we just present the proof for the $k = 1$ case. For $k \geq 2$ case, interested readers can find the detailed proof in the original paper [153].

By definition, when $k = 1$, $\phi_1 \in Q_{0,1}$. Then by the Taylor expansion, we obtain

$$\begin{aligned} &\int_K (u_1 - (\Pi_h \mathbf{u})_1)\phi_1 dxdy \\ &= \int_K (u_1 - (\Pi_h \mathbf{u})_1)[\phi_1(x_c, y_c) + (y - y_c)\partial_y \phi_1(x_c, y_c)]dxdy. \end{aligned} \tag{5.34}$$

Denote the functions

$$A(x) = \frac{1}{2}[(x - x_c)^2 - h_x^2], \quad B(y) = \frac{1}{2}[(y - y_c)^2 - h_y^2]. \tag{5.35}$$

Note that in the proof below we will constantly use the facts that:

$$A(x) = 0 \quad \text{on } x = x_c \pm h_x, \quad B(y) = 0 \quad \text{on } y = y_c \pm h_y. \tag{5.36}$$

Using integration by parts and the identity $\partial_{yy} B(y) = 1$, (5.30) and (5.36), we have

$$\begin{aligned} &\int_K (u_1 - (\Pi_h \mathbf{u})_1) dxdy = \int_K (u_1 - (\Pi_h \mathbf{u})_1)\partial_{yy} B(y) dxdy \\ &= \int_{x=x_c-h_x}^{x_c+h_x} (u_1 - (\Pi_h \mathbf{u})_1)\partial_y B(y)|_{y=y_c-h_y}^{y_c+h_y} dx - \int_K (u_1 - (\Pi_h \mathbf{u})_1)_y \partial_y B(y) dxdy \\ &= \int_K (u_1 - (\Pi_h \mathbf{u})_1)_{yy} B(y) dxdy = \int_K \partial_{yy} u_1 \cdot B(y) dxdy, \end{aligned}$$

where in the last step we used the fact that $(\Pi_h \mathbf{u})_1 \in Q_{0,1}$.

Similarly, by the identity $y - y_c = \frac{1}{6}\partial_y^3(B^2(y))$ and integration by parts, we obtain

$$\begin{aligned}
&\int_K (u_1 - (\Pi_h \mathbf{u})_1)(y - y_c)dxdy = \int_K (u_1 - (\Pi_h \mathbf{u})_1) \cdot \frac{1}{6}\partial_y^3(B^2(y))dxdy \\
&= \int_{x=x_c-h_x}^{x_c+h_x} (u_1 - (\Pi_h \mathbf{u})_1) \cdot \frac{1}{6}\partial_y^2(B^2(y))|_{y=y_c-h_y}^{y_c+h_y} dx \\
&\quad - \int_K (u_1 - (\Pi_h \mathbf{u})_1)_y \cdot \frac{1}{6}\partial_y^2(B^2(y))dxdy \\
&= \int_K (u_1 - (\Pi_h \mathbf{u})_1)_{yy} \cdot \frac{1}{6}\partial_y(B^2(y))dxdy = \int_K \partial_{yy}u_1 \cdot \frac{1}{6}(B^2(y))_y dxdy.
\end{aligned}$$

Substituting the above integral identities into (5.34) and using the inverse estimate, we have

$$\begin{aligned}
&\int_K (u_1 - (\Pi_h \mathbf{u})_1)\phi_1 dxdy \\
&= \int_K \partial_{yy}u_1 \cdot B(y) \cdot \phi_1(x_c, y_c)dxdy + \int_K \partial_{yy}u_1 \cdot \frac{1}{6}(B^2(y))_y \cdot \partial_y\phi_1(x_c, y_c)dxdy \\
&= \int_K \partial_{yy}u_1 \cdot B(y) \cdot [\phi_1(x, y) - (y - y_c)\partial_y\phi_1(x, y)]dxdy \\
&\quad + \int_K \partial_{yy}u_1 \cdot \frac{1}{3}B(y) \cdot (y - y_c)\partial_y\phi_1(x, y)dxdy \\
&= O(h_y^2)||\partial_{yy}u_1||_{0,K}||\phi_1||_{0,K}.
\end{aligned}$$

Using the same arguments, we can prove

$$\int_K (u_2 - (\Pi_h \mathbf{u})_2)\phi_2 dxdy = O(h_x^2)||\partial_{xx}u_2||_{0,K}||\phi_2||_{0,K},$$

which completes our proof for the $k = 1$ case. □

With Lemmas 5.4–5.6, we can see that Theorems 5.1 and 5.2 hold true for 2-D rectangular elements. Below we want to show that for the lowest-order edge element (i.e., $k = 1$ in U_h and $\mathbf{V}_h$), we have one-order higher superconvergence in the L^∞-norm at rectangular element centers.

Lemma 5.7. *Let $K = [x_c - h_x, x_c + h_x] \times [y_c - h_y, y_c + h_y]$ be an arbitrary rectangular element. Then for any $\mathbf{u} \in H(curl; K)$ and $\Pi_h \mathbf{u}|_K \in Q_{0,1} \times Q_{1,0}$, we have*

$$(\mathbf{u} - \Pi_K^c \mathbf{u})(x_c, y_c) = O(h^2). \tag{5.37}$$

Proof. For the lowest-order edge element $Q_{0,1} \times Q_{1,0}$, the interpolation $\Pi_K^c \mathbf{u}$ of any $\mathbf{u} \in H(\text{curl}; K)$ can be written as (cf. Example 3.6):

$$\Pi_K^c \mathbf{u}(x, y) = \sum_{j=1}^{4} \left(\int_{l_j} \mathbf{u} \cdot \tau_j \, dl \right) \mathbf{N}_j(x, y), \tag{5.38}$$

where we denote l_j the four edges of the element, which start from the bottom edge and are oriented counterclockwise. Furthermore, we denote τ_j for the unit tangent vector along l_j. Recall that the edge element basis functions $\mathbf{N}_j$ are as follows (cf. Example 3.6):

$$\mathbf{N}_1 = \begin{pmatrix} \frac{(y_c+h_y)-y}{4h_x h_y} \\ 0 \end{pmatrix}, \quad \mathbf{N}_2 = \begin{pmatrix} 0 \\ \frac{x-(x_c-h_x)}{4h_x h_y} \end{pmatrix},$$

$$\mathbf{N}_3 = \begin{pmatrix} \frac{(y_c-h_y)-y}{4h_x h_y} \\ 0 \end{pmatrix}, \quad \mathbf{N}_4 = \begin{pmatrix} 0 \\ \frac{x-(x_c+h_x)}{4h_x h_y} \end{pmatrix}.$$

By (5.38) and the notation $\mathbf{u} = (u_1, u_2)'$, we have

$$\begin{aligned}
&\Pi_K^c \mathbf{u}(x_c, y_c) \\
&= \int_{l_1} u_1(x, y_c - h_y) dx \cdot \begin{pmatrix} \frac{(y_c+h_y)-y_c}{4h_x h_y} \\ 0 \end{pmatrix} + \int_{l_2} u_2(x_c + h_x, y) dy \cdot \begin{pmatrix} 0 \\ \frac{x_c-(x_c-h_x)}{4h_x h_y} \end{pmatrix} \\
&\quad - \int_{l_3} u_1(x, y_c + h_y) dx \cdot \begin{pmatrix} \frac{(y_c-h_y)-y_c}{4h_x h_y} \\ 0 \end{pmatrix} - \int_{l_4} u_2(x_c - h_x, y) dy \cdot \begin{pmatrix} 0 \\ \frac{x_c-(x_c+h_x)}{4h_x h_y} \end{pmatrix},
\end{aligned}$$

from which we obtain the first component

$$\begin{aligned}
&\frac{1}{4h_x} \left(\int_{x_c-h_x}^{x_c+h_x} u_1(x, y_c - h_y) dx + \int_{x_c-h_x}^{x_c+h_x} u_1(x, y_c + h_y) dx \right) \\
&= \frac{1}{4h_x} \left(\int_{x_c-h_x}^{x_c+h_x} [u_1(x_c, y_c - h_y) + (x - x_c)\partial_x u_1(x_c, y_c - h_y) + O(h_x^2)] dx \right. \\
&\quad \left. + \int_{x_c-h_x}^{x_c+h_x} [u_1(x_c, y_c + h_y) + (x - x_c)\partial_x u_1(x_c, y_c + h_y) + O(h_x^2)] dx \right) \\
&= \frac{1}{2} [u_1(x_c, y_c - h_y) + u_1(x_c, y_c + h_y)] + O(h_x^2),
\end{aligned}$$

where we used the Taylor expansion and the fact that $\int_{x_c-h_x}^{x_c+h_x} (x - x_c) dx = 0$. Using the Taylor expansion one more time, we can easily see that

$$
\begin{aligned}
&((\Pi_K^c \mathbf{u})_1 - u_1)(x_c, y_c) \\
&= \frac{1}{2}[u_1(x_c, y_c - h_y) + u_1(x_c, y_c + h_y)] - u_1(x_c, y_c) + O(h_x^2) \\
&= O(h_x^2) + O(h_y^2).
\end{aligned}
$$

By the same arguments, we can obtain the same estimate for the second component:

$$((\Pi_K^c \mathbf{u})_2 - u_2)(x_c, y_c) = O(h_x^2) + O(h_y^2),$$

which completes the proof. □

With the above preparations, finally we can prove the following L^∞ superconvergence result.

Theorem 5.4. *Let (x_c, y_c) be the center of a rectangular element $K = [x_c - h_x, x_c + h_x] \times [y_c - h_y, y_c + h_y]$, and $\mathbf{E}^h$ and H^h be the lowest-order finite element solution of (5.9)–(5.12), i.e., $\mathbf{E}^h|_K \in Q_{0,1} \times Q_{1,0}$ and $H^h|_K \in Q_{0,0}$. Under the assumption that the L^2 norms of $\Pi_h \mathbf{E} - \mathbf{E}^h$ and $P_h H - H^h$ are almost uniformly distributed, i.e.,*

$$\int_K |\Pi_h \mathbf{E} - \mathbf{E}^h|^2 dK \le \frac{C}{N} \int_\Omega |\Pi_h \mathbf{E} - \mathbf{E}^h|^2 dK, \tag{5.39}$$

$$\int_K |P_h H - H^h|^2 dK \le \frac{C}{N} \int_\Omega |P_h H - H^h|^2 dK, \tag{5.40}$$

where N denotes the total number of elements over Ω. Then on a quasi-uniform mesh we have the L^∞ superconvergence

$$|(\mathbf{E} - \mathbf{E}^h)(x_c, y_c)| + |(H - H^h)(x_c, y_c)| \le C h^2. \tag{5.41}$$

Proof. Using the fact that the m-point Gaussian quadrature holds exactly for all polynomials up to degree $2m - 1$, and the Cauchy-Schwarz inequality, for the first component of error $\Pi_h \mathbf{E} - \mathbf{E}^h$ we easily have

$$
\begin{aligned}
&|(\Pi_h \mathbf{E} - \mathbf{E}^h)_1(x_c, y_c)| = |\frac{1}{|K|} \int_K (\Pi_h \mathbf{E} - \mathbf{E}^h)_1 dxdy| \\
&\le \frac{1}{|K|} (\int_K |(\Pi_h \mathbf{E} - \mathbf{E}^h)_1|^2 dxdy)^{1/2} (\int_K 1^2 dxdy)^{1/2} \\
&\le \frac{1}{|K|^{1/2}} (\frac{1}{N} \int_\Omega |(\Pi_h \mathbf{E} - \mathbf{E}^h)_1|^2 dxdy)^{1/2} \\
&\le \frac{1}{(N|K|)^{1/2}} \cdot Ch^2 \le Ch^2,
\end{aligned} \tag{5.42}
$$

where we used Theorem 5.1 and the fact that $N|K| \approx \text{meas}(\Omega) = O(1)$. Similar estimate can be obtained for the second component, i.e.,

$$|(\Pi_h \mathbf{E} - \mathbf{E}^h)_2(x_c, y_c)| = O(h^2),$$

from which and Lemma 5.4, we obtain

$$(\mathbf{E} - \mathbf{E}^h)(x_c, y_c) = (\mathbf{E} - \Pi_h \mathbf{E})(x_c, y_c) + (\Pi_h \mathbf{E} - \mathbf{E}^h)(x_c, y_c) = O(h^2).$$

Note that for any function $f(x, y)$, by Taylor expansion, we have

$$\begin{aligned} &\frac{1}{|K|}\int_K f(x,y)dxdy - f(x_c,y_c) = \frac{1}{|K|}\int_K (f(x,y) - f(x_c,y_c))dxdy \\ &= \frac{1}{|K|}\int_K [(x-x_c)\partial_x f(x_c,y_c) + (y-y_c)\partial_y f(x_c,y_c) + O(h^2)]dxdy \\ &= O(h^2), \end{aligned} \tag{5.43}$$

using which, the fact that $\int_K (P_h H - H)dxdy = 0$ and similar arguments used in (5.42), we have

$$\begin{aligned} &(H - H^h)(x_c, y_c) \approx \frac{1}{|K|}\int_K (H - H^h)(x,y)dxdy + O(h^2) \\ &= \frac{1}{|K|}\int_K (P_h H - H^h)(x,y)dxdy + O(h^2) \\ &\le \frac{1}{|K|}\Big(\int_K |P_h H - H^h|^2 dxdy\Big)^{1/2}\Big(\int_K 1^2 dxdy\Big)^{1/2} + O(h^2) \le Ch^2, \end{aligned}$$

which concludes the proof. □

By similar arguments, for the fully-discrete scheme (5.16)–(5.19), under the constraints (5.39) and (5.40), we can prove

$$\max_{1\le m\le M}(|(\mathbf{E}^m - \mathbf{E}_h^m)(x_c, y_c)| + |(H^m - H_h^m)(x_c, y_c)|) \le C(h^2 + \tau^2).$$

Without imposing the constraints (5.39) and (5.40), we can similarly prove the discrete l_2 superconvergence as Theorem 5.3. More specifically, we have

Theorem 5.5. *Let $\mathbf{x}_c^K = (x_c, y_c)$ be the center of a rectangular element $K = [x_c - h_x, x_c + h_x] \times [y_c - h_y, y_c + h_y]$, $(\mathbf{E}^m, H^m)$ and $(\mathbf{E}_h^m, H_h^m)$ be the 2-D analytical and numerical solutions of (5.1)–(5.4) and (5.16)–(5.19), respectively. Then we have*

$$\max_{1\le m\le M}\left(||\mathbf{E}^m - \mathbf{E}_h^m||_{l^2} + ||H^m - H_h^m||_{l^2}\right) \le C(\tau^2 + h^2),$$

where we denote $||u||_{l_2} = \left(\sum_e |u(\mathbf{x}_c^K)|^2 \cdot |K| \right)^{\frac{1}{2}}$, *and* $|K|$ *for the area of element* K.

Numerical results demonstrating L^∞ convergence rate $O(h^2)$ at rectangular element centers are indeed observed for the lowest-order rectangular edge element. Detailed results are presented in Chap. 7.

5.5.2 Superconvergence on Triangular Edge Elements

In this section, we would like to show that some superconvergence results as Sect. 5.5.1 hold true for the lowest-order triangular edge element. We assume that the domain Ω is partitioned by a family of regular triangular meshes T_h with maximum mesh size h, in which case the mixed finite element spaces used to solve (5.16)–(5.19) are:

$$U_h = \{\psi_h \in L^2(\Omega) : \ \psi_h = \text{ piecewise constant}, \ \forall\ K \in T_h\},$$
$$\mathbf{V}_h = \{\phi_h \in H(\text{curl}; \Omega) : \ \phi_h|_K = \text{span}(\lambda_i \nabla\lambda_j - \lambda_j \nabla\lambda_i), \ i, j = 1, 2, 3, \ \forall\ K \in T_h\},$$

where λ_i is the barycentric coordinate at the i-th vertex of the triangle K.

By the Stokes' formula, it is easy to see that:

Lemma 5.8. *For any* $\mathbf{u} \in H(curl; K)$, *we have*

$$\int_K \nabla \times (\mathbf{u} - \Pi_h \mathbf{u}) dxdy = 0.$$

Note that for any $\phi_h|_K \in \mathbf{V}_h$, $\nabla \times \phi_h$ is a constant, hence we easily have the following result.

Lemma 5.9. *For any* $w \in L^2(K)$ *and* $\phi_h|_K \in \mathbf{V}_h$, *we have*

$$\int_K (w - P_h w) \cdot \nabla \times \phi_h dxdy = 0.$$

Since there exists no natural superconvergence point for the numerical solution of (5.16)–(5.19) obtained with the lowest-order triangular edge element, we consider a special triangular mesh formed by parallelograms such as Fig. 5.2.

Below is a superclose result between a function and its Nédélec interpolation on a parallelogram.

Theorem 5.6 ([154, Theorem 3.3]). *On a parallelogram* $\diamondsuit$ *formed by two triangles, if* $\mathbf{u} \in H(curl; \diamondsuit) \cap H^3(\diamondsuit)$ *and* $\phi_h \in \mathbf{V}_h$, *then we have*

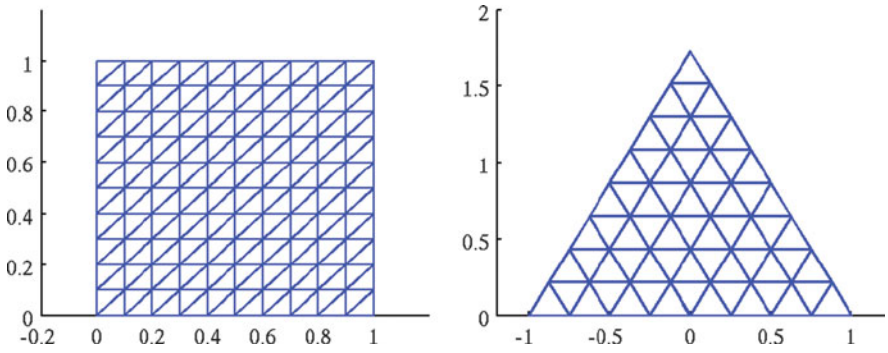

Fig. 5.2 Exemplary triangular meshes formed by parallelograms

$$\int_{\diamond} (\mathbf{u} - (\Pi_h \mathbf{u})) \cdot \phi_h \, \mathrm{d}x\mathrm{d}y = O(h^2)||\partial^3 \mathbf{u}||_{0,\diamond}||\phi_h||_{0,\diamond}. \tag{5.44}$$

Note that by the standard interpolation estimate [217], we only have

$$\int_{\diamond} (\mathbf{u} - (\Pi_h \mathbf{u})) \cdot \phi_h = O(h)||\mathbf{u}||_{H(\mathrm{curl};\diamond)}||\phi_h||_{0,\diamond},$$

which is one order less than (5.44).

Through some technical calculation, a pointwise superconvergence result at the center of each parallelogram can be proved by taking an average of the interpolations from those two neighboring triangles.

Theorem 5.7 ([154, Theorem 3.4]). *Assume that (x_c, y_c) is the center of one parallelogram $\diamond$ formed by two triangles L and R, then for any $\mathbf{u} = (u_1, u_2) \in C^2(\overline{\diamond})$, we have*

$$[\mathbf{u} - \frac{1}{2}((\Pi_h \mathbf{u})|_L + (\Pi_h \mathbf{u})|_R)](x_c, y_c) = O(h^2).$$

Another interesting result for the lowest-order triangular edge element is that the average of a function over a parallelogram is equal to the function value at the parallelogram center.

Lemma 5.10. *Consider a parallelogram $\diamond$ formed by vertices $A(x_c - l_3\cos\alpha, y_c - l_3\sin\alpha)$, $B(x_c - l_3\cos\alpha + 2l_1, y_c - l_3\sin\alpha)$, $C(x_c + l_3\cos\alpha, y_c + l_3\sin\alpha)$, and $D(x_c + l_3\cos\alpha - 2l_1, y_c + l_3\sin\alpha)$, where $O(x_c, y_c)$ denotes the midpoint of AC, $\alpha = \angle CAB$, $2l_1, 2l_2$ and $2l_3$ are the lengths of AB, BC and CA, respectively. The following holds true (for any parallelogram in Fig. 5.3):*

$$\frac{1}{|\diamond|}\int_{\diamond} \mathbf{u}_h dx dy = \mathbf{u}_h(x_c, y_c) \quad \forall\, \mathbf{u}_h \in \mathbf{V}_h. \tag{5.45}$$

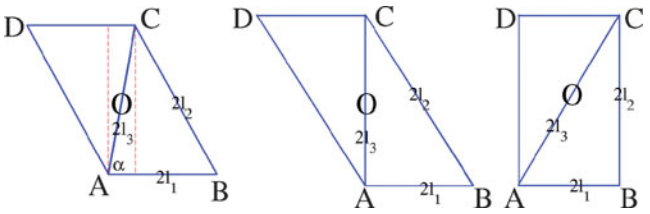

Fig. 5.3 The exemplary parallelograms

Proof. By definition of $\mathbf{V}_h$, the first component of $\mathbf{u}_h \in \mathbf{V}_h$ can be written as (cf. [154, Lemma 2.1]):

$$u_h^1 = c_1 + c_2 y,$$

where c_1 and c_2 are some constants. Below we just prove (5.45) for the first component on the general parallelogram (the left one in Fig. 5.3), since proofs of the other cases are easier.

From Fig. 5.3, we can write the line equations of AD and BC respectively:

$$l_{AD}: \ y - (y_c - l_3 \sin\alpha) = \frac{l_3 \sin\alpha}{l_3 \cos\alpha - 2l_1}[x - (x_c - l_3 \cos\alpha)],$$

$$l_{BC}: \ y - (y_c - l_3 \sin\alpha) = \frac{l_3 \sin\alpha}{l_3 \cos\alpha - 2l_1}[x - (x_c - l_3 \cos\alpha + 2l_1)],$$

solving which for x, we obtain

$$x_{l_{AD}} = \frac{l_3 \cos\alpha - 2l_1}{l_3 \sin\alpha}[y - (y_c - l_3 \sin\alpha)] + (x_c - l_3 \cos\alpha),$$

$$x_{l_{BC}} = \frac{l_3 \cos\alpha - 2l_1}{l_3 \sin\alpha}[y - (y_c - l_3 \sin\alpha)] + (x_c - l_3 \cos\alpha + 2l_1).$$

Therefore, we have

$$\int_{\diamond} u_h^1 dxdy = \int_{\diamond} (c_1 + c_2 y) dxdy$$

$$= \int_{y_c - l_3 \sin\alpha}^{y_c + l_3 \sin\alpha} \int_{x_{l_{AD}}}^{x_{l_{BC}}} (c_1 + c_2 y) dxdy = \int_{y_c - l_3 \sin\alpha}^{y_c + l_3 \sin\alpha} 2l_1 (c_1 + c_2 y) dy$$

$$= 2l_1[c_1 \cdot 2l_3 \sin\alpha + c_2 \cdot 2y_c l_3 \sin\alpha] = 2l_1 \cdot 2l_3 \sin\alpha (c_1 + c_2 y_c) = |\diamond| u_h^1(x_c, y_c).$$

By the same technique, we can prove that $\int_{\diamond} u_h^2 dxdy = |\diamond| u_h^2(x_c, y_c)$, which concludes our proof. □

Using the above results, we can prove that the averaged solutions have pointwise superconvergence at parallelogram centers (cf. [154, Theorem 4.3]).

Theorem 5.8. *Let (x_c, y_c) be the center of a parallelogram $\diamond$ shown in Fig. 5.3, and $\mathbf{E}_h^m$ and H_h^m be the finite element solution of (5.16)–(5.19) at time level t_m. If the L^2 estimates of $\Pi_h \mathbf{E}^m - \mathbf{E}_h^m$ and $P_h H^m - H_h^m$ are almost uniformly distributed over Ω, i.e.,*

$$\int_{\diamond} |\Pi_h \mathbf{E}^m - \mathbf{E}_h^m|^2 dxdy \leq \frac{C}{N} \int_{\Omega} |\Pi_h \mathbf{E}^m - \mathbf{E}_h^m|^2 dxdy, \tag{5.46}$$

$$\int_{\diamond} |P_h H^m - H_h^m|^2 dxdy \leq \frac{C}{N} \int_{\Omega} |P_h H^m - H_h^m|^2 dxdy, \tag{5.47}$$

where N denotes the total number of elements over Ω, then we have

$$\max_{m \geq 1} (|(\mathbf{E}^m - \mathbf{E}_{*h}^m)(x_c, y_c)| + |(H^m - H_{*h}^m)(x_c, y_c)|) \leq C(h^2 + \tau^2),$$

where $\mathbf{E}_{*h}^m$ and H_{*h}^m are the averaged values at the parallelogram centers:

$$\mathbf{E}_{*h}^m = \frac{1}{2}(\mathbf{E}_h^m|_L + \mathbf{E}_h^m|_R)(x_c, y_c), \quad H_{*h}^m = \frac{1}{2}(H_h^m|_L + H_h^m|_R)(x_c, y_c).$$

Chapter 6
A Posteriori Error Estimation

In this chapter, we present some basic techniques for developing a posteriori error estimation for solving Maxwell's equations. It is known that the a posteriori error estimation plays a very important role in adaptive finite element method. In Sect. 6.1, we provide a brief overview of a posteriori error estimation. Then in Sect. 6.2, through time-harmonic Maxwell's equations, we demonstrate the fundamental ideas on how to obtain the upper and lower posteriori error estimates. In Sect. 6.3, we present a posteriori error estimator obtained for a cold plasma model described by integro-differential Maxwell's equations.

6.1 A Brief Overview of A Posteriori Error Analysis

How to use the computational solution to guide where to refine or coarsen the local mesh grid and/or how to choose the proper orders of the basis function in different regions becomes an essential problem in the adaptive finite element method. Since the pioneering work of Babuska and Rheinboldt in the late 1970s [16], the adaptive finite element method has been well developed as evidenced by the vast literature in this area. If an error estimate for the unknown exact solution is totally based on the available computational result, then this error estimate is called **a posteriori error estimator**. How to develop a robust a posteriori error estimator plays an important role in developing an effective adaptive finite element method. Due to the intelligent work of many researchers over the past three decades, the study of a posteriori error estimator for standard elliptic, parabolic and second order hyperbolic problems seems mature (e.g., review papers [32,64,111,126,227], books [4, 20, 21, 252, 287, 297], and references cited therein).

On the other hand, works on a posteriori error estimators for Maxwell's equations are quite limited. The analysis of a posteriori error estimators for the edge elements was initiated by Monk in 1998 [216] and Beck et al. in 2000 [31]. So far, there are only about two dozens of papers devoted to Maxwell's equations in free space

J. Li and Y. Huang, *Time-Domain Finite Element Methods for Maxwell's Equations in Metamaterials*, Springer Series in Computational Mathematics 43,
DOI 10.1007/978-3-642-33789-5_6,

[48, 72, 82, 139, 147, 148, 159, 253, 305, 306]. For example, Monk [216] obtained a posteriori error estimator for a scattering problem interacting with a bounded inhomogeneous and anisotropic scatterer. Beck et al. [31] developed a residual-based a posteriori error estimator for the model problem (6.1) and (6.2) shown below. In this seminar paper, they obtained both the lower and upper bounds. In 2003, Nicaise et al. [225] considered residual-based a posteriori error estimator for the same model. Then in 2005, Nicaise [224] developed a posteriori Zienkiewicz-Zhu type error estimators for the same problem. Recently, Houston et al. [147] developed a posteriori error estimator for a mixed discontinuous Galerkin approximation of time-harmonic Maxwell's equations.

In the following two sections, we present details on those basic techniques of how to derive a posteriori error estimator for Maxwell's equations in free space and cold plasma, respectively. The rest of this chapter is mainly based on papers [72, 182].

6.2 A Posteriori Error Estimator for Free Space Model

6.2.1 Preliminaries

When discretizing the time-domain free space Maxwell's equations in time, we end up solving the following problem at each time step [31, 72]:

$$\nabla \times (\alpha(\mathbf{x})\nabla \times \mathbf{u}) + \beta(\mathbf{x})\mathbf{u} = \mathbf{f} \quad \text{in } \Omega, \tag{6.1}$$

$$\mathbf{u} \times \mathbf{n} = \mathbf{0} \quad \text{on } \partial\Omega, \tag{6.2}$$

where $\mathbf{u}$ is the time approximation of either the electric field $\mathbf{E}$ or the magnetic field $\mathbf{H}$, $\mathbf{f} \in H(\text{div}; \Omega)$ is a source function, while $\alpha(\mathbf{x})$ and $\beta(\mathbf{x})$ are the underlying medium parameters. Here for simplicity, we only consider the perfect conductor boundary condition (6.2). Throughout this chapter, Ω is assumed to be a bounded, simply-connected domain in R^3 with connected Lipschitz polyhedral boundary, whose unit outward normal vector is denoted as $\mathbf{n}$.

For simplicity, we assume that $\alpha(\mathbf{x})$ and $\beta(\mathbf{x})$ are piecewise positive constant functions on Ω, and Ω is composed of two disjoint subdomains Ω_1 and Ω_2. More specifically,

$$\alpha = \alpha_i \qquad \text{and } \beta = \beta_i \quad \text{in } \Omega_i,$$

where both Ω_1 and Ω_2 are simply-connected Lipschitz polyhedra.

It is easy to obtain a weak formulation of (6.1) and (6.2): Find $\mathbf{u} \in H_0(\text{curl}; \Omega)$ such that

$$a(\mathbf{u}, \mathbf{v}) = (\mathbf{f}, \mathbf{v}) \quad \forall\, \mathbf{v} \in H_0(\text{curl}; \Omega), \tag{6.3}$$

where the bilinear form $a(\mathbf{u}, \mathbf{v})$ is given by

$$a(\mathbf{u}, \mathbf{v}) = (\alpha\nabla \times \mathbf{u}, \nabla \times \mathbf{v}) + (\beta\mathbf{u}, \mathbf{v}) \quad \forall\, \mathbf{u}, \mathbf{v} \in H_0(\text{curl}; \Omega).$$

Recall that the space $H_0(\text{curl};\Omega)$ is defined as

$$H_0(\text{curl};\Omega) = \left\{\mathbf{u} \in (L^2(\Omega))^3 : \ \nabla \times \mathbf{u} \in (L^2(\Omega))^3 \text{ and } \mathbf{u} \times \mathbf{n} = \mathbf{0} \text{ on } \partial\Omega\right\}.$$

The well-posedness of (6.3) is guaranteed by the Lax-Milgram lemma.

To develop a finite element method for solving (6.3), we consider a shape-regular mesh T_h that partitions the domain Ω into disjoint tetrahedral elements $\{K\}$, such that $\overline{\Omega} = \bigcup_{K\in T_h} K$. Following the same notations defined in Sect. 4.2.2, we denote the diameter of K by h_K, the mesh size by $h = \max_{K\in T_h} h_K$, the set of all interior faces by F_h^I, the set of all boundary faces by F_h^B, and the set of all faces by $F_h = F_h^I \cup F_h^B$. Furthermore, we denote ω_K for the union of all elements K having a common face with K, and ω_F for the union of two elements sharing F.

With the above preparation, we can develop the finite element approximation of (6.3): Find $\mathbf{u} \in \mathbf{V}_h^0$ such that

$$a(\mathbf{u}_h, \mathbf{v}_h) = (\mathbf{f}, \mathbf{v}_h) \quad \forall\ \mathbf{v}_h \in \mathbf{V}_h^0, \tag{6.4}$$

where we use the lowest order edge element space (cf. Example 3.8):

$$\mathbf{V}_h^0 = \left\{\ \mathbf{v} \in H_0(\text{curl};\Omega) : \ \mathbf{v}|_K = \mathbf{a}_K \times \mathbf{x} + \mathbf{b}_K, \ \ \mathbf{a}_K, \mathbf{b}_K \in R^3, \ \forall\ K \in T_h\right\}.$$

Below are some fundamental results (cf. [145, 217]) needed for deriving a posterior error estimator.

Lemma 6.1. *The space $H_0(curl;\Omega)$ admits the following β-orthogonal decomposition*

$$H_0(curl;\Omega) = H_0^0(curl;\Omega) \oplus H_0^{\perp}(curl;\Omega), \tag{6.5}$$

where

$$H_0^0(curl;\Omega) \equiv \{\ \mathbf{v} \in H_0(curl;\Omega) : \ \nabla \times \mathbf{v} = \mathbf{0}\}$$

and

$$H_0^{\perp}(curl;\Omega) \equiv \left\{\ \mathbf{v} \in H_0(curl;\Omega) : \ (\beta\mathbf{v}, \mathbf{v}^0) = 0, \ \mathbf{v}^0 \in H_0^0(curl;\Omega)\right\}.$$

Lemma 6.2. *If the domain Ω is simply connected with connected boundary, we have*

$$H_0^0(curl;\Omega) = \nabla H_0^1(\Omega) \tag{6.6}$$

and

$$||\mathbf{v}||_0 \le C\,||\nabla \times \mathbf{v}||_0 \quad \forall\ \mathbf{v} \in H_0^{\perp}(curl;\Omega), \tag{6.7}$$

where the constant $C > 0$ depends on Ω only.

Lemma 6.3. *[145, Lemma 2.4] Assume that Ω is a bounded Lipschitz domain, then for any $\mathbf{v} \in H_0(curl;\Omega)$, there exists the regular decomposition*

$$\mathbf{v} = \mathbf{w} + \nabla\phi, \tag{6.8}$$

where $\mathbf{w} \in H_0(curl;\Omega) \cap (H^1(\Omega))^3$ and $\phi \in H_0^1(\Omega)$. Moreover, there is a positive constant C_{hip} depending only on Ω such that

$$||\mathbf{w}||_1 \le C_{hip}||\mathbf{v}||_{curl}, \quad ||\phi||_1 \le C_{hip}||\mathbf{v}||_{curl}, \tag{6.9}$$

here and below we define the norm $||\mathbf{v}||_{curl} = (||\mathbf{v}||_0^2 + ||\nabla \times \mathbf{v}||_0^2)^{1/2}$.

Lemma 6.4. *Let D_K (resp. D_F) denote the union of elements in T_h with non-empty intersection with K (resp. F). Furthermore, we denote a generic constant $C > 0$, which depends only on the shape regularity of the mesh.*

(i) [31] for any $\mathbf{w} \in H_0(curl;\Omega) \cap (H^1(\Omega))^3$, there exists the quasi-interpolation $\Pi_h\mathbf{w} \in \mathbf{V}_h^0$ such that

$$||\mathbf{w} - \Pi_h\mathbf{w}||_{0,K} \le Ch_K|\mathbf{w}|_{1,D_K} \quad \forall\ K \in T_h, \tag{6.10}$$

$$||\mathbf{w} - \Pi_h\mathbf{w}||_{0,F} \le Ch_F^{\frac{1}{2}}|\mathbf{w}|_{1,D_F}, \quad \forall\ F \in F_h. \tag{6.11}$$

(ii) [127, Sect. I.A.3] let S_0^h be the continuous piecewise linear finite element subspace of $H_0^1(\Omega)$. Then for any $\phi \in H_0^1(\Omega)$, there exists a continuous piecewise linear approximation $I_h\phi \in S_0^h$ such that

$$||\phi - I_h\phi||_{0,K} \le Ch_K|\phi|_{1,D_K} \quad \forall\ K \in T_h, \tag{6.12}$$

$$\beta_K^{\frac{1}{2}}||\phi - I_h\phi||_{0,K} \le Ch_K||\beta^{\frac{1}{2}}\nabla\phi||_{0,D_K} \quad \forall\ K \in T_h, \tag{6.13}$$

$$||\phi - I_h\phi||_{0,F} \le Ch_F^{\frac{1}{2}}|\phi|_{1,D_F} \quad \forall\ F \in F_h, \tag{6.14}$$

$$\beta_F^{\frac{1}{2}}||\phi - I_h\phi||_{0,F} \le Ch_F^{\frac{1}{2}}||\beta^{\frac{1}{2}}\nabla\phi||_{0,D_F} \quad \forall\ F \in F_h. \tag{6.15}$$

6.2.2 An Upper Bound of A Posterior Error Estimator

Before presenting the error estimate, we need to introduce some notations:

$$\Lambda_K^\alpha \equiv \frac{\alpha_K}{\alpha_m}, \quad \Lambda_F^\alpha \equiv \frac{\alpha_F}{\alpha_m}, \quad \Lambda_K^{\beta\alpha} \equiv \frac{\beta_K}{\alpha_m}, \quad \Lambda_F^{\beta\alpha} \equiv \frac{\beta_F}{\alpha_m}, \quad \forall\ K \in T_h,\ F \in F_h,$$

where $\alpha_m = \min\{\alpha_1, \alpha_2\}$. Furthermore, we define the element residuals

$$\mathbf{R}_K(\mathbf{u}_h) = \mathbf{f} - \beta\mathbf{u}_h \quad \text{in } K \in T_h,$$

and the jump residuals: for any $F \in F_h$,

$$\mathbf{J}_{F1}(\mathbf{u}_h) \equiv -[\alpha(\nabla \times \mathbf{u}_h) \times \mathbf{n}]_F \text{ and } \mathbf{J}_{F2}(\mathbf{u}_h) \equiv [(\mathbf{f} - \beta \mathbf{u}_h) \cdot \mathbf{n}]_F.$$

For simplicity, in the rest of this section, we write

$$[g(\mathbf{n})]_F = g(\mathbf{n}_F)|_{K_1} + g(\mathbf{n}_F)|_{K_2}$$

with $g(\mathbf{n})$ being either $\alpha(\nabla \times \mathbf{u}_h) \times \mathbf{n}$ or $(\mathbf{f} - \beta \mathbf{u}_h) \cdot \mathbf{n}$, and K_1 and K_2 are the two neighboring elements sharing the face F with unit outward normal vector $\mathbf{n}_F$. Furthermore, without confusion, we often use the short notation

$$\mathbf{R}_K = \mathbf{R}_K(\mathbf{u}_h), \ \mathbf{J}_{F1} = \mathbf{J}_{F1}(\mathbf{u}_h) \text{ and } \mathbf{J}_{F2} = \mathbf{J}_{F2}(\mathbf{u}_h).$$

Denote the solution error $\mathbf{e} = \mathbf{u} - \mathbf{u}_h$. Below we will bound the energy norm $||\mathbf{e}||_a = \sqrt{a(\mathbf{e}, \mathbf{e})}$ from above and below by the local error indicators

$$\eta_h^2(K) = \eta_{1,K}^2 + \sum_{F \subset \partial K, F \in F_h} \eta_{1,F}^2 + \eta_{2,K}^2 + \sum_{F \subset \partial K, F \in F_h} \eta_{2,F}^2, \tag{6.16}$$

where

$$\eta_{1,K}^2 = \Lambda_K^{\alpha} ||h_K \alpha_K^{-1/2} \mathbf{R}_K||_{0,K}^2, \quad \eta_{1,F}^2 = \Lambda_F^{\alpha} ||h_F^{1/2} \alpha_F^{-1/2} \mathbf{J}_{F1}||_{0,F}^2,$$
$$\eta_{2,K}^2 = \max(1, \Lambda_K^{\beta\alpha}) ||h_K \beta_K^{-1/2} \mathrm{div}\mathbf{f}||_{0,K}^2, \ \eta_{2,F}^2 = \max(1, \Lambda_F^{\beta\alpha}) ||h_F^{1/2} \beta_F^{-1/2} \mathbf{J}_{F2}||_{0,F}^2.$$

The upper bound of the error $\mathbf{u} - \mathbf{u}_h$ is given below.

Theorem 6.1.

$$||\mathbf{u} - \mathbf{u}_h||_a^2 \leq C_{up} \sum_{K \in T_h} \eta_h^2(K), \tag{6.17}$$

where the constant $C_{up} > 0$ depends only on the shape regularity of the mesh.

Proof. It is easy to see that the error $\mathbf{e} \in H_0(\mathrm{curl}; \Omega)$ satisfies the error equation

$$a(\mathbf{e}, \mathbf{v}) = r(\mathbf{v}) \quad \forall \ \mathbf{v} \in H_0(\mathrm{curl}; \Omega), \tag{6.18}$$

where the residual

$$r(\mathbf{v}) = (\mathbf{f} - \beta \mathbf{u}_h, \mathbf{v}) - (\alpha \ \nabla \times \mathbf{u}_h, \nabla \times \mathbf{v}) \quad \forall \ \mathbf{v} \in H_0(\mathrm{curl}; \Omega).$$

Using (6.4), we have the Galerkin orthogonality relation

$$a(\mathbf{e}, \mathbf{v}_h) = r(\mathbf{v}_h) = 0 \quad \forall \ \mathbf{v}_h \in \mathbf{V}_h^0. \tag{6.19}$$

Using the orthogonal decomposition (6.5) in the error equation (6.18), we have

$$||\mathbf{e}||_a^2 = r(\mathbf{e}) = r(\mathbf{e}^\perp) + r(\mathbf{e}^0), \tag{6.20}$$

where $\mathbf{e}^\perp \in H_0^\perp(\text{curl}; \Omega)$ and $\mathbf{e}^0 \in H_0^0(\text{curl}; \Omega)$. Then by the decomposition (6.8), we have

$$r(\mathbf{e}^\perp) = r(\mathbf{w}) + r(\nabla\phi), \tag{6.21}$$

where $\mathbf{w} \in H_0(\text{curl}; \Omega) \cap (H^1(\Omega))^3$ and $\phi \in H_0^1(\Omega)$.

The proof is completed by combining the estimates of $r(\mathbf{w}), r(\nabla\phi)$ and $r(\mathbf{e}^0)$ proved in Lemmas 6.5–6.7 shown below. □

Lemma 6.5.

$$r(\mathbf{w}) \le C\Big(\sum_{K\in T_h} \eta_{1,K}^2 + \sum_{F\in F_h} \eta_{1,F}^2\Big)^{1/2} ||\mathbf{e}^\perp||_a.$$

Proof. By the Galerkin orthogonality (6.19), we have

$$r(\mathbf{w}) = r(\mathbf{w} - \Pi_h\mathbf{w}) = (\mathbf{f} - \beta\mathbf{u}_h, \mathbf{w} - \Pi_h\mathbf{w}) - (\alpha\nabla\times\mathbf{u}_h, \nabla\times(\mathbf{w} - \Pi_h\mathbf{w})).$$

Then using integration by parts and the property $\nabla\times(\alpha\ \nabla\times\mathbf{u}_h) = \mathbf{0}$, we obtain

$$\begin{aligned}
r(\mathbf{w}) &= \sum_{K\in T_h} (\mathbf{f} - \beta\mathbf{u}_h - \nabla\times(\alpha\nabla\times\mathbf{u}_h), \mathbf{w} - \Pi_h\mathbf{w}) - \sum_{F\in F_h} ([\alpha\nabla\times\mathbf{u}_h\times\mathbf{n}]_F, \mathbf{w} - \Pi_h\mathbf{w})_F \\
&= \sum_{K\in T_h} (\mathbf{R}_K, \mathbf{w} - \Pi_h\mathbf{w}) + \sum_{F\in F_h} (\mathbf{J}_{F1}, \mathbf{w} - \Pi_h\mathbf{w})_F \\
&\le \sum_{K\in T_h} ||\mathbf{R}_K||_{0,K} ||\mathbf{w} - \Pi_h\mathbf{w}||_{0,K} + \sum_{F\in F_h} ||\mathbf{J}_{F1}||_{0,F} ||\mathbf{w} - \Pi_h\mathbf{w}||_{0,F} \\
&\le C[(\sum_{K\in T_h} h_K^2 ||\mathbf{R}_K||_{0,K}^2)^{1/2} |\mathbf{w}|_1 + (\sum_{F\in F_h} h_F ||\mathbf{J}_{F1}||_{0,F}^2)^{1/2} |\mathbf{w}|_1] \\
&\le C\Big(\sum_{K\in T_h} \eta_{1,K}^2 + \sum_{F\in F_h} \eta_{1,F}^2\Big)^{1/2} ||\alpha^{1/2}\nabla\times\mathbf{e}^\perp||_0,
\end{aligned}$$

which concludes the proof. In the above, we used the approximation property (6.10) and (6.11), and the estimate (6.9). □

Lemma 6.6.

$$r(\nabla\phi) \le C\Big(\sum_{K\in T_h} \eta_{2,K}^2 + \sum_{F\in F_h} \eta_{2,F}^2\Big)^{1/2} ||\mathbf{e}^\perp||_a.$$

Proof. It is known that (cf. (2.6) of [72]):

$$\mathbf{V}_h^0 \cap H_0^0(\text{curl}; \Omega) = \nabla S_0^h, \tag{6.22}$$

which implies that $\nabla\phi_h$ belongs to $\mathbf{V}_h^0$ for any $\phi_h \in S_0^h$.

By the Galerkin orthogonality (6.19), integration by parts and the fact that $\mathrm{div}(\beta \mathbf{u}_h) = 0$ on each element K, we have

$$
\begin{aligned}
r(\nabla\phi) &= r(\nabla(\phi - \Pi_h\phi)) = (\mathbf{f} - \beta\mathbf{u}_h, \nabla(\phi - \Pi_h\phi)) \\
&= \sum_{K\in T_h} (\mathrm{div}(\beta\mathbf{u}_h) - \mathrm{div}\mathbf{f}, \phi - \Pi_h\phi)_K + \sum_{F\in F_h} ([\mathbf{f} - \beta\mathbf{u}_h \cdot \mathbf{n}]_F, \phi - \Pi_h\phi)_F \\
&= \sum_{K\in T_h} (-\mathrm{div}\mathbf{f}, \phi - \Pi_h\phi)_K + \sum_{F\in F_h} (\mathbf{J}_{F_2}, \phi - \Pi_h\phi)_F \\
&\le \sum_{K\in T_h} ||\mathrm{div}\mathbf{f}||_{0,K} ||\phi - \Pi_h\phi||_{0,K} + \sum_{F\in F_h} ||\mathbf{J}_{F_2}||_{0,F} ||\phi - \Pi_h\phi||_{0,F} \\
&\le C[(\sum_{K\in T_h} h_K^2 ||\mathrm{div}\mathbf{f}||_{0,K}^2)^{\frac{1}{2}} |\phi|_1 + (\sum_{F\in F_h} h_F ||\mathbf{J}_{F_2}||_{0,F}^2)^{\frac{1}{2}} |\phi|_1] \\
&\le C[(\sum_{K\in T_h} h_K^2 ||\mathrm{div}\mathbf{f}||_{0,K}^2)^{\frac{1}{2}} ||\mathbf{e}^\perp||_0 + (\sum_{F\in F_h} h_F ||\mathbf{J}_{F_2}||_{0,F}^2)^{\frac{1}{2}} ||\mathbf{e}^\perp||_0] \\
&\le C[(\sum_{K\in T_h} \Lambda_K^{\beta\alpha} ||h_K \beta_K^{-\frac{1}{2}} \mathrm{div}\mathbf{f}||_{0,K}^2 + \sum_{F\in F_h} \Lambda_F^{\beta\alpha} ||h_F^{\frac{1}{2}} \beta_F^{-\frac{1}{2}} \mathbf{J}_{F2}||_{0,F}^2)^{\frac{1}{2}} ||\alpha^{\frac{1}{2}} \nabla \times \mathbf{e}^\perp||_0 \\
&\le C(\sum_{K\in T_h} \eta_{2,K}^2 + \sum_{F\in F_h} \eta_{2,F}^2)^{\frac{1}{2}} ||\alpha^{\frac{1}{2}} \nabla \times \mathbf{e}^\perp||_0,
\end{aligned}
$$

which concludes the proof. Here we used the approximation property (6.12)–(6.14), and the estimates (6.9) and (6.7). □

Lemma 6.7.

$$
r(\mathbf{e}^0) \le C(\sum_{K\in T_h} \eta_{2,K}^2 + \sum_{F\in F_h} \eta_{2,F}^2)^{1/2} ||\beta^{1/2}\mathbf{e}^0||_0.
$$

Proof. By (6.6), we know that there exists some $\psi \in H_0^1(\Omega)$ such that $\mathbf{e}^0 = \nabla\psi$. Hence similar to the proof of Lemma 6.6, we have

$$
\begin{aligned}
r(\mathbf{e}^0) &= r(\nabla(\psi - \Pi_h\psi)) \\
&\le C(\sum_{K\in T_h} ||h_K \beta_K^{-\frac{1}{2}} \mathrm{div}\mathbf{f}||_{0,K}^2 + \sum_{F\in F_h} ||h_F^{\frac{1}{2}} \beta_F^{-\frac{1}{2}} \mathbf{J}_{F2}||_{0,F}^2)^{\frac{1}{2}} ||\beta^{\frac{1}{2}} \nabla\psi||_0 \\
&\le C(\sum_{K\in T_h} \eta_{2,K}^2 + \sum_{F\in F_h} \eta_{2,F}^2)^{\frac{1}{2}} ||\beta^{\frac{1}{2}} \mathbf{e}^0||_0,
\end{aligned}
$$

which completes the proof. □

6.2.3 A Lower Bound of A Posterior Error Estimator

To obtain a lower bound of the error, we need to use the bubble function technique originally introduced by Verfurth [286] for the elliptic problem. We denote b_K for the standard polynomial bubble function on an element K, and b_F for the standard polynomial bubble function on an interior element face F, shared by two elements K and K'. For simplicity, in the following we denote $UF = \{K, K'\}$ for the union of elements K and K'. For a tetrahedron K, an exemplary element bubble function $b_K = 256\Pi_{i=1}^4 \lambda_i$, and face bubble function $b_F = 27\Pi_{i=1}^3 \lambda_i$, where λ_i is the standard basis function in S_0^h at vertex x_i.

With these notation, we have the following classical estimates.

Lemma 6.8. *For any polynomial function v on K, there exists a constant $C > 0$ independent of v and h_K such that*

$$||b_K v||_{0,K} \leq C||v||_{0,K}, \quad ||v||_{0,K} \leq C||b_K^{\frac{1}{2}} v||_{0,K}, \tag{6.23}$$

$$||\nabla(b_K v)||_{0,K} \leq Ch_K^{-1}||v||_{0,K}. \tag{6.24}$$

On the other hand, for any polynomial function w on F, there exists a constant $C > 0$ independent of w and h_F such that

$$||w||_{0,F} \leq C||b_F^{\frac{1}{2}} w||_{0,F}, \tag{6.25}$$

$$||Ex(b_F w)||_{0,K} \leq Ch_F^{\frac{1}{2}}||w||_{0,F} \quad \forall\ K \in UF, \tag{6.26}$$

$$||\nabla Ex(b_F w)||_{0,K} \leq Ch_F^{-\frac{1}{2}}||w||_{0,F} \quad \forall\ K \in UF, \tag{6.27}$$

where $Ex(b_F w) \in H_0^1((\overline{K} \cup \overline{K}')^\circ)$ is an extension of $b_F w$ such that $Ex(b_F w)|_F = b_F w$.

The same estimates as (6.23)–(6.27) hold true for vector functions. Moreover, for a vector ploynomial function $\mathbf{v}$ on K, there exists a constant $C > 0$ independent of $\mathbf{v}$ and h_K such that

$$||\nabla \times (b_K \mathbf{v})||_{0,K} \leq Ch_K^{-1}||\mathbf{v}||_{0,K}. \tag{6.28}$$

Similarly, for any vector polynomial function $\mathbf{w}$ on F, there exists a constant $C > 0$ independent of $\mathbf{w}$ and h_F such that

$$||\nabla \times Ex(b_F \mathbf{w})||_{0,K} \leq Ch_F^{-\frac{1}{2}}||\mathbf{w}||_{0,F} \quad \forall\ K \in UF, \tag{6.29}$$

where $Ex(b_F \mathbf{w}) \in H_0^1((\overline{K} \cup \overline{K}')^\circ)^3$ is an extension of $b_F \mathbf{w}$ such that $Ex(b_F \mathbf{w})|_F = b_F \mathbf{w}$.

Proof. The proof of (6.23), (6.25), and (6.26) can be found in [286, Lemma 4.1]. The proof of (6.24) and (6.27) can be obtained from Eqs. (2.35) and (2.39) of [4], respectively. The proof of (6.28) and (6.29) can be obtained by similar arguments as the proof of (6.24) and (6.27). □

Before deriving the lower bound of the error, let us introduce a few more notations. Let $\overline{\mathbf{R}}_K$ be the integral mean of $\mathbf{R}_K$ over element K, and $\overline{\text{div}\mathbf{f}}_K$ be the integral mean of $\text{div}\mathbf{f}_K$ over element K. Let $\Lambda^{\alpha}_{\omega_K} = \max_{K' \in \omega_K}(\Lambda^{\alpha}_{K'})$ and $\Lambda^{\beta\alpha}_{\omega_K} = \max_{K' \in \omega_K}(\Lambda^{\beta\alpha}_{K'}, 1)$.

First, we have the following lower bound for the local error estimator $\eta^2_{1,K}$.

Lemma 6.9.

$$\eta^2_{1,K} \leq C\Lambda^{\alpha}_K[||\alpha^{\frac{1}{2}} \nabla \times (\mathbf{u} - \mathbf{u}_h)||^2_{0,K} + h^2_K \beta_K \alpha_K^{-1} ||\beta^{\frac{1}{2}}(\mathbf{u} - \mathbf{u}_h)||^2_{0,K} + h^2_K \alpha_K^{-1} ||\overline{\mathbf{R}}_K - \mathbf{R}_K||^2_{0,K}]. \tag{6.30}$$

Proof. Using (6.23), the facts that $b_K \overline{\mathbf{R}}_K \in H^1_0(\Omega)^3$ and $\nabla \times (\alpha \nabla \times \mathbf{u}_h) = 0$, and integration by parts, we have

$$\begin{aligned} C||\overline{\mathbf{R}}_K||^2_{0,K} &\leq (\overline{\mathbf{R}}_K, b_K \overline{\mathbf{R}}_K)_K = (\mathbf{R}_K, b_K \overline{\mathbf{R}}_K)_K + (\overline{\mathbf{R}}_K - \mathbf{R}_K, b_K \overline{\mathbf{R}}_K)_K \\ &= (\mathbf{f} - \beta \mathbf{u}_h - \nabla \times (\alpha \nabla \times \mathbf{u}_h), b_K \overline{\mathbf{R}}_K)_K + (\overline{\mathbf{R}}_K - \mathbf{R}_K, b_K \overline{\mathbf{R}}_K)_K \\ &= (\alpha \nabla \times (\mathbf{u} - \mathbf{u}_h), \nabla \times (b_K \overline{\mathbf{R}}_K))_K + (\beta(\mathbf{u} - \mathbf{u}_h), b_K \overline{\mathbf{R}}_K)_K \\ &\quad + (\overline{\mathbf{R}}_K - \mathbf{R}_K, b_K \overline{\mathbf{R}}_K)_K \\ &\leq C[\alpha^{\frac{1}{2}} h_K^{-1} ||\alpha^{\frac{1}{2}} \nabla \times (\mathbf{u} - \mathbf{u}_h)||_{0,K} + \beta_K^{\frac{1}{2}} ||\beta^{\frac{1}{2}}(\mathbf{u} - \mathbf{u}_h)||_{0,K} \\ &\quad + ||\overline{\mathbf{R}}_K - \mathbf{R}_K||_{0,K}]||\overline{\mathbf{R}}_K||_{0,K}, \end{aligned}$$

where in the last step we used the standard inverse estimate and (6.23).

Combining the above estimate with the triangle inequality, we obtain

$$\begin{aligned} ||\mathbf{R}_K||_{0,K} &\leq ||\overline{\mathbf{R}}_K||_{0,K} + ||\mathbf{R}_K - \overline{\mathbf{R}}_K||_{0,K} \\ &\leq C[\alpha^{\frac{1}{2}} h_K^{-1} ||\alpha^{\frac{1}{2}} \nabla \times (\mathbf{u} - \mathbf{u}_h)||_{0,K} + \beta_K^{\frac{1}{2}} ||\beta^{\frac{1}{2}}(\mathbf{u} - \mathbf{u}_h)||_{0,K} \\ &\quad + ||\overline{\mathbf{R}}_K - \mathbf{R}_K||_{0,K}]. \end{aligned} \tag{6.31}$$

Recall the definition of $\eta_{1,K}$, we have

$$\begin{aligned} \eta^2_{1,K} &= \Lambda^{\alpha}_K h^2_K \alpha_K^{-1} ||\mathbf{R}_K||^2_{0,K} \\ &\leq C\Lambda^{\alpha}_K[||\alpha^{\frac{1}{2}} \nabla \times (\mathbf{u} - \mathbf{u}_h)||^2_{0,K} + h^2_K \beta_K \alpha_K^{-1} ||\beta^{\frac{1}{2}}(\mathbf{u} - \mathbf{u}_h)||^2_{0,K} \\ &\quad + h^2_K \alpha_K^{-1} ||\overline{\mathbf{R}}_K - \mathbf{R}_K||^2_{0,K}], \end{aligned}$$

which completes the proof. □

For the local error estimator $\eta_{1,F}^2$, we have the following lower bound.

Lemma 6.10.

$$\eta_{1,F}^2 \leq C\Lambda_F^\alpha[\sum_{K\in\omega_F}||\alpha^{\frac{1}{2}}\nabla\times(\mathbf{u}-\mathbf{u}_h)||_{0,K}^2+\sum_{K\in\omega_F}h_K^2\beta_K\alpha_K^{-1}||\beta^{\frac{1}{2}}(\mathbf{u}-\mathbf{u}_h)||_{0,K}^2 + \sum_{K\in\omega_F}h_K^2\alpha_K^{-1}||\mathbf{R}_K-\overline{\mathbf{R}}_K||_{0,K}^2]. \tag{6.32}$$

Proof. Using (6.25), the fact $\nabla\times(\alpha\,\nabla\times\mathbf{u}_h)=0$, and integration by parts, we have

$$\begin{aligned}
&C||\mathbf{J}_{F1}||_{0,F}^2 \leq (\mathbf{J}_{F1}, b_F\mathbf{J}_{F1})_F = -([\alpha\,\nabla\times\mathbf{u}_h\times\mathbf{n}]_F, b_F\mathbf{J}_{F1})_F\\
&= \sum_{K\in\omega_F}(\nabla\times(\alpha\,\nabla\times\mathbf{u}_h), b_F\mathbf{J}_{F1})_K - \sum_{K\in\omega_F}(\alpha\,\nabla\times\mathbf{u}_h, \nabla\times(b_F\mathbf{J}_{F1}))_K\\
&= r(b_F\mathbf{J}_{F1}) - \sum_{K\in\omega_F}(\mathbf{R}_K, b_F\mathbf{J}_{F1})_K\\
&= \sum_{K\in\omega_F}[(\alpha\,\nabla\times(\mathbf{u}-\mathbf{u}_h), \nabla\times(b_F\mathbf{J}_{F1}))_K - (\beta(\mathbf{u}-\mathbf{u}_h), b_F\mathbf{J}_{F1})_K]\\
&\quad - \sum_{K\in\omega_F}(\mathbf{R}_K, b_F\mathbf{J}_{F1})_K\\
&\leq C[\sum_{K\in\omega_F}\alpha_K^{\frac{1}{2}}h_K^{-1}||\alpha^{\frac{1}{2}}\nabla\times(\mathbf{u}-\mathbf{u}_h)||_{0,K}||b_F\mathbf{J}_{F1}||_{0,K}\\
&\quad + \sum_{K\in\omega_F}\beta_K^{\frac{1}{2}}||\beta^{\frac{1}{2}}(\mathbf{u}-\mathbf{u}_h)||_{0,K}||b_F\mathbf{J}_{F1}||_{0,K} + \sum_{K\in\omega_F}||\mathbf{R}_K||_{0,K}||b_F\mathbf{J}_{F1}||_{0,K}]\\
&\leq C[\sum_{K\in\omega_F}\alpha_K^{\frac{1}{2}}h_K^{-\frac{1}{2}}||\alpha^{\frac{1}{2}}\nabla\times(\mathbf{u}-\mathbf{u}_h)||_{0,K} + \sum_{K\in\omega_F}h_K^{\frac{1}{2}}\beta_K^{\frac{1}{2}}||\beta^{\frac{1}{2}}(\mathbf{u}-\mathbf{u}_h)||_{0,K}\\
&\quad + \sum_{K\in\omega_F}h_K^{\frac{1}{2}}||\mathbf{R}_K-\overline{\mathbf{R}}_K||_{0,K}]||\mathbf{J}_{F1}||_{0,F}],
\end{aligned}$$

where in the above derivation we used the standard inverse estimate, estimates (6.31) and (6.27).

Hence by the definion of $\eta_{1,F}$, we obtain

$$\begin{aligned}
\eta_{1,F}^2 &= \Lambda_F^\alpha h_F\alpha_F^{-1}||\mathbf{J}_{F1}||_{0,F}^2\\
&\leq C\Lambda_F^\alpha[\sum_{K\in\omega_F}||\alpha^{\frac{1}{2}}\nabla\times(\mathbf{u}-\mathbf{u}_h)||_{0,K}^2 + \sum_{K\in\omega_F}h_K^2\beta_K\alpha_K^{-1}||\beta^{\frac{1}{2}}(\mathbf{u}-\mathbf{u}_h)||_{0,K}^2\\
&\quad + \sum_{K\in\omega_F}h_K^2\alpha_K^{-1}||\mathbf{R}_K-\overline{\mathbf{R}}_K||_{0,K}^2],
\end{aligned}$$

which concludes the proof. □

For the local error estimator $\eta_{2,K}^2$, we have the following lower bound.

Lemma 6.11.

$$\eta_{2,K}^2 \leq C \max(\Lambda_K^{\beta\alpha}, 1)[||\beta^{\frac{1}{2}}(\mathbf{u}-\mathbf{u}_h)||_{0,K}^2 + h_K^2\beta_K^{-1}||\overline{div\mathbf{f}}_K - div\mathbf{f}||_{0,K}^2].$$

Proof. Similar to the proof of (6.31), by using the facts that

$$\mathrm{div}(\beta\mathbf{u}_h) = 0 \ \text{ and } \ \mathrm{div}\mathbf{f} = \mathrm{div}(\beta\mathbf{u}),$$

we easily have

$$\begin{aligned} C||\overline{\mathrm{div}\mathbf{f}}_K||_{0,K}^2 &\leq (\overline{\mathrm{div}\mathbf{f}}_K, b_K\overline{\mathrm{div}\mathbf{f}}_K)_K \\ &= (\mathrm{div}(\mathbf{f}-\beta\mathbf{u}_h), b_K\overline{\mathrm{div}\mathbf{f}}_K)_K + (\overline{\mathrm{div}\mathbf{f}}_K - \mathrm{div}\mathbf{f}, b_K\overline{\mathrm{div}\mathbf{f}}_K)_K \\ &= -(\beta(\mathbf{u}-\mathbf{u}_h), \nabla(b_K\overline{\mathrm{div}\mathbf{f}}_K))_K + (\overline{\mathrm{div}\mathbf{f}}_K - \mathrm{div}\mathbf{f}, b_K\overline{\mathrm{div}\mathbf{f}}_K)_K \\ &\leq C[\beta_K^{\frac{1}{2}}h_K^{-1}||\beta^{\frac{1}{2}}(\mathbf{u}-\mathbf{u}_h)||_{0,K} + ||\overline{\mathrm{div}\mathbf{f}}_K - \mathrm{div}\mathbf{f}||_{0,K}]||\overline{\mathrm{div}\mathbf{f}}_K||_{0,K}, \end{aligned}$$

which, along with the triangle inequality, leads to

$$||\mathrm{div}\mathbf{f}||_{0,K} \leq C[\beta_K^{\frac{1}{2}}h_K^{-1}||\beta^{\frac{1}{2}}(\mathbf{u}-\mathbf{u}_h)||_{0,K} + ||\overline{\mathrm{div}\mathbf{f}}_K - \mathrm{div}\mathbf{f}||_{0,K}].$$

Recall the definition of $\eta_{2,K}^2$, we obtain

$$\eta_{2,K}^2 \leq C \max(\Lambda_K^{\beta\alpha}, 1)[||\beta^{\frac{1}{2}}(\mathbf{u}-\mathbf{u}_h)||_{0,K}^2 + h_K^2\beta_K^{-1}||\overline{\mathrm{div}\mathbf{f}}_K - \mathrm{div}\mathbf{f}||_{0,K}^2],$$

which concludes the proof. □

Finally, we can prove the following lower bound for the local error estimator $\eta_{2,F}^2$.

Lemma 6.12.

$$\eta_{2,F}^2 \leq C \max(\Lambda_F^{\beta\alpha}, 1)[\sum_{K\in\omega_F} ||\beta^{\frac{1}{2}}(\mathbf{u}-\mathbf{u}_h)||_{0,K}^2 + \sum_{K\in\omega_F} h_K^2\beta_K^{-1}||div\mathbf{f} - \overline{div\mathbf{f}}_K||_{0,K}^2]. \tag{6.33}$$

Proof. Applying the extension operator Ex to the jump $\mathbf{J}_{F_2}$, we obtain

$$\begin{aligned} C||\mathbf{J}_{F2}||_{0,F}^2 &\leq (\mathbf{J}_{F2}, b_F Ex(\mathbf{J}_{F2}))_F = ([(\mathbf{f}-\beta\mathbf{u}_h)\cdot\mathbf{n}]_F, b_F Ex(\mathbf{J}_{F2}))_F \\ &= (\mathbf{f}-\beta\mathbf{u}_h, \nabla(b_F Ex(\mathbf{J}_{F2})))_{\omega_F} + (\mathrm{div}(\mathbf{f}-\beta\mathbf{u}_h), b_F Ex(\mathbf{J}_{F2}))_{\omega_F} \\ &= (\beta(\mathbf{f}-\mathbf{u}_h), \nabla(b_F Ex(\mathbf{J}_{F2})))_{\omega_F} + (\mathrm{div}\mathbf{f}, b_F Ex(\mathbf{J}_{F2}))_{\omega_F} \\ &\leq C[\sum_{K\in\omega_F} \beta_K^{\frac{1}{2}}h_K^{-\frac{1}{2}}||\beta^{\frac{1}{2}}(\mathbf{f}-\mathbf{u}_h)||_{0,K} + \sum_{K\in\omega_F} h_K^{\frac{1}{2}}||\mathrm{div}\mathbf{f}||_{0,K}]||\mathbf{J}_{F2}||_{0,F}, \end{aligned}$$

where we used (6.26). Using the estimate of div$\mathbf{f}$, we have

$$||\mathbf{J}_{F2}||_{0,F} \leq C[\sum_{K\in\omega_F} \beta_K^{\frac{1}{2}} h_K^{-\frac{1}{2}} ||\beta^{\frac{1}{2}}(\mathbf{f}-\mathbf{u}_h)||_{0,K} + \sum_{K\in\omega_F} h_K^{\frac{1}{2}} ||\mathrm{div}\mathbf{f} - \overline{\mathrm{div}\mathbf{f}}_K||_{0,K}],$$

from which and the definition of $\eta_{2,F}^2$ we obtain

$$\begin{aligned}\eta_{2,F}^2 &= \max(\Lambda_F^{\beta\alpha}, 1) h_F^{-1}\beta_F^{-1}||\mathbf{J}_{F2}||_{0,F}^2 \\ &\leq C\max(\Lambda_F^{\beta\alpha}, 1)[\sum_{K\in\omega_F} ||\beta^{\frac{1}{2}}(\mathbf{u}-\mathbf{u}_h)||_{0,K}^2 + \sum_{K\in\omega_F} h_K^2\beta_K^{-1}||\mathrm{div}\mathbf{f} - \overline{\mathrm{div}\mathbf{f}}_K||_{0,K}^2],\end{aligned}$$

which completes the proof. □

Combining Lemmas 6.9–6.12, we obtain the following lower bound of a posterior error estimator.

Theorem 6.2.

$$\begin{aligned}\sum_{K\in T_h} \eta_h^2(K) \leq C_{low} \sum_{K\in T_h} &[\Lambda_{\omega_K}^{\alpha} ||\alpha^{\frac{1}{2}}\nabla\times(\mathbf{u}-\mathbf{u}_h)||_{0,\omega_K}^2 \\ &+\Lambda_{\omega_K}^{\alpha} \sum_{K'\in\omega_K} h_{K'}^2\beta_{K'}\alpha_{K'}^{-1}||\beta^{\frac{1}{2}}(\mathbf{u}-\mathbf{u}_h)||_{0,K'}^2 \\ &+\Lambda_{\omega_K}^{\beta\alpha} \sum_{K'\in\omega_K} ||\beta^{\frac{1}{2}}(\mathbf{u}-\mathbf{u}_h)||_{0,K'}^2 + \Lambda_{\omega_K}^{\alpha} \sum_{K'\in\omega_K} h_{K'}^2\alpha_{K'}^{-1}||\mathbf{R}_{K'} - \overline{\mathbf{R}}_{K'}||_{0,K'}^2 \\ &+\Lambda_{\omega_K}^{\beta\alpha} \sum_{K'\in\omega_K} h_{K'}^2\beta_{K'}^{-1}||div\mathbf{f} - \overline{div\mathbf{f}}_{K'}||_{0,K'}^2].\end{aligned}$$

6.2.4 Zienkiewicz-Zhu Error Estimator

Another simple and effective posterior error estimator is the so-called Zienkiewicz-Zhu (ZZ) estimator introduced by Zienkiewicz and Zhu [309] and later improved by many researchers (cf. [172, 296, 298] and references cited therein). The basic idea is to use some post-processing procedure to compute an improved gradient of the numerical solution first, then use the difference between this recovered gradient and the original gradient for the error estimator. In practical implementation, a gradient (or flux) is often computed, hence it is cheap to implement the ZZ error estimator. Moreover, the estimator has been proved to be very robust for a variety of problems, and has been quite popular. In this section, we present a nice Zienkiewicz-Zhu error estimator obtained by Nicaise [224] for the Maxwell's equations (6.1) and (6.2).

We denote $\mathcal{N}$ the set of all (interior or boundary) nodes of T_h, $\omega_{\mathbf{x}}$ the union of all elements sharing node $\mathbf{x}$, and the jump of a function $\mathbf{v}$ across a face F as:

$$[|\mathbf{v}(\mathbf{y})|] = \lim_{\epsilon \to +0} (\mathbf{v}(\mathbf{y} + \epsilon \mathbf{n}_F) - \mathbf{v}(\mathbf{y} - \epsilon \mathbf{n}_F)), \quad \mathbf{y} \in F,$$

where $\mathbf{n}_F$ is the unit outward vector to F.

Before we define a ZZ type recovered operator, let us first recall the barycentric coordinate $\lambda_{\mathbf{x}}$ at any node $\mathbf{x}$ defined in Chap. 2, i.e., $\lambda_{\mathbf{x}}$ is a continuous piecewise linear function on T_h such that

$$\lambda_{\mathbf{x}}(\mathbf{y}) = \delta_{\mathbf{x},\mathbf{y}}, \quad \forall\, \mathbf{y} \in \mathcal{N},$$

where $\delta_{\mathbf{x},\mathbf{y}} = 1$ if $\mathbf{x} = \mathbf{y}$, and 0 otherwise. Moreover, let us denote W_h the space of piecewise linear vector fields on T_h, and $V_h = W_h \cap C(\Omega, \mathscr{R}^3)$.

With the above notation, a ZZ type recovered operator $R_{ZZ} : W_h \to V_h$ can be defined by [224]: $\mathbf{v}_h \to \sum_{\mathbf{x} \in \mathcal{N}} (R_{ZZ}\mathbf{v}_h)(\mathbf{x}) \lambda_{\mathbf{x}}$, where

$$(R_{ZZ}\mathbf{v}_h)(\mathbf{x}) = \sum_{K \in \omega_{\mathbf{x}}} \mu_{K,\mathbf{x}} \mathbf{v}_h|_K(\mathbf{x}), \quad \mathbf{x} \in \mathcal{N}, \tag{6.34}$$

where $\mu_{K,\mathbf{x}} \geq 0$ are the weights, which can be freely chosen such that $\sum_{K \in \omega_{\mathbf{x}}} \mu_{K,\mathbf{x}} = 1$. Furthermore, the local and global ZZ estimators are defined as:

$$\eta_{Z,K}^2 = ||R_{ZZ}\mathbf{u}_h - \mathbf{u}_h||_{0,K}^2 + ||R_{ZZ}(\mathrm{curl}_h \mathbf{u}_h) - \mathrm{curl}_h \mathbf{u}_h||_{0,K}^2,$$
$$\eta_Z^2 = \sum_{K \in T_h} \eta_{Z,K}^2,$$

where curl_h is calculated elementwisely.

Nicaise [224] proved that the above defined ZZ estimator is equivalent to a residual type error estimator. Furthermore, both lower and upper bounds for the ZZ estimators are obtained.

Theorem 6.3. *For problem (6.1) and (6.2), the error* $\mathbf{u} - \mathbf{u}_h$ *is bounded locally from below and globally from above:*

$$\eta_{Z,K} \leq C[||\mathbf{u} - \mathbf{u}_h||_{H(curl;\omega_K)} + \sum_{K' \subset \omega_K} \xi_{K'}],$$
$$||\mathbf{u} - \mathbf{u}_h||_{H(curl;\Omega)} \leq C[\eta_Z + \eta_{el} + \xi],$$

where

$$\xi_K^2 = h_K^2 ||r_K - R_K||_{0,K}^2, \quad \xi^2 = \sum_{K \in T_h} \xi_K^2, \quad \eta_{el}^2 = \sum_{K \in T_h} h_K^2 ||r_K||_{0,K}^2.$$

Here R_K is the exact residual defined by

$$R_K = f - (\nabla \times (\alpha \nabla \times \mathbf{u}_h) + \beta \mathbf{u}_h) \quad \forall\ K \in T_h,$$

and r_K is the corresponding approximated residual.

The proof of Theorem 6.3 is quite technical, interested readers can consult the original paper [224, Theorem 3.9].

6.3 A Posteriori Error Estimator for Cold Plasma Model

In this section, we develop a posteriori error estimator for a semi-discrete DG scheme used to solve the cold plasma model discussed in Sect. 4.2.2. For simplicity, we assume that Ω is partitioned into disjoint tetrahedral elements $\{K\}$ such that $\overline{\Omega} = \bigcup_{K \in T_h} K$. Hence the according finite element space is given by

$$\mathbf{V}_h = \{\mathbf{v} \in (L^2(\Omega))^3 : \mathbf{v}|_K \in (P_l(K))^3,\ K \in T_h\},\ \ l \geq 1, \tag{6.35}$$

Note that all results below hold true for a mesh of affine hexahedral elements, in which case on each element K, $\mathbf{v}|_K$ is a polynomial of degree at most l in each variable.

To simplify the presentation, we assume that all physical parameters in the governing equation (4.5) are one (i.e., $C_v = \nu = \omega_p = 1$) and adding a source term $\mathbf{f}$ to the right hand side of (4.5), in which case the governing equation is simplified as:

$$\mathbf{E}_{tt} + \nabla \times \nabla \times \mathbf{E} + \mathbf{E} - \mathbf{J}(\mathbf{E}) = \mathbf{f}, \tag{6.36}$$

where the polarization current density $\mathbf{J}$ is

$$\mathbf{J}(\mathbf{E}) \equiv \mathbf{J}(\mathbf{x}, t; \mathbf{E}) = \int_0^t e^{-(t-s)} \mathbf{E}(\mathbf{x}, s) ds. \tag{6.37}$$

We can form a semi-discrete DG scheme for (6.36): For any $t \in (0, T)$, find $\mathbf{E}^h(\cdot, t) \in \mathbf{V}_h$ such that

$$(\mathbf{E}^h_{tt}, \phi) + a_h(\mathbf{E}^h, \phi) - (\mathbf{J}(\mathbf{E}^h), \phi) = (\mathbf{f}, \phi),\ \ \forall\ \phi \in \mathbf{V}_h, \tag{6.38}$$

subject to the initial conditions

$$\mathbf{E}^h|_{t=0} = \Pi_2 \mathbf{E}_0, \quad \mathbf{E}^h_t|_{t=0} = \Pi_2 \mathbf{E}_1, \tag{6.39}$$

where Π_2 denotes the standard L_2-projection onto $\mathbf{V}_h$. Moreover, the bilinear form a_h is defined on $\mathbf{V}_h \times \mathbf{V}_h$ as

$$a_h(\mathbf{u},\mathbf{v}) = \sum_{K\in T_h}\int_K (\nabla\times\mathbf{u}\cdot\nabla\times\mathbf{v}+\mathbf{u}\cdot\mathbf{v})d\mathbf{x} - \sum_{F\in F_h}\int_F [[\mathbf{u}]]_T\cdot\{\{\nabla\times\mathbf{v}\}\}dA$$
$$- \sum_{F\in F_h}\int_F [[\mathbf{v}]]_T\cdot\{\{\nabla\times\mathbf{u}\}\}dA + \sum_{F\in F_h}\int_F a[[\mathbf{u}]]_T\cdot[[\mathbf{v}]]_T dA.$$

Here $[[\mathbf{v}]]$ and $\{\{\mathbf{v}\}\}$ are the standard notation for the tangential jumps and averages of $\mathbf{v}$ across interior faces defined in Sect. 4.2.2. Finally, a is a penalty function, which is defined on each face $F \in F_h$ as:

$$a|_F = \gamma\hbar^{-1},$$

where $\hbar|_F = \min\{h_{K^+}, h_{K^-}\}$ for an interior face $F = \partial K^+ \cap \partial K^-$, and $\hbar|_F = h_K$ for a boundary face $F = \partial K \cap \partial\Omega$. The penalty parameter γ is a positive constant.

Following Sect. 4.2.2, we denote the space $\mathbf{V}(h) = H_0(\text{curl};\Omega) + \mathbf{V}_h$ and define the DG energy norm by

$$||\mathbf{v}||_h^2 = ||\mathbf{v}||_{0,\Omega}^2 + \sum_{K\in T_h}||\nabla\times\mathbf{v}||_{0,K}^2 + \sum_{F\in F_h}||a^{1/2}[[\mathbf{v}]]_T||_{0,F}^2.$$

In order to carry out the posteriori analysis, we introduce an auxiliary bilinear form $\tilde{a}_h$ on $\mathbf{V}(h)\times\mathbf{V}(h)$ defined as

$$\tilde{a}_h(\mathbf{u},\mathbf{v}) = \sum_{K\in T_h}\int_K (\nabla\times\mathbf{u}\cdot\nabla\times\mathbf{v}+\mathbf{u}\cdot\mathbf{v})d\mathbf{x} - \sum_{K\in T_h}\int_K \mathscr{L}(\mathbf{u})\cdot(\nabla\times\mathbf{v})dx$$
$$- \sum_{K\in T_h}\int_K \mathscr{L}(\mathbf{v})\cdot(\nabla\times\mathbf{u})dx + \sum_{F\in F_h}\int_F a[[\mathbf{u}]]_T\cdot[[\mathbf{v}]]_T dA,$$

where the lifting operator $\mathscr{L}(\mathbf{v}) \in \mathbf{V}_h$ for any $\mathbf{v} \in \mathbf{V}(h)$ is defined by

$$\int_\Omega \mathscr{L}(\mathbf{v})\cdot\mathbf{w}dx = \sum_{F\in F_h}\int_F [[\mathbf{v}]]_T\cdot\{\{\mathbf{w}\}\}dA \quad \forall\ \mathbf{w}\in\mathbf{V}_h, \tag{6.40}$$

from which it is easy to see that the lifting operator $\mathscr{L}(\mathbf{v})$ can be bounded as follows [148]:

$$||\mathscr{L}(\mathbf{v})||_{0,\Omega}^2 \le \alpha^{-1}C_{lift}\sum_{F\in F_h}||a^{\frac{1}{2}}[[\mathbf{v}]]_T||_{0,F}^2. \tag{6.41}$$

In the rest two subsections, we present detailed derivation of upper and lower bounds of the posteriori error estimator.

6.3.1 Upper Bound of the Posteriori Error Estimator

One of the main tools used in the posteriori error estimate for DG methods is to find a conforming finite element function close to the discontinuous one. For this purpose, we define the conforming finite element space

$$\mathbf{V}_h^c = \mathbf{V}_h \cap H_0(\text{curl}; \Omega), \tag{6.42}$$

i.e., $\mathbf{V}_h^c$ is the second family of Nédélec element [223]. Moreover, we have the following approximation property [148].

Lemma 6.13. *For any $\mathbf{v}^h \in \mathbf{V}_h$, there exists a conforming approximation $\mathbf{v}_c^h \in \mathbf{V}_h^c$ such that*

$$\sum_{K\in T_h} ||\nabla \times (\mathbf{v}^h - \mathbf{v}_c^h)||_{0,K}^2 \le C_{app} \sum_{F\in F_h} h_F^{-1} ||[[\mathbf{v}^h]]_T||_{0,F}^2,$$

$$||\mathbf{v}^h - \mathbf{v}_c^h||_{0,\Omega}^2 \le C_{app} \sum_{F\in F_h} h_F ||[[\mathbf{v}^h]]_T||_{0,F}^2,$$

and

$$||\mathbf{v}^h - \mathbf{v}_c^h||_h^2 \le (2\alpha^{-1} C_{app} + 1) \sum_{F\in F_h} ||a^{\frac{1}{2}}[[\mathbf{v}^h]]_T||_{0,F}^2,$$

where the constant $C_{app} > 0$ depends only on the shape regularity of the mesh and the approximation order l in space $\mathbf{V}_h$.

Before we state the posteriori error estimator, let us introduce some local error indicators. Let $\mathbf{f}_h \in \mathbf{V}_h$ be some approximation of $\mathbf{f}$, and

$$\eta_{R_K}^2 = h_K^2 ||\mathbf{f}_h - \mathbf{E}_{tt}^h - \nabla \times \nabla \times \mathbf{E}^h - \mathbf{E}^h + \mathbf{J}(\mathbf{E}^h)||_{0,K}^2,$$

which measures the residual of the approximated governing Maxwell's equations (6.36).

We denote

$$\eta_{T_K}^2 = \frac{1}{2} \sum_{F\in\partial K\setminus\Gamma} h_F ||[[\nabla \times \mathbf{E}^h]]_T||_{0,F}^2$$

for the face residual about the jump of $\nabla \times \mathbf{E}^h$.

To measure the tangential jumps of the approximate solution $\mathbf{E}^h$, we denote

$$\eta_{J_K}^2 = \frac{1}{2} \sum_{F\in\partial K\setminus\Gamma} ||a^{\frac{1}{2}}[[\mathbf{E}^h]]_T||_{0,F}^2.$$

Noting that $\nabla \cdot \nabla \times (\nabla \times \mathbf{E}^h) = 0$, hence

$$\eta_{D_K}^2 = h_K^2 ||\nabla \cdot (\mathbf{f}_h - \mathbf{E}_{tt}^h - \mathbf{E}^h + \mathbf{J}(\mathbf{E}^h))||_{0,K}^2$$

measures the error in the divergence of the governing Maxwell's equations (6.36).

Furthermore, we denote

$$\eta_{N_K}^2 = \frac{1}{2}\sum_{F\in\partial K\setminus\Gamma} h_K ||[[\mathbf{f}_h - \mathbf{E}_{tt}^h - \mathbf{E}^h + \mathbf{J}(\mathbf{E}^h)]]_N||_{0,F}^2$$

for measuring the normal jump of $\mathbf{f}_h - \mathbf{E}_{tt}^h - \mathbf{E}^h + \mathbf{J}(\mathbf{E}^h)$ over the interior faces.

Similarly, we can define the following local estimators:

$$\eta_{R_K^t}^2 = h_K^2 ||(\mathbf{f}_h - \mathbf{E}_{tt}^h - \nabla\times\nabla\times\mathbf{E}^h - \mathbf{E}^h + \mathbf{J}(\mathbf{E}^h))_t||_{0,K}^2,$$

$$\eta_{T_K^t}^2 = \frac{1}{2}\sum_{F\in\partial K\setminus\Gamma} h_F ||[[\nabla\times\mathbf{E}_t^h]]_T||_{0,F}^2,$$

$$\eta_{J_K^t}^2 = \frac{1}{2}\sum_{F\in\partial K\setminus\Gamma} ||a^{\frac{1}{2}}[[\mathbf{E}_t^h]]_T||_{0,F}^2,$$

$$\eta_{D_K^t}^2 = h_K^2 ||\nabla\cdot(\mathbf{f}_h - \mathbf{E}_{tt}^h - \mathbf{E}^h + \mathbf{J}(\mathbf{E}^h))_t||_{0,K}^2,$$

$$\eta_{N_K^t}^2 = \frac{1}{2}\sum_{F\in\partial K\setminus\Gamma} h_K ||[[(\mathbf{f}_h - \mathbf{E}_{tt}^h - \mathbf{E}^h + \mathbf{J}(\mathbf{E}^h))_t]]_N||_{0,F}^2.$$

Theorem 6.4. *Let* $\mathbf{E}$ *be the solution of (6.36) and* $\mathbf{E}^h$ *be the DG solution of (6.38) with* $\gamma \geq \gamma_{min}$. *Then the following estimation holds:*

$$\begin{aligned}
&||\mathbf{E}-\mathbf{E}^h||_h^2(t) + ||(\mathbf{E}-\mathbf{E}^h)_t||_h^2(t)\\
\leq\ & C[||\mathbf{E}-\mathbf{E}^h||_h^2(0) + ||(\mathbf{E}-\mathbf{E}^h)_t||_h^2(0)]\\
&+C\int_0^t \sum_{F\in F_h} h_F(||[[\mathbf{E}_{tt}^h]]_T||_{0,F}^2 + ||[[\mathbf{E}_t^h]]_T||_{0,F}^2 + ||[[\mathbf{E}^h]]_T||_{0,F}^2)dt\\
&+C\sum_{F\in F_h}[||a^{\frac{1}{2}}[[\mathbf{E}^h]]_T||_{0,F}^2(t) + ||a^{\frac{1}{2}}[[\mathbf{E}_t^h]]_T||_{0,F}^2(t)\\
&+||a^{\frac{1}{2}}[[\mathbf{E}^h]]_T||_{0,F}^2(0) + ||a^{\frac{1}{2}}[[\mathbf{E}_t^h]]_T||_{0,F}^2(0)]\\
&+C[||\mathbf{f}-\mathbf{f}_h||_{0,\Omega}^2(t) + \sum_{K\in T_h}(\eta_{R_K}^2 + \eta_{T_K}^2 + \eta_{J_K}^2 + \eta_{D_K}^2 + \eta_{N_K}^2)(t)]\\
&+C[||\mathbf{f}-\mathbf{f}_h||_{0,\Omega}^2(0) + \sum_{K\in T_h}(\eta_{R_K}^2 + \eta_{T_K}^2 + \eta_{J_K}^2 + \eta_{D_K}^2 + \eta_{N_K}^2)(0)]\\
&+C\int_0^t[\sum_{K\in T_h}(\eta_{R_K^t}^2 + \eta_{T_K^t}^2 + \eta_{J_K^t}^2 + \eta_{D_K^t}^2 + \eta_{N_K^t}^2) + ||(\mathbf{f}-\mathbf{f}_h)_t||_0^2]dt.
\end{aligned}$$

Proof. Denote $\mathbf{w} = \mathbf{E} - \mathbf{E}_c^h \in H_0(\text{curl}; \Omega)$, where $\mathbf{E}_c^h$ is the conforming approximation of $\mathbf{E}^h$. Then for any $\phi \in H_0(\text{curl}; \Omega)$, we have

$$\begin{aligned}(\mathbf{w}_{tt}, \phi) + \tilde{a}_h(\mathbf{w}, \phi) &= ((\mathbf{E} - \mathbf{E}^h + \mathbf{E}^h - \mathbf{E}_c^h)_{tt}, \phi) + \tilde{a}_h(\mathbf{E} - \mathbf{E}^h + \mathbf{E}^h - \mathbf{E}_c^h, \phi) \\ &= (\mathbf{E}_{tt} - \mathbf{E}_{tt}^h, \phi) + \tilde{a}_h(\mathbf{E}, \phi) - \tilde{a}_h(\mathbf{E}^h, \phi) \\ &\quad + ((\mathbf{E}^h - \mathbf{E}_c^h)_{tt}, \phi) + \tilde{a}_h(\mathbf{E}^h - \mathbf{E}_c^h, \phi). \end{aligned} \tag{6.43}$$

Using the fact that $\tilde{a}_h(\mathbf{u}, \mathbf{v}) = \int_\Omega (\nabla \times \mathbf{u} \cdot \nabla \times \mathbf{v} + \mathbf{u} \cdot \mathbf{v}) dx$ on $H_0(\text{curl}; \Omega) \times H_0(\text{curl}; \Omega)$, we can write the weak formulation of (6.36) as: Find $\mathbf{E} \in H_0(\text{curl}; \Omega)$ such that

$$(\mathbf{E}_{tt}, \phi) + \tilde{a}_h(\mathbf{E}, \phi) - (\mathbf{J}(\mathbf{E}), \phi) = (\mathbf{f}, \phi) \quad \forall\, \phi \in H_0(\text{curl}; \Omega). \tag{6.44}$$

Using the fact that $\tilde{a}_h = a_h$ on $\mathbf{V}_h \times \mathbf{V}_h$, we can rewrite the semi-discrete scheme (6.38) as

$$(\mathbf{E}_{tt}^h, \phi_h) + \tilde{a}_h(\mathbf{E}^h, \phi_h) - (\mathbf{J}(\mathbf{E}^h), \phi_h) = (\mathbf{f}, \phi_h), \quad \forall\, \phi_h \in \mathbf{V}_h. \tag{6.45}$$

From (6.44) and (6.45), we have

$$\begin{aligned} &(\mathbf{E}_{tt} - \mathbf{E}_{tt}^h, \phi) + \tilde{a}_h(\mathbf{E}, \phi) - \tilde{a}_h(\mathbf{E}^h, \phi) \\ &= (\mathbf{f} + \mathbf{J}(\mathbf{E}) - \mathbf{E}_{tt}^h, \phi) - \tilde{a}_h(\mathbf{E}^h, \phi_h) - \tilde{a}_h(\mathbf{E}^h, \phi - \phi_h) \\ &= (\mathbf{f} + \mathbf{J}(\mathbf{E}^h) - \mathbf{E}_{tt}^h, \phi - \phi_h) + (\mathbf{J}(\mathbf{E} - \mathbf{E}^h), \phi) - \tilde{a}_h(\mathbf{E}^h, \phi - \phi_h), \end{aligned}$$

substituting which into (6.43), we obtain

$$\begin{aligned}(\mathbf{w}_{tt}, \phi) + \tilde{a}_h(\mathbf{w}, \phi) &= (\mathbf{f} + \mathbf{J}(\mathbf{E}^h) - \mathbf{E}_{tt}^h, \phi - \phi_h) + (\mathbf{J}(\mathbf{w} + \mathbf{E}_c^h - \mathbf{E}^h), \phi) \\ &\quad - \tilde{a}_h(\mathbf{E}^h, \phi - \phi_h) + ((\mathbf{E}^h - \mathbf{E}_c^h)_{tt}, \phi) + \tilde{a}_h(\mathbf{E}^h - \mathbf{E}_c^h, \phi). \end{aligned} \tag{6.46}$$

Choosing $\phi = \mathbf{w}_t$ in (6.46), then integrating both sides from 0 to t, and multiplying both sides by 2, we obtain

$$||\mathbf{w}(t)||_h^2 + ||\mathbf{w}_t(t)||_0^2 \le ||\mathbf{w}(0)||_h^2 + ||\mathbf{w}_t(0)||_0^2 + \sum_{i=1}^{5} Err_i. \tag{6.47}$$

With careful estimates of all $Err_i, i = 1, \cdots, 5$ (cf. [182]), we have

$$
\begin{aligned}
&||\mathbf{w}(t)||_h^2 + ||\mathbf{w}_t(t)||_h^2 \\
&\le C(||\mathbf{w}(0)||_h^2 + ||\mathbf{w}_t(0)||_h^2) + C\int_0^t (||\mathbf{w}(t)||_h^2 + ||\mathbf{w}_t(t)||_h^2)dt \\
&+ C\int_0^t \sum_{F\in F_h} h_F(||[[\mathbf{E}_{tt}^h]]_T||_{0,F}^2 + ||[[\mathbf{E}_t^h]]_T||_{0,F}^2 + ||[[\mathbf{E}^h]]_T||_{0,F}^2)dt \\
&+ \delta_1||\mathbf{w}(t)||_h^2 + \frac{C}{\delta_1}\sum_{F\in F_h} ||a^{\frac{1}{2}}[[\mathbf{E}^h]]_T||_{0,F}^2 + C\sum_{F\in F_h} ||a^{\frac{1}{2}}[[\mathbf{E}^h(0)]]_T||_{0,F}^2 \\
&+ \delta_2||\mathbf{w}(t)||_{curl}^2 + \frac{C}{\delta_2}[||\mathbf{f}-\mathbf{f}_h||_{0,\Omega}^2(t) + \sum_{K\in T_h}(\eta_{R_K}^2 + \eta_{T_K}^2 + \eta_{J_K}^2 + \eta_{D_K}^2 + \eta_{N_K}^2)(t)] \\
&+ \delta_3||\mathbf{w}(0)||_{curl}^2 + \frac{C}{\delta_3}[||\mathbf{f}-\mathbf{f}_h||_{0,\Omega}^2(0) + \sum_{K\in T_h}(\eta_{R_K}^2 + \eta_{T_K}^2 + \eta_{J_K}^2 + \eta_{D_K}^2 + \eta_{N_K}^2)(0)] \\
&+ C\int_0^t [\sum_{K\in T_h}(\eta_{R_K^t}^2 + \eta_{T_K^t}^2 + \eta_{J_K^t}^2 + \eta_{D_K^t}^2 + \eta_{N_K^t}^2) + ||(\mathbf{f}-\mathbf{f}_h)_t||_0^2]dt. \qquad (6.48)
\end{aligned}
$$

By the definition of $||\cdot||_h$ and Lemma 6.13, we easily have

$$||(\mathbf{E}^h - \mathbf{E}_c^h)(t)||_h^2 \le C\sum_{F\in F_h} ||a^{\frac{1}{2}}[[\mathbf{E}^h]]_T||_{0,F}^2,$$

and

$$||(\mathbf{E}^h - \mathbf{E}_c^h)_t(t)||_h^2 \le C\sum_{F\in F_h} ||a^{\frac{1}{2}}[[\mathbf{E}_t^h]]_T||_{0,F}^2,$$

which, along with (6.48), the triangle inequality, and the Gronwall inequality (choosing δ_1 and δ_2 small enough), concludes the proof. □

6.3.2 *Lower Bound of the Local Error Estimator*

Theorem 6.5. *Let* $\mathbf{E}$ *be the solution of (6.36) and* $\mathbf{E}^h$ *be the DG solution of (6.38) with* $\gamma \ge \gamma_{min}$. *Then the following local bounds hold:*

$$
\begin{aligned}
(i)\quad \eta_{R_K} &\le C[h_K||(\mathbf{E}-\mathbf{E}^h)_{tt}||_{0,K} + h_K||\mathbf{E}-\mathbf{E}^h||_{0,K} + h_K\int_0^t ||\mathbf{E}-\mathbf{E}^h||_{0,K}(s)ds \\
&\quad + h_K||\mathbf{f}_h - \mathbf{f}||_{0,K} + ||\nabla\times(\mathbf{E}-\mathbf{E}^h)||_{0,K}], \\
(ii)\quad \eta_{T_K} &\le C\sum_{K\in UF}[h_K||(\mathbf{E}-\mathbf{E}^h)_{tt}||_{0,K} + h_K||\mathbf{E}-\mathbf{E}^h||_{0,K} \\
&\quad + h_K\int_0^t ||\mathbf{E}^h - \mathbf{E}||_{0,K}(s)ds + h_K||\mathbf{f}_h - \mathbf{f}||_{0,K} + ||\nabla\times(\mathbf{E}^h - \mathbf{E})||_{0,K}],
\end{aligned}
$$

$$(iii)\quad \eta_{D_K} \le C(||\mathbf{f}_h - \mathbf{f}||_{0,K} + ||(\mathbf{E}-\mathbf{E}^h)_{tt}||_{0,K} + ||\mathbf{E}-\mathbf{E}^h||_{0,K} + \int_0^t ||\mathbf{E}^h - \mathbf{E}||_{0,K}(s)ds),$$

$$(iv)\quad \eta_{N_K} \le C \sum_{K\in UF} (||\mathbf{f}_h - \mathbf{f}||_{0,K} + ||(\mathbf{E}-\mathbf{E}^h)_{tt}||_{0,K} + ||\mathbf{E}-\mathbf{E}^h||_{0,K} + \int_0^t ||\mathbf{E}^h - \mathbf{E}||_{0,K}(s)ds).$$

Proof. To give readers some ideas about how to prove these lower bounds, below we just show the proofs of (i) and (iv). Proofs of the rest can be found in the original paper [182].

(i) Let $\mathbf{v}_h = \mathbf{f}_h - \mathbf{E}^h_{tt} - \nabla\times\nabla\times\mathbf{E}^h - \mathbf{E}^h + \mathbf{J}(\mathbf{E}^h)$, and $\mathbf{v}_b = b_K\mathbf{v}_h$.
Using the governing equation (6.36), we have

$$\begin{aligned}
||b_K^{\frac{1}{2}}\mathbf{v}_h||^2_{0,K} &= \int_K (\mathbf{f}_h - \mathbf{E}^h_{tt} - \nabla\times\nabla\times\mathbf{E}^h - \mathbf{E}^h + \mathbf{J}(\mathbf{E}^h))\cdot\mathbf{v}_b dx \\
&= \int_K [(\mathbf{E}-\mathbf{E}^h)_{tt} + \nabla\times\nabla\times(\mathbf{E}-\mathbf{E}^h) + (\mathbf{E}-\mathbf{E}^h) - \mathbf{J}(\mathbf{E}-\mathbf{E}^h)]\cdot\mathbf{v}_b dx \\
&\quad + \int_K (\mathbf{f}_h - \mathbf{f})\cdot\mathbf{v}_b dx \\
&= \int_K [(\mathbf{E}-\mathbf{E}^h)_{tt} + (\mathbf{E}-\mathbf{E}^h) - \mathbf{J}(\mathbf{E}-\mathbf{E}^h) + (\mathbf{f}_h - \mathbf{f})]\cdot\mathbf{v}_b dx \\
&\quad + \int_K (\nabla\times(\mathbf{E}-\mathbf{E}^h))\cdot(\nabla\times\mathbf{v}_b)dx,
\end{aligned}$$

where in the last step we used integration by parts and the fact that $\mathbf{v}_b = 0$ on ∂K.

Then by Lemma 6.8, we have

$$\begin{aligned}
||\mathbf{v}_h||_{0,K} \le C[&||(\mathbf{E}-\mathbf{E}^h)_{tt}||_{0,K} + ||\mathbf{E}-\mathbf{E}^h||_{0,K} + \int_0^t ||\mathbf{E}-\mathbf{E}^h||_{0,K}(s)ds \\
&+||\mathbf{f}_h - \mathbf{f}||_{0,K} + h_K^{-1}||\nabla\times(\mathbf{E}-\mathbf{E}^h)||_{0,K}],
\end{aligned}$$

which leads to

$$\begin{aligned}
\eta_{R_K} \le C[&h_K||(\mathbf{E}-\mathbf{E}^h)_{tt}||_{0,K} + h_K||\mathbf{E}-\mathbf{E}^h||_{0,K} + h_K\int_0^t ||\mathbf{E}-\mathbf{E}^h||_{0,K}(s)ds \\
&+h_K||\mathbf{f}_h - \mathbf{f}||_{0,K} + ||\nabla\times(\mathbf{E}-\mathbf{E}^h)||_{0,K}],
\end{aligned}$$

which completes the proof of (i).

(iv) Let $\mathbf{v}_h = [[\mathbf{f}_h - \mathbf{E}^h_{tt} - \mathbf{E}^h + \mathbf{J}(\mathbf{E}^h)]]_N$, and $\mathbf{v}_b = b_F\mathbf{v}_h$. Using the facts that $[[\mathbf{f} - \mathbf{E}_{tt} - \mathbf{E} + \mathbf{J}(\mathbf{E})]]_N = 0$ on interior faces and $\nabla\cdot(\mathbf{f} - \mathbf{E}_{tt} - \mathbf{E} + \mathbf{J}(\mathbf{E})) = 0$ in K, we have

$$
\begin{aligned}
||b_F^{\frac{1}{2}}\mathbf{v}_h||_{0,F}^2 &= \int_F [[\mathbf{f}_h - \mathbf{E}_{tt}^h - \mathbf{E}^h + \mathbf{J}(\mathbf{E}^h)]]_N \cdot \mathbf{v}_b ds \\
&= \int_F [[\mathbf{f}_h - \mathbf{f} + (\mathbf{E} - \mathbf{E}^h)_{tt} + \mathbf{E} - \mathbf{E}^h + \mathbf{J}(\mathbf{E}^h - \mathbf{E})]]_N \cdot \mathbf{v}_b ds \\
&= \sum_{K \in UF} \int_K \nabla \cdot (\mathbf{f}_h - \mathbf{E}_{tt}^h - \mathbf{E}^h + \mathbf{J}(\mathbf{E}^h))\mathbf{v}_b dx \\
&\quad + \sum_{K \in UF} \int_K (\mathbf{f}_h - \mathbf{f} + (\mathbf{E} - \mathbf{E}^h)_{tt} + \mathbf{E} - \mathbf{E}^h + \mathbf{J}(\mathbf{E}^h - \mathbf{E})) \cdot \nabla \mathbf{v}_b dx \\
&\leq C \sum_{K \in UF} h_K^{-1} \eta_{D_K} ||\mathbf{v}_b||_{0,K} + \sum_{K \in UF} [||\mathbf{f}_h - \mathbf{f}||_{0,K} + ||(\mathbf{E} - \mathbf{E}^h)_{tt}||_{0,K} \\
&\quad + ||\mathbf{E} - \mathbf{E}^h||_{0,K} + \int_0^t ||\mathbf{E}^h - \mathbf{E}||_{0,K}(s) ds]] |\nabla \mathbf{v}_b||_{0,K}.
\end{aligned}
$$

Using Lemma 6.4 and the estimate (iii), we have

$$
\begin{aligned}
&\eta_{N_K} = h_F^{\frac{1}{2}} ||\mathbf{v}_h||_{0,F} \\
&\leq C \sum_{K \in UF} (\eta_{D_K} + ||\mathbf{f}_h - \mathbf{f}||_{0,K} + ||(\mathbf{E} - \mathbf{E}^h)_{tt}||_{0,K} \\
&\quad + ||\mathbf{E} - \mathbf{E}^h||_{0,K} + \int_0^t ||\mathbf{E}^h - \mathbf{E}||_{0,K}(s) ds) \\
&\leq C \sum_{K \in UF} (||\mathbf{f}_h - \mathbf{f}||_{0,K} + ||(\mathbf{E} - \mathbf{E}^h)_{tt}||_{0,K} + ||\mathbf{E} - \mathbf{E}^h||_{0,K} + \int_0^t ||\mathbf{E}^h - \mathbf{E}||_{0,K}(s) ds),
\end{aligned}
$$

which concludes the proof of (iv). □

Chapter 7
A Matlab Edge Element Code for Metamaterials

In this chapter, we demonstrate the practical implementation of a mixed finite element method (FEM) for a 2-D Drude metamaterial model (5.1)–(5.4).

Let us recall that the basic procedures of using FEM to solve a partial differential equation (PDE):

1. Discretize the computational domain into finite elements;
2. Rewrite the PDE in a weak formulation, then choose proper finite element spaces and form the finite element scheme from the weak formulation;
3. Calculate those element matrices on each element;
4. Assemble element matrices to form a global linear system;
5. Implement the boundary conditions and solve the linear system;
6. Postprocess the numerical solution.

Compared to many books on finite element programming [21, 62, 158, 240, 252], there are only several books devoted to Maxwell's equations [42, 97, 98, 141, 162, 267]. To our best knowledge, no existing book provides complete source codes for solving time-domain Maxwell's equations using edge elements. Hence, in this chapter, we will present implementation details on using edge elements to solve the Drude metamaterial model (5.1)–(5.4). More specifically, in Sect. 7.1, we present a simple grid-generation algorithm and its implementation. Section 7.2 formulates the finite element scheme for the Drude model (5.1)–(5.4). In Sect. 7.3, we discuss how to calculate those element matrices involved. Then in Sect. 7.4 we discuss the finite element assembly procedure and how to implement the Dirichlet boundary condition. Since edge elements do not yield numerical solutions at mesh nodes automatically, in Sect. 7.5 we present a postprocessing step to retrieve the numerical solutions at element centers. Finally, in Sect. 7.6 we present an example problem to show how our algorithm gets implemented in MATLAB. Detailed MATLAB source codes with many comments are provided. We summarize this chapter in Sect. 7.7.

J. Li and Y. Huang, *Time-Domain Finite Element Methods for Maxwell's Equations in Metamaterials*, Springer Series in Computational Mathematics 43,
DOI 10.1007/978-3-642-33789-5_7, © Springer-Verlag Berlin Heidelberg 2013

7.1 Mesh Generation

For simplicity, we assume that the physical domain is a rectangle $\Omega \equiv [lowx, highx] \times [lowy, highy]$, which is subdivided into $nelex \times neley$ uniform rectangular elements. Here $nelex$ and $neley$ denote the numbers of elements in the x and y directions, respectively. A simple MATLAB code below accomplishes this task, where the x and y coordinates of all nodes are stored in the first and second rows of array $no2xy(1:2, 1:np)$, respectively, where np denotes the total number of points in the mesh.

```
dx=(highx-lowx)/nelex; dy=(highy-lowy)/neley;

nx = nelex+1; ny = neley+1;
np = (nelex+1)*(neley+1);  % total # of grid points
no2xy = zeros(2,np);
for j=1:ny
    for i=1:nx
        ipt=nx*(j-1)+i;
        no2xy(1,ipt)=dx*(i-1);
        no2xy(2,ipt)=dy*(j-1);
     end
end
```

Similar to the classical nodal based finite element method, we need to build up a connectivity matrix $el2no(i, j)$ to describe the relation between local nodes and global nodes. For the lowest-order rectangular edge element, $el2no(i, j)$ denotes the global label of the i-th node of the j-th element, where $i = 1, 2, 3, 4$, $j = 1, \cdots, numel$, and $numel$ denotes the total number of elements. For consistency, the four nodes of each element are ordered counterclockwise. This task is achieved by the following MATLAB code.

```
numel=(nelex)*(neley);  % total number of elements
el2no=zeros(4,numel);

idx=1;
for i=1:neley
  for j=1:nelex
     el2no(:,idx)=[j+(i-1)*nx; j+(i-1)*nx+1; \ldots
                   j+nx*i+1; j+nx*i];
     idx = idx+1;
  end
end
```

Since unknowns in edge element space are associated with edges in the mesh, we need to number the edges and associate an orientation direction with each edge. To do this, we assume that each edge is defined by its start and end points, and each

edge is assigned a global edge number. This task can be done efficiently based on a sorting technique originally proposed by Jin [162, p. 332] and implemented for triangular edge elements in MATLAB [42, p. 125]. Below is our implementation to create an array $el2ed(i, j)$, which stores the i-th edge of the j-th element, where $i = 1, \cdots, 4$, and $j = 1, \cdots, numel$.

```
% the total number of edges including boundary edges
numed=nelex*(ny)+neley*(nx);
for i=1:numel
   for j=1:4
      if (j==1 | j==2 | j==3)
         edges((i-1)*4+j,:)=[el2no(j,i) el2no(j+1,i)];
      else
         edges((i-1)*4+j,:)=[el2no(j,i) el2no(1,i)];
      end
   end
end

edges=sort(edges,2);
[ed2no,trash,el2ed]=unique(edges,'rows');
el2ed=reshape(el2ed,4,numel);
```

The complete MATLAB source code *create_mesh.m* is shown below:

```
function create_mesh

globals2D;

% give the rectangle info
lowx=0;  highx=1.0; lowy=0;  highy=1.0;
nelex=20; neley=20;
dx=(highx-lowx)/nelex; dy=(highy-lowy)/neley;

% generate a rectangular mesh
nx = nelex+1; % number of points in the x direction
ny = neley+1; % number of points in the y direction
np = nx*ny; % total number of grid points
no2xy = zeros(2,np);
for j=1:ny
    for i=1:nx
        ipt=nx*(j-1)+i;
        no2xy(1,ipt)=dx*(i-1);   no2xy(2,ipt)=dy*(j-1);
    end
end

numel=(nelex)*(neley);  % the number of total elements
% 4 nodes (counterclockwise) for each element!
el2no=zeros(4,numel);
idx=1;
for i=1:neley % number of columns to go through
    for j=1:nelex
```

```
        el2no(:,idx)=[j+(i-1)*nx; j+(i-1)*nx+1; ...
                      j+nx*i+1; j+nx*i];
        idx = idx+1;
    end
end

% the total number of edges including boundary edges
numed=nelex*(ny)+neley*(nx);
for i=1:numel
    for j=1:4 % for each element in each column
        if (j==1 | j==2 | j==3)
            edges((i-1)*4+j,:)=[el2no(j,i) el2no(j+1,i)];
        else
            edges((i-1)*4+j,:)=[el2no(j,i) el2no(1,i)];
        end
    end
end
edges=sort(edges,2);
[ed2no,trash,el2ed]=unique(edges,'rows');
el2ed=reshape(el2ed,4,numel);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Indicators: 1 for interior edges; 0 for boundary edges.
ed_id=zeros(numed,1);
for i=1:numed
    v1x = no2xy(1,ed2no(i,1));
    v1y = no2xy(2,ed2no(i,1));
    v2x = no2xy(1,ed2no(i,2));
    v2y = no2xy(2,ed2no(i,2));
    if (v1x==lowx & v2x==lowx) | (v1x==highx & v2x==highx) |...
            (v1y==lowy & v2y==lowy) | (v1y==highy & v2y==highy)
        ed_id(i)=1;
    end
end

eint = find(ed_id == 0); % get labels for all interior edges
iecnt = length(eint);    % total number of interior edges

% compare reference element edge directions vs global
% edge directions to get the orientations for all edges
%        3
%   ----<------
%   |          |
% 4 v          ^  2
%   |          |
%   ----->-----
%         1
edori = ones(numel,4);
for i=1:numel
 %   edori(i,:)=[1 1 -1 -1];
    for j=1:4
        edn = el2ed(j,i);
        n1=ed2no(edn,1); n2=ed2no(edn,2);
        if j < 4
```

```
          m1=el2no(j,i); m2=el2no(j+1,i);
        else
          m1=el2no(j,i); m2=el2no(1,i);
        end
        if m1 > m2
            edori(i,j)=-1;
        end
    end
end
return
```

7.2 The Finite Element Scheme

For the non-dimensionalized Drude model derived in Sect. 4.4 with added source terms $\mathbf{f}$ and $\mathbf{g}$, we can obtain its weak formulation: For any $t \in (0, T]$, find the solutions $\mathbf{E} \in H_0(\text{curl}; \Omega)$, $\mathbf{J} \in H(\text{curl}; \Omega)$, H and $K \in L^2(\Omega)$ such that

$$(\mathbf{E}_t, \boldsymbol{\phi}) - (\mathbf{H}, \nabla \times \boldsymbol{\phi}) + (\mathbf{J}, \boldsymbol{\phi}) = (\mathbf{f}, \boldsymbol{\phi}), \quad \forall\, \boldsymbol{\phi} \in H_0(\text{curl}; \Omega), \tag{7.1}$$

$$(H_t, \psi) + (\nabla \times \mathbf{E}, \psi) + (K, \psi) = (g, \psi), \quad \forall\, \psi \in L^2(\Omega), \tag{7.2}$$

$$(\mathbf{J}_t, \tilde{\boldsymbol{\phi}}) + \Gamma_e(\mathbf{J}, \tilde{\boldsymbol{\phi}}) - \omega_e^2(\mathbf{E}, \tilde{\boldsymbol{\phi}}) = 0, \quad \forall\, \tilde{\boldsymbol{\phi}} \in H(\text{curl}; \Omega), \tag{7.3}$$

$$(K_t, \tilde{\psi}) + \Gamma_m(K, \tilde{\psi}) - \omega_m^2(H, \tilde{\psi}) = 0, \quad \forall\, \tilde{\psi} \in L^2(\Omega), \tag{7.4}$$

subject to the perfect conducting boundary condition (3.59) and initial conditions (3.60) and (3.61).

To construct a finite element scheme for (7.1)–(7.4), we first discretize the physical domain Ω into rectangular elements $K \in T_h$. On this mesh T_h, we construct the lowest-order Raviart-Thomas-Nédélec finite element spaces:

$$\mathbf{U}_h = \{u_h \in L^2(\Omega) : \ u_h|_K \in Q_{0,0}, \quad \forall\, K \in T_h\}, \tag{7.5}$$

$$\mathbf{V}_h = \{\mathbf{v}_h \in H(\text{curl}; \Omega) : \ \mathbf{v}_h|_K \in Q_{0,1} \times Q_{1,0}, \quad \forall\, K \in T_h\}. \tag{7.6}$$

To take care of the perfect conducting boundary condition (3.59), we introduce a subspace of $\mathbf{V}_h$:

$$\mathbf{V}_h^0 = \{\mathbf{v}_h \in \mathbf{V}_h : \ \mathbf{v}_h \times \mathbf{n} = \mathbf{0} \ \text{ on } \partial\Omega\}.$$

Similar to (5.16)–(5.19), we can formulate a Crank-Nicolson mixed finite element scheme for solving (7.1)–(7.4): For $k \geq 1$, find $\mathbf{E}_h^k \in \mathbf{V}_h^0, \mathbf{J}_h^k \in \mathbf{V}_h, H_h^k, K_h^k \in \mathbf{U}_h$ such that

$$(\delta_\tau \mathbf{E}_h^k, \boldsymbol{\phi}_h) - (\overline{\mathbf{H}}_h^k, \nabla \times \boldsymbol{\phi}_h) + (\overline{\mathbf{J}}_h^k, \boldsymbol{\phi}_h) = (\mathbf{f}^{k-\frac{1}{2}}, \boldsymbol{\phi}_h), \ \forall\, \boldsymbol{\phi}_h \in \mathbf{V}_h^0, \tag{7.7}$$

$$(\delta_\tau H_h^k, \psi_h) + (\nabla \times \overline{\mathbf{E}}_h^k, \psi_h) + (\overline{K}_h^k, \psi_h) = (g^{k-\frac{1}{2}}, \psi_h), \ \forall\, \psi_h \in \mathbf{U}_h, \tag{7.8}$$

$$(\delta_\tau \mathbf{J}_h^k, \tilde{\boldsymbol{\phi}}_h) + \Gamma_e(\overline{\mathbf{J}}_h^k, \tilde{\boldsymbol{\phi}}_h) - \omega_e^2(\overline{\mathbf{E}}_h^k, \tilde{\boldsymbol{\phi}}_h) = 0, \quad \forall\ \tilde{\boldsymbol{\phi}}_h \in \mathbf{V}_h, \tag{7.9}$$

$$(\delta_\tau K_h^k, \tilde{\psi}_h) + \Gamma_m(\overline{K}_h^k, \tilde{\psi}_h) - \omega_m^2(\overline{H}_h^k, \tilde{\psi}_h) = 0, \quad \forall\ \tilde{\psi}_h \in \mathbf{U}_h, \tag{7.10}$$

subject to the initial approximations

$$\mathbf{E}_h^0(\mathbf{x}) = \Pi_h \mathbf{E}_0(\mathbf{x}), \quad H_h^0(\mathbf{x}) = P_h H_0(\mathbf{x}), \tag{7.11}$$

$$\mathbf{J}_h^0(\mathbf{x}) = \Pi_h \mathbf{J}_0(\mathbf{x}), \quad K_h^0(\mathbf{x}) = P_h K_0(\mathbf{x}). \tag{7.12}$$

As usual, we denote P_h for the standard $L^2(\Omega)$-projection operator onto $\mathbf{U}_h$, and Π_h for the Nédélec interpolation operator.

In practical implementation, we first solve (7.9) and (7.10) for $\mathbf{J}_h^k$ and K_h^k:

$$\mathbf{J}_h^k = \frac{2-\tau\Gamma_e}{2+\tau\Gamma_e}\mathbf{J}_h^{k-1} + \frac{\tau\omega_e^2}{2+\tau\Gamma_e}(\mathbf{E}_h^k + \mathbf{E}_h^{k-1}), \tag{7.13}$$

$$K_h^k = \frac{2-\tau\Gamma_m}{2+\tau\Gamma_m}K_h^{k-1} + \frac{\tau\omega_m^2}{2+\tau\Gamma_m}(H_h^k + H_h^{k-1}) \tag{7.14}$$

then substituting (7.13) and (7.14) into (7.7) and (7.8), respectively, we obtain

$$\begin{aligned}
(i) \quad & (1 + \frac{\tau^2\omega_e^2}{2(2+\tau\Gamma_e)})(\mathbf{E}_h^k, \boldsymbol{\phi}_h) - \frac{\tau}{2}(H_h^k, \nabla\times\boldsymbol{\phi}_h) \\
&= (1 - \frac{\tau^2\omega_e^2}{2(2+\tau\Gamma_e)})(\mathbf{E}_h^{k-1}, \boldsymbol{\phi}_h) + \frac{\tau}{2}(H_h^{k-1}, \nabla\times\boldsymbol{\phi}_h) \\
&\quad -\frac{2\tau}{2+\tau\Gamma_e}(\mathbf{J}_h^{k-1}, \boldsymbol{\phi}_h) + \tau(\mathbf{f}^{k-\frac{1}{2}}, \boldsymbol{\phi}_h), \\
(ii) \quad & (1 + \frac{\tau^2\omega_m^2}{2(2+\tau\Gamma_m)})(H_h^k, \psi_h) + \frac{\tau}{2}(\nabla\times\mathbf{E}_h^k, \psi_h) \\
&= (1 - \frac{\tau^2\omega_m^2}{2(2+\tau\Gamma_m)})(H_h^{k-1}, \psi_h) - \frac{\tau}{2}(\nabla\times\mathbf{E}_h^{k-1}, \psi_h) \\
&\quad -\frac{2\tau}{2+\tau\Gamma_m}(K_h^{k-1}, \psi_h) + \tau(g^{k-\frac{1}{2}}, \psi_h).
\end{aligned}$$

We can simply rewrite the above system as:

$$A\mathbf{E}^k - B\mathbf{H}^k = \tilde{\mathbf{f}}, \tag{7.15}$$

$$B'\mathbf{E}^k + C\mathbf{H}^k = \tilde{\mathbf{g}}, \tag{7.16}$$

where A, B and C represent the corresponding coefficient matrices. Here B' denote the transpose of B. Solving for $\mathbf{H}^k$ from (7.16), then substituting it into (7.15), we obtain

$$\mathbf{H}^k = C^{-1}(\tilde{\mathbf{g}} - B'\mathbf{E}^k), \quad \mathbf{E}^k = (A + BC^{-1}B')^{-1}(\tilde{\mathbf{f}} + BC^{-1}\tilde{\mathbf{g}}). \tag{7.17}$$

In summary, the algorithm can be implemented as follows: At each time step, we first solve for $\mathbf{E}_h^k$ from (7.17) , then H_h^k; and finally update $\mathbf{J}_h^k$ and K_h^k using (7.13) and (7.14), respectively.

7.3 Calculation of Element Matrices

On a rectangle $K = [x_a, x_b] \times [y_a, y_b]$, we use a scaled edge element basis functions for the space $\mathbf{V}_h$:

$$\hat{\mathbf{N}}_1 = \begin{pmatrix} \frac{y_b - y}{y_b - y_a} \\ 0 \end{pmatrix}, \quad \hat{\mathbf{N}}_2 = \begin{pmatrix} 0 \\ \frac{x - x_a}{x_b - x_a} \end{pmatrix}, \quad \hat{\mathbf{N}}_3 = \begin{pmatrix} \frac{y_a - y}{y_b - y_a} \\ 0 \end{pmatrix}, \quad \hat{\mathbf{N}}_4 = \begin{pmatrix} 0 \\ \frac{x - x_b}{x_b - x_a} \end{pmatrix},$$

where the edges are oriented counterclockwisely, starting from the bottom edge.

In (7.15), the matrix A is really a multiple of a global mass matrix $matM = (\mathbf{N}_j, \mathbf{N}_i)$, while B is a multiple of matrix $matBM = (1, \nabla \times \mathbf{N}_i)$. The matrix C in (7.16) is just a diagonal matrix, whose elements are the areas of all elements. Matrices $matM$ and $matBM$ can be constructed from the corresponding matrices on each element. These element matrices can be obtained directly by the following lemmas.

Lemma 7.1. *The mass matrix* $M^e = (M_{ij}^e) = (\int_{x_a}^{x_b} \int_{y_a}^{y_b} \mathbf{N}_i \cdot \mathbf{N}_j dxdy)$ *is given by*

$$M^e = \frac{(x_b - x_a)(y_b - y_a)}{6} \begin{bmatrix} 2 & 0 & -1 & 0 \\ 0 & 2 & 0 & -1 \\ -1 & 0 & 2 & 0 \\ 0 & -1 & 0 & 2 \end{bmatrix}.$$

Proof. It is easy to see that M^e is symmetric, and $M_{12}^e = M_{23}^e = M_{34}^e = 0$. Furthermore, we have

$$M_{11}^e = \int_{x_a}^{x_b} \int_{y_a}^{y_b} \left(\frac{y_b - y}{y_b - y_a} \right)^2 dxdy = \frac{1}{3}(x_b - x_a)(y_b - y_a),$$

$$M_{13}^e = \int_{x_a}^{x_b} \int_{y_a}^{y_b} \frac{(y_b - y)(y_a - y)}{(y_b - y_a)^2} dxdy = \frac{-1}{6}(x_b - x_a)(y_b - y_a),$$

and

$$M_{24}^e = \int_{x_a}^{x_b} \int_{y_a}^{y_b} \frac{(x - x_a)(x - x_b)}{(x_b - x_a)^2} dxdy = \frac{-1}{6}(x_b - x_a)(y_b - y_a),$$

which completes the proof. □

Lemma 7.2. *The corresponding element curl matrix* $B^e = (B_j^e) = (\int_{x_a}^{x_b}\int_{y_a}^{y_b} \nabla \times \boldsymbol{N}_j\, dxdy)$ *is given by*

$$B^e = [x_b - x_a \; y_b - y_a \; x_b - x_a \; y_b - y_a].$$

Proof. Note that

$$\begin{aligned} B_1^e &= \int_{x_a}^{x_b}\int_{y_a}^{y_b} \left(\frac{\partial \mathbf{N}_1^{(2)}}{\partial x} - \frac{\partial \mathbf{N}_1^{(1)}}{\partial y}\right) dxdy \\ &= \int_{x_a}^{x_b}\int_{y_a}^{y_b} \frac{1}{y_b - y_a} dxdy = x_b - x_a. \end{aligned}$$

Similarly, we can prove the other components. □

7.4 Assembly Process and Boundary Conditions

The global mass matrix $matM$ and curl matrix $matBM$ can be formed by assembling the contributions from each element matrix. More specifically, we just need to loop through all the edges of all elements in the mesh to find the global label for each edge, and put the contribution into the right location in the global matrix. This is different from the classical nodal-based finite element method, which needs to loop through all the nodes of all elements in the mesh. We like to remark that during the assembly process, the orientation of each edge (stored as ± 1 in array $edori(1 : numel, 1 : 4)$) needs to be considered before each component is added to the global matrix.

The detailed assembly process for both $matB$ and $matBM$ is realized in the following code.

```
for i=1:numel   % loop through elements
  for j=1:4     % loop through edges
    ed1 = el2ed(j,i);
    matBM(ed1,i) = matBM(ed1,i) + edori(i,j)*Curl(j);

    for k=j:4   % loop through edges
      ed2 = el2ed(k,i);
      matM(ed1,ed2) = matM(ed1,ed2) \ldots
                    + edori(i,j)*edori(i,k)*Mref(j,k);
      matM(ed2,ed1) = matM(ed1,ed2);
    end
  end
end
```

Since our boundary condition $\mathbf{n} \times \mathbf{E} = \mathbf{0}$ is a natural boundary condition, we don't have to impose it explicitly.

After assembly, we have to solve the system (7.17) for the unknown coefficients of electric field $\mathbf{E}$. Since the coefficient matrix is symmetric and well conditioned, we just use the simple direct solver provided by MATLAB. Interested readers can use more advanced solvers, such as the Generalized Minimal Residual (GRMES) method, the Bi-Conjugate Gradient (Bi-CG) method, the Bi-Conjugate Gradient Stabilized (Bi-CGSTAB) method [28], multigrid method and the preconditioner method [129, 136, 146].

The complete MATLAB source code *form_mass_matrix.m*, which accomplishes the construction of the global matrix, is shown below.

```
function [rhsEF,rhsEE,rhsEJ,rhsEH,rhsHH,rhsHK,rhsHG,H0,K0] = ...
          form_mass_matrix(HH,KK,gRHS,f1RHS,f2RHS,Ex,Ey,Jx,Jy)

globals2D;

numel=(nelex)*(neley);
one = ones(1,4);
rhsEF=zeros(numed,1);          % for (f,N_i)
rhsEE=zeros(numed,1);          % for (E0,N_i)
rhsEJ=zeros(numed,1);          % for (J0,N_i)
rhsEH=zeros(numed,1);    % for (H0,curl N_i)
rhsHH=zeros(numel,1);        % for (H0, psi_i)
rhsHK=zeros(numel,1);        % for (K0, psi_i)
rhsHG=zeros(numel,1);        % for (g, psi_i)
% store the initial value at each element center
H0 = zeros(numel,1);
K0 = zeros(numel,1);

matM = sparse(numed,numed);  % zero matrix of numedges x numedges
matBM = sparse(numed,numel);
area = zeros(numel,1);
for i=1:numel
   % coordinates of this element from 1st node & 3rd node
   xae=no2xy(1,el2no(1,i));  xbe=no2xy(1,el2no(3,i));
   yae=no2xy(2,el2no(1,i));  ybe=no2xy(2,el2no(3,i));

   midpt(i,1) = 0.5*(min(no2xy(1,el2no(:,i))) ...
                + max(no2xy(1,el2no(:,i))));
   midpt(i,2) = 0.5*(min(no2xy(2,el2no(:,i))) ...
                + max(no2xy(2,el2no(:,i))));
   H0(i) = HH(midpt(i,1),midpt(i,2),0);  % element center value
   K0(i) = KK(midpt(i,1),midpt(i,2),0);  % element center value
   rhs_g = gRHS(midpt(i,1),midpt(i,2),0.5*dt);

   % the coordinates of the four vertex
   xe(1)=xae; ye(1)=yae;
   xe(2)=xbe; ye(2)=yae;
   xe(3)=xbe; ye(3)=ybe;
   xe(4)=xae; ye(4)=ybe;
   area(i) = (ybe-yae)*(xbe-xae);   % for non-uniform rectangles
```

```
for j=1:4   % loop through edges
   ed1 = el2ed(j,i);

   % evaluate the RHS: \int_O fRHS * N_j
   % we used Gaussian integration: cf. my book p.190!
   rhs_ef=0;  rhs_ee=0;  rhs_ej=0;

   for ii=1:2  % loop over gauss points in eta
      for jj=1:2   % loop over gauss points in psi
        eta = gauss(ii);  psi = gauss(jj);
        % Q1 function: countcockwise starting at left-low corner
        NJ=0.25*(one + psi*psiJ).*(one + eta*etaJ);
        % derivatives of shape functions in reference coordinates
        NJpsi = 0.25*psiJ.*(one + eta*etaJ);    % 1x4 array
        NJeta = 0.25*etaJ.*(one + psi*psiJ);    % 1x4 array
        % derivatives of x and y wrt psi and eta
        xpsi = NJpsi*xe';  ypsi = NJpsi*ye';
        xeta = NJeta*xe';  yeta = NJeta*ye';
        % Jinv = [yeta, -xeta; -ypsi, xpsi];      % 2x2 array
        jcob = xpsi*yeta - xeta*ypsi;

        xhat=dot(xe,NJ);  yhat=dot(ye,NJ);

        if j==1
          bas1=(ybe-yhat)/(ybe-yae);
          rhs_ef = rhs_ef + f1RHS(xhat,yhat,0.5*dt)*bas1*jcob;
          rhs_ee = rhs_ee + Ex(xhat,yhat,0)*bas1*jcob;
          rhs_ej = rhs_ej + Jx(xhat,yhat,0)*bas1*jcob;
        elseif j==2
          bas2=(xhat-xae)/(xbe-xae);
          rhs_ef = rhs_ef + f2RHS(xhat,yhat,0.5*dt)*bas2*jcob;
          rhs_ee = rhs_ee + Ey(xhat,yhat,0)*bas2*jcob;
          rhs_ej = rhs_ej + Jy(xhat,yhat,0)*bas2*jcob;
        elseif j==3
          bas3=-(yhat-yae)/(ybe-yae);
          rhs_ef = rhs_ef + f1RHS(xhat,yhat,0.5*dt)*bas3*jcob;
          rhs_ee = rhs_ee + Ex(xhat,yhat,0)*bas3*jcob;
          rhs_ej = rhs_ej + Jx(xhat,yhat,0)*bas3*jcob;
        else
          bas4=-(xbe-xhat)/(xbe-xae);
          rhs_ef = rhs_ef + f2RHS(xhat,yhat,0.5*dt)*bas4*jcob;
          rhs_ee = rhs_ee + Ey(xhat,yhat,0)*bas4*jcob;
          rhs_ej = rhs_ej + Jy(xhat,yhat,0)*bas4*jcob;
        end
    end
   end

    % assemble the edge contribution into global rhs vector
    rhsEF(ed1)=rhsEF(ed1)+edori(i,j)*rhs_ef;
    rhsEE(ed1)=rhsEE(ed1)+edori(i,j)*rhs_ee;
    rhsEJ(ed1)=rhsEJ(ed1)+edori(i,j)*rhs_ej;
    rhsEH(ed1)=        edori(i,j)*H0(i)*Curl(j);

    matBM(ed1,i) = matBM(ed1,i) + edori(i,j)*Curl(j);
```

```
        for k=j:4
          ed2 = el2ed(k,i);
          matM(ed1,ed2) = matM(ed1,ed2) ...
                          + edori(i,j)*edori(i,k)*Mref(j,k);
          matM(ed2,ed1) = matM(ed1,ed2);
        end
    end   % end of 1st edge loop
    rhsHH(i)=H0(i)*area(i);
    rhsHK(i)=K0(i)*area(i);
    rhsHG(i)=rhs_g*area(i);
end
return
```

By similar techniques, we have to assemble the right hand side vector in each time step. This task is realized in the driver function *Drude_cn.m* shown in Sect. 7.6.

7.5 Postprocessing

Once we obtain the unknown coefficients of electric field $\mathbf{E}$, we can use them to construct the numerical electric field $\mathbf{E}_h$ at any point, which can be used to compare with the analytic electric field $\mathbf{E}$ for error estimates. This reconstruction part can be realized in the following code, where the numerical electric field $\mathbf{E}$ is calculated at each element center.

```
solvec = zeros(numed,1);
% extract the coefficients of E field
solvec(eint)=znew(1:iecnt);
for i=1:numel
    % coordinates of this element from 1st & 3rd nodes
    xae=no2xy(1,el2no(1,i));  xbe=no2xy(1,el2no(3,i));
    yae=no2xy(2,el2no(1,i));  ybe=no2xy(2,el2no(3,i));
    % basis functions
    bas1=(ybe-midpt(i,2))/(ybe-yae);
    bas3=-(midpt(i,2)-yae)/(ybe-yae);
    bas2=(midpt(i,1)-xae)/(xbe-xae);
    bas4=-(xbe-midpt(i,1))/(xbe-xae);

    %construct the numerical E fields
    Ex_num(i)=edori(i,1)*solvec(el2ed(1,i))*bas1 + \ldots
              edori(i,3)*solvec(el2ed(3,i))*bas3;
    Ey_num(i)=edori(i,2)*solvec(el2ed(2,i))*bas2 + \ldots
              edori(i,4)*solvec(el2ed(4,i))*bas4;
 end
```

Considering that H is a piecewise constant, the numerical magnetic field H can be directly obtained by the following code.

```
for i=1:numel
    HH_num(i) = znew(iecnt+i);
end
```

Once we have the numerical solutions, we can postprocess the solutions in various ways. For example, we can plot the electric field $\mathbf{E}$ by simple commands as follows:

```
figure(1);clf(1);
quiver(midpt(:,1)',midpt(:,2)',Ex_num,Ey_num),
titstr=strcat('Numerical E field at midpoints');
title(titstr),
axis([lowx highx lowy highy]);
```

Similarly, we can do a surface plot for the scale magnetic field H as shown below:

```
figure(4);clf(4);
for j=1:neley
    for i=1:nelex
        % change 1-D vector into 2-D array
        U2d(i,j)=HH_num(nelex*(j-1)+i);
    end
end

surf(1:nelex, 1:neley, U2d');
titstr=strcat('Numerical H field');
title(titstr);
xlabel('X'); ylabel('Y');
```

A sample MATLAB code *postprocessing.m* demonstrating our postprocessing implementation is given below:

```
function postprocessing(HH,Ex,Ey,znew,tt,numel)

globals2D;

%plot the numerical field
solvec = zeros(numed,1);
solvec(eint)=znew(1:iecnt);   %coefficients of E field
 for i=1:numel
    % coordinates of this element from 1st node & 3rd node
    xae=no2xy(1,el2no(1,i));  xbe=no2xy(1,el2no(3,i));
    yae=no2xy(2,el2no(1,i));  ybe=no2xy(2,el2no(3,i));
    % basis functions: cf p111
    bas1=(ybe-midpt(i,2))/(ybe-yae);
    bas3=-(midpt(i,2)-yae)/(ybe-yae);
    bas2=(midpt(i,1)-xae)/(xbe-xae);
```

```
    bas4=-(xbe-midpt(i,1))/(xbe-xae);

    %calculate the numerical and exact E fields
    Ex_num(i)=edori(i,1)*solvec(el2ed(1,i))*bas1 + ...
              edori(i,3)*solvec(el2ed(3,i))*bas3;
    Ey_num(i)=edori(i,2)*solvec(el2ed(2,i))*bas2 + ...
              edori(i,4)*solvec(el2ed(4,i))*bas4;

    Ex_ex(i) = Ex(midpt(i,1),midpt(i,2),tt);
    Ey_ex(i) = Ey(midpt(i,1),midpt(i,2),tt);

    % calculate the numerical and exact H fields (a scalar)
    HH_ex(i) = HH(midpt(i,1),midpt(i,2),tt);
    HH_num(i) = znew(iecnt+i);
 end

 figure(1);clf(1);
 quiver(midpt(:,1)',midpt(:,2)',Ex_num,Ey_num),
 titstr=strcat('Numerical E field at midpoints');
 title(titstr),
 axis([lowx highx lowy highy]);

figure(2);clf(2);
quiver(midpt(:,1)',midpt(:,2)',Ex_ex, Ey_ex),
titstr=strcat('Analytical E field at midpoints');
title(titstr),
axis([lowx highx lowy highy]);

% plot Hz at the last time step
      timestep=int2str(nt);
      figure(3);clf(3);
      pcolor((reshape(HH_num(1:numel),nelex,neley))');
      hold on;
   %%   line(boxlinex,boxliney,'Color','w');
   %   hold off
      shading flat;
    %  caxis([-1.0 1.0]);
    %%  axis([1 ie 1 je]);
      colorbar;
      axis image;
     % axis off;
      titstr=strcat('Numerical H field');
      title(titstr);
      xlabel('X'); ylabel('Y');

      figure(4);clf(4);
      for j=1:neley
        for i=1:nelex
          % change 1-D vector into 2-D array
          U2d(i,j)=HH_num(nelex*(j-1)+i);
          H2d(i,j)=HH_ex(nelex*(j-1)+i)-U2d(i,j);
          Ex2d(i,j)=Ex_ex(nelex*(j-1)+i)-Ex_num(nelex*(j-1)+i);
        end
      end
```

```
      surf(1:nelex, 1:neley, U2d');
      title(titstr);
      xlabel('X'); ylabel('Y');

      figure(5);clf(5);
      surf(1:nelex, 1:neley, H2d');
      title('H field pointwise error');
      xlabel('X'); ylabel('Y');

      figure(6);clf(6);
      surf(1:nelex, 1:neley, Ex2d');
      title('Electric component Ex pointwise error');
      xlabel('X'); ylabel('Y');

%Debug: check the last 4 element solutions
for i=numel-4:numel
    display(' H exact, numer ='),HH_ex(i),HH_num(i)
    display('Ex exact, numer ='),Ex_ex(i),Ex_num(i)
end

display('Number of interior edges, numel, DOF = '), ...
                size(eint),numel,size(znew)

% calculate the max pointwise error
err_Ex=max(abs(Ex_num-Ex_ex)),
err_Ey=max(abs(Ey_num-Ey_ex)),
err_H=max(abs(HH_num-HH_ex)),
```

7.6 Numerical Results

In this section, we use an example to demonstrate our implementation of the scheme (7.1)–(7.4). To check the convergence rate, we construct the following exact solutions for the 2-D transverse electrical model (assuming that $\Gamma_m = \Gamma_e, \omega_m = \omega_e$) on the domain $\Omega = (0, 1)^2$:

$$\mathbf{E} \equiv \begin{pmatrix} E_x \\ E_y \end{pmatrix} = \begin{pmatrix} \sin \pi y \\ \sin \pi x \end{pmatrix} e^{-\Gamma_e t},$$

$$H = \frac{1}{\pi}(\cos \pi x - \cos \pi y) e^{-\Gamma_e t} (\omega_e^2 t - \Gamma_e).$$

The corresponding electric and magnetic currents are

$$\mathbf{J} \equiv \begin{pmatrix} J_x \\ J_y \end{pmatrix} = \begin{pmatrix} \sin \pi y \\ \sin \pi x \end{pmatrix} \omega_e^2 t e^{-\Gamma_e t},$$

and

$$K = \frac{1}{\pi}(\cos \pi x - \cos \pi y)e^{-\Gamma_e t}\omega_e^2(\frac{1}{2}\omega_e^2 t^2 - \Gamma_e t),$$

respectively. The corresponding source term $\mathbf{f} \equiv 0$, while g is given by

$$g = \frac{1}{\pi}(\cos \pi x - \cos \pi y)e^{-\Gamma_e t}(-2\Gamma_e\omega_e^2 t + \Gamma_e^2 + \omega_e^2 + \pi^2 + \frac{1}{2}\omega_e^4 t^2).$$

Notice that our solution $\mathbf{E}$ satisfies the conditions

$$\mathbf{n} \times \mathbf{E} = \mathbf{0} \quad \text{on } \partial\Omega, \qquad \nabla \cdot \mathbf{E} = 0 \quad \text{in } \Omega.$$

Our complete codes for solving this problem are composed of five MATLAB functions:

1. *Drude_cn.m*: the driver function;
2. *globals2D.m*: define all the global variables and constants;
3. *create_mesh.m, form_mass_matrix.m, postprocessing.m*: the other supporting functions explained above.

In the driver function *Drude_cn.m*, we assign the time step size, the total number of time steps of the simulation, and define the exact solutions used to calculate the error estimates. Below is our implementation of *Drude_cn.m*:

```
%-------------------------------------------------------------
% Author: Prof. Jichun Li
%
% Solve Drude metamaterial model using Crank-Nicolson type
% mixed FEM by rectangular edge element.
%
% Drive function (this one): Drude_cn.m
% Other supporting functions:
%    1. globals2D.m: define global variables and constants
%    2. create_mesh.m:  generate rectangular mesh
%    3. form_mass_matrix.m: create the global mass matrix
%                      and prepare for time marching
%    4. postprocessing.m: compare numerical and analytical
%            solutions, calculate errors and do plottings
%-------------------------------------------------------------
clear all,
globals2D;   % all global variables and constants
format long;

%%%%%%%%%% set up the exact solutions  %%%%%%%%%%%%%
gama=1.0e0; wpem=1.0e0;
% exact electric field and electric polarization
Ex = @(x,y,t)sin(pi*y).*exp(-gama*t);
Ey = @(x,y,t)sin(pi*x).*exp(-gama*t);
Jx = @(x,y,t)sin(pi*y).*exp(-gama*t)*wpem^2*t;
Jy = @(x,y,t)sin(pi*x).*exp(-gama*t)*wpem^2*t;
% exact magnetic field and magnetic polarization
```

```
HH = @(x,y,t)(cos(pi*x)-cos(pi*y))...
       /pi.*exp(-gama*t)*(wpem^2*t-gama);
KK = @(x,y,t)(cos(pi*x)-cos(pi*y))/pi.*exp(-gama*t)...
       *wpem^2*(0.5*(wpem*t)^2-gama*t);

% exact RHS
f1RHS = @(x,y,t)0.0;
f2RHS = @(x,y,t)0.0;
gRHS  = @(x,y,t)(cos(pi*x)-cos(pi*y))/pi.*exp(-gama*t)...
        *(-2*gama*wpem^2*t + gama^2 + wpem^2 ...
        + pi^2 + 0.5*wpem^2*(wpem*t)^2);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
dt=1.e-8, nt=10;
id_bc = 1;    %Indicator: 1 for Dirichlet BC; 0 otherwise.

% create a rectangular mesh on [lowx,highx]x[lowy,highy]
create_mesh;
dim = iecnt + numel;   % total number of unknowns

% local mass matrix
Mref = (dx*dy/6)*[2 0 -1 0;0 2 0 -1;-1 0 2 0;0 -1 0 2];
Curl = [dx;dy;dx;dy];      % (1, curl N_i)

matM = sparse(numed,numed);
matBM = sparse(numed,numel);
area = zeros(numel,1);

% form matrix matM and matrix matBM
[rhsEF,rhsEE,rhsEJ,rhsEH,rhsHH,rhsHK,rhsHG,H0,K0] = ...
    form_mass_matrix(HH,KK,gRHS,f1RHS,f2RHS,Ex,Ey,Jx,Jy);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

if id_bc == 1 % For Dirichlet BC, use internal edge values
    matM = matM(eint,eint);
    matBM=matBM(eint,:);
    rhsEF=rhsEF(eint);
    rhsEE=rhsEE(eint);
    rhsEJ=rhsEJ(eint);
    rhsEH=rhsEH(eint);
end

E0=inv(matM)*rhsEE;   % get initial values of E and J
J0=inv(matM)*rhsEJ;

cst1 = 1+(dt*wpem)^2/(2*(2+dt*gama));
cst2 = 0.5*dt; cst3 = 1-(dt*wpem)^2/(2*(2+dt*gama));
cst4 = 2*dt/(2+dt*gama);
cst5=(2-dt*gama)/(2+dt*gama);
cst6=dt*wpem^2/(2+dt*gama);

 rhs_glb=[cst3*rhsEE + cst2*rhsEH - cst4*rhsEJ + dt*rhsEF; ...
        cst3*rhsHH - cst2*matBM'*E0 - cst4*rhsHK + dt*rhsHG];

 % the global coefficient matrix (A -B; B' D)
```

```
 % Ae  - Bh = f_1
 % B'e + Dh = f_2
 % solve the system: h = D^{-1}(f_2-B'e);
 %                  e = (A+BD^{-1}B')^{-1}(f_1+BD^{-1}f_2)
 % coefficient matrix for unknown E is: A+B*inv(D)*B'

 %mat4E = cst1*matM + matBM*matBM'./area*(cst2*cst2)/cst3;
 for k=1:numel
    tmp(:,k) = (matBM(:,k)/area(k));
 end
 mat4E = cst1*matM + matBM*tmp'*(cst2*cst2)/cst3;

% rhs4E = rhs_glb(1:iecnt) ...
%         + matBM./area*rhs_glb(iecnt+1:dim)*cst2/cst3;
 rhs4E = rhs_glb(1:iecnt) ...
          +tmp(:,1:numel)*rhs_glb(iecnt+1:dim)*cst2/cst3;

 %%%%%%%%%%%%%% begin time marching %%%%%%%%%%%%%%%%%%
 one = ones(1,4);
 tic   % start to measure the elapsed time
 for n=1:nt
   %first solve the system for E_h^k, then for H_h^k
   znew(1:iecnt,1) = mat4E\rhs4E;

   sum1 = zeros(numel,1);
   for ii=1:numel
       sum1(ii)=sum1(ii)+matBM(1:iecnt,ii)'*znew(1:iecnt);
   end

   znew(iecnt+1:dim,1) = (rhs_glb(iecnt+1:dim,1) ...
        -cst2*sum1(1:numel,1))./(cst3*area(1:numel,1));
   %znew(iecnt+1:dim)=(rhs_glb(iecnt+1:dim)...
   %                     -dt*matBM'*znew(1:iecnt))/a1;

   % for safety, re-assign to zero
   rhsEF=zeros(numed,1);          % for (f,N_i)
   rhsEH=zeros(numed,1);        % for (H0,curl N_i)
   rhsHH=zeros(numel,1);        % for (H0, psi_i)
   rhsHK=zeros(numel,1);        % for (K0, psi_i)
   rhsHG=zeros(numel,1);        % for (g, psi_i)

   if n <= (nt-1)
     % update all degrees of freedom
     tt = (n+0.5)*dt;
     En = znew(1:iecnt);
     Hn = znew(iecnt+1:dim);
     Jn = cst6*(En + E0) + cst5*J0;
     Kn = cst6*(Hn + H0) + cst5*K0;

     for i=1:numel
       xae=no2xy(1,el2no(1,i));  xbe=no2xy(1,el2no(3,i));
       yae=no2xy(2,el2no(1,i));  ybe=no2xy(2,el2no(3,i));

       % the coordinates of the four vertex
```

```
      xe(1)=xae; ye(1)=yae;
      xe(2)=xbe; ye(2)=yae;
      xe(3)=xbe; ye(3)=ybe;
      xe(4)=xae; ye(4)=ybe;

      for j=1:4   % loop through edges
       ed1 = el2ed(j,i);
       rhs_ef=0;

       for ii=1:2  % loop over gauss points in eta
          for jj=1:2   % loop over gauss points in psi
            eta = gauss(ii);  psi = gauss(jj);
            % Q1 function:
            NJ=0.25*(one + psi*psiJ).*(one + eta*etaJ);
            % derivatives of shape functions
            NJpsi=0.25*psiJ.*(one + eta*etaJ);  % 1x4 array
            NJeta=0.25*etaJ.*(one + psi*psiJ);  % 1x4 array
            % derivatives of x and y wrt psi and eta
            xpsi = NJpsi*xe';  ypsi = NJpsi*ye';
            xeta = NJeta*xe';  yeta = NJeta*ye';
          % Jinv = [yeta, -xeta; -ypsi, xpsi];  % 2x2 array
            jcob = xpsi*yeta - xeta*ypsi;

            xhat=dot(xe,NJ); yhat=dot(ye,NJ);

           if j==1
             bas1=(ybe-yhat)/(ybe-yae);
             rhs_ef=rhs_ef + f1RHS(xhat,yhat,tt)*bas1*jcob;
           elseif j==2
             bas2=(xhat-xae)/(xbe-xae);
             rhs_ef=rhs_ef + f2RHS(xhat,yhat,tt)*bas2*jcob;
           elseif j==3
             bas3=-(yhat-yae)/(ybe-yae);
             rhs_ef=rhs_ef + f1RHS(xhat,yhat,tt)*bas3*jcob;
           else
             bas4=-(xbe-xhat)/(xbe-xae);
             rhs_ef=rhs_ef + f2RHS(xhat,yhat,tt)*bas4*jcob;
           end
         end
       end
      % assemble the edge contribution into global rhs vector
      rhsEF(ed1)=rhsEF(ed1)+edori(i,j)*rhs_ef;
      rhsEH(ed1)=         edori(i,j)*Hn(i)*Curl(j);
    end   % end of 1st edge loop
    rhsHH(i)=Hn(i)*area(i);
    rhsHK(i)=Kn(i)*area(i);
    rhsHG(i)=gRHS(midpt(i,1),midpt(i,2),tt)*area(i);
  end  % end of element loop

     if id_bc == 1  % Dirichlet BC
       rhsEF=rhsEF(eint);
       rhsEH=rhsEH(eint);
     end
   % form new RHS
```

Table 7.1 The pointwise errors at element centers with $\Gamma_e = \omega_e = 1, \tau = 10^{-8}$ after 1 time step

Meshes	E_x errors	H_z errors
10×10	4.10388426568e−003	2.55501841905e−010
20×20	1.02758690447e−003	6.44169162455e−011
40×40	2.57051528841e−004	1.61380908636e−011
80×80	6.43804719980e−005	4.19719814459e−012
160×160	1.63183158637e−005	1.66422431391e−012

Table 7.2 The pointwise errors at element centers with $\Gamma_e = \omega_e = 1, \tau = 10^{-8}$ after 100 time step

Meshes	E_x errors	H_z errors	DOFs	CPU time (sec)
10×10	4.10387149e−03	2.55493626e−10	280	5.49
20×20	1.02756605e−03	6.44204689e−11	1,160	22.17
40×40	2.57014726e−04	1.61460844e−11	4,720	99.71
80×80	6.43118232e−05	4.11581879e−12	19,040	604.19
160×160	1.61859982e−05	1.33271171e−12	76,480	4479.43

```
rhs_glb=[cst3*matM*E0+cst2*rhsEH-cst4*matM*J0+dt*rhsEF;...
         cst3*rhsHH-cst2*matBM'*E0-cst4*rhsHK+dt*rhsHG];
rhs4E = rhs_glb(1:iecnt) ...
        + tmp(:,1:numel)*rhs_glb(iecnt+1:dim)*cst2/cst3;

   % update all dof
   E0 = En; J0 = Jn; H0 = Hn; K0 = Kn;
  end    % end of the BIG 'if' loop
  display('step n='),n
end    % end of time marching
toc    % end measurement of elapsed time
 %%%%%%%%%%%%%% end of time marching %%%%%%%%%%%%%%%
% compare the analytic and numerical solutions at T
postprocessing(HH,Ex,Ey,znew,nt*dt,numel);
return
```

Exemplary results are shown in Tables 7.1 and 7.2 (where DOFs denote the total number of degrees of freedom) and in Fig. 7.1. Tables 7.1 and 7.2 clearly show the pointwise convergence rate $O(h^2)$ at element centers, where h is the mesh size. Note that $O(h^2)$ is better than the theoretical approximation result, i.e., superconvergence happens at the rectangular element centers as we proved in Chap. 5.

7.7 Bibliographical Remarks

The number of books covering finite element programming for Maxwell's equations is quite limited. For example, the classic books by Jin [162] and by Silverster and Ferrari [267] describe the basic finite element techniques for Maxwell's equations. [267] even provides all the source code in Fortran. The recent book by Hesthaven and Warburton [141] introduced the nodal discontinuous Galerkin (DG) method for

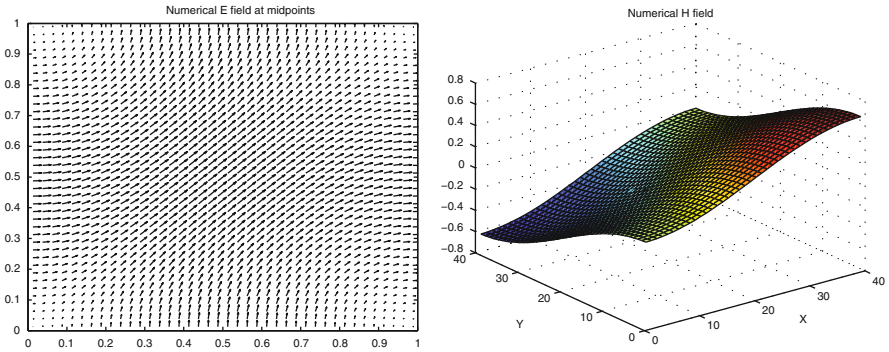

Fig. 7.1 Numerical solution obtained on 20×20 mesh with $\Gamma_e = \omega_e = 1, \tau = 10^{-10}$ after 100 time steps: (*Left*) The electric field; (*Right*) The magnetic field

conservation laws and Maxwell's equations. This book provides a very nice package for readers to experience the DG method for solving various problems, including time-domain Maxwell's equations in free space. Other recent contributions to this area are the books by Demkowicz et al. [97, 98], in which they detailed the hp-finite element method for solving elliptic problems and time-harmonic Maxwell's equations. A self-contained 2-D package (covering grid generation, solver, and visualisation) is included in [97]. Readers can find a few other hp Maxwell packages mentioned in the Foreword of [97].

Chapter 8
Perfectly Matched Layers

One common problem in computational electromagnetics is how to simulate wave propagation on an unbounded domain accurately and efficiently. One typical technique is to use the absorbing boundary conditions (ABCs) to truncate the unbounded domain to a bounded domain. The solution computed with an ABC on a bounded domain should be a good approximation to the solution originally given on the unbounded domain. Hence constructing a good ABC is quite challenging.

Over the years various ABCs have been developed in computational electromagnetics, and the most effective ABC seems to be the perfectly matched layer (PML) introduced by Berenger in 1994 [34]. Since 1994, the PML has been intensively investigated, and many interesting results have been obtained (cf. [10], [276, Chap. 7] and references therein).

In this chapter, we present some interesting PMLs developed over the last 20 years. More specifically, in Sect. 8.1, we discuss three ways for obtaining the PMLs matched to the free space. Then in Sect. 8.2, we extend the discussion to lossy media. Finally, we describe some PMLs developed for the dispersive media and metamaterials in Sect. 8.3.

8.1 PMLs Matched to the Free Space

8.1.1 Berenger Split PMLs

In 1994, Berenger [34] proposed the first time-domain perfectly matched layer for modeling electromagnetic wave propagation in unbounded free space. The basic idea is to introduce a specially designed layer to absorb the electromagnetic waves without any reflection from the interfaces between the free space and the special layer. Later in 1996, Berenger [35] extended the PML medium technique to 3-D. Below we shall introduce the 3-D Berenger PML, since the 2-D PML can be directly obtained from the 3-D model as special cases. Following [35], Berenger's PML is

J. Li and Y. Huang, *Time-Domain Finite Element Methods for Maxwell's Equations in Metamaterials*, Springer Series in Computational Mathematics 43,
DOI 10.1007/978-3-642-33789-5_8,

based on a splitted form of Maxwell's equations: the six field components are split into 12 subcomponents (i.e., $\mathbf{E}_x = E_{xy} + E_{xz}, \mathbf{E}_y = E_{yx} + E_{yz}, \mathbf{E}_z = E_{zx} + E_{zy}$, $\mathbf{H}_x = H_{xy} + H_{xz}, \mathbf{H}_y = H_{yx} + H_{yz}, \mathbf{H}_z = H_{zx} + H_{zy}$), in which case the original six Cartesian equations are split into 12 subequations as follows:

$$\epsilon_0 \frac{\partial E_{xy}}{\partial t} + \sigma_y E_{xy} = \frac{\partial (H_{zx} + H_{zy})}{\partial y} \tag{8.1a}$$

$$\epsilon_0 \frac{\partial E_{xz}}{\partial t} + \sigma_z E_{xz} = -\frac{\partial (H_{yz} + H_{yx})}{\partial z} \tag{8.1b}$$

$$\epsilon_0 \frac{\partial E_{yz}}{\partial t} + \sigma_z E_{yz} = \frac{\partial (H_{xy} + H_{xz})}{\partial z} \tag{8.1c}$$

$$\epsilon_0 \frac{\partial E_{yx}}{\partial t} + \sigma_x E_{yx} = -\frac{\partial (H_{zx} + H_{zy})}{\partial x} \tag{8.1d}$$

$$\epsilon_0 \frac{\partial E_{zx}}{\partial t} + \sigma_x E_{zx} = \frac{\partial (H_{yz} + H_{yx})}{\partial x} \tag{8.1e}$$

$$\epsilon_0 \frac{\partial E_{zy}}{\partial t} + \sigma_y E_{zy} = -\frac{\partial (H_{xy} + H_{xz})}{\partial y} \tag{8.1f}$$

$$\mu_0 \frac{\partial H_{xy}}{\partial t} + \sigma_y^* H_{xy} = -\frac{\partial (E_{zx} + E_{zy})}{\partial y} \tag{8.1g}$$

$$\mu_0 \frac{\partial H_{xz}}{\partial t} + \sigma_z^* H_{xz} = \frac{\partial (E_{yz} + E_{yx})}{\partial z} \tag{8.1h}$$

$$\mu_0 \frac{\partial H_{yz}}{\partial t} + \sigma_z^* H_{yz} = -\frac{\partial (E_{xy} + E_{xz})}{\partial z} \tag{8.1i}$$

$$\mu_0 \frac{\partial H_{yx}}{\partial t} + \sigma_x^* H_{yx} = \frac{\partial (E_{zx} + E_{zy})}{\partial x} \tag{8.1j}$$

$$\mu_0 \frac{\partial H_{zx}}{\partial t} + \sigma_x^* H_{zx} = -\frac{\partial (E_{yx} + E_{yz})}{\partial x} \tag{8.1k}$$

$$\mu_0 \frac{\partial H_{zy}}{\partial t} + \sigma_y^* H_{zy} = \frac{\partial (E_{xy} + E_{xz})}{\partial y} \tag{8.1l}$$

where parameters $\sigma_i, \sigma_i^*, i = x, y, z$, are the homogeneous electric and magnetic conductivities in the i direction.

Replacing each component of (8.1) by a plane wave solution, for example,

$$E_{xy} = \tilde{E}_{xy} e^{j(\omega t - \mathbf{k}\cdot\mathbf{r})}, \tag{8.2}$$

we can obtain 12 equations expressed in terms of the angular frequency ω, the wave number $\mathbf{k} = (k_x, k_y, k_z)$, the position vector $\mathbf{r} = (x, y, z)$, and 12 components $\tilde{E}_{xy}, \tilde{E}_{xz}, \cdots, \tilde{H}_{zx}, \tilde{H}_{zy}$. By introducing stretching parameters s_i and s_i^* [35]:

$$s_i = 1 + \frac{\sigma_i}{j\omega\epsilon_0}, \quad s_i^* = 1 + \frac{\sigma_i^*}{j\omega\mu_0}, \quad i = x, y, z, \tag{8.3}$$

we can further reduce the resulting 12 equations into just 6 equations:

$$\omega\epsilon_0\tilde{E}_x = -\frac{k_y}{s_y}\tilde{H}_z + \frac{k_z}{s_z}\tilde{H}_y \tag{8.4a}$$

$$\omega\epsilon_0\tilde{E}_y = -\frac{k_z}{s_z}\tilde{H}_x + \frac{k_x}{s_x}\tilde{H}_z \tag{8.4b}$$

$$\omega\epsilon_0\tilde{E}_z = -\frac{k_x}{s_x}\tilde{H}_y + \frac{k_y}{s_y}\tilde{H}_x \tag{8.4c}$$

$$\omega\mu_0\tilde{H}_x = \frac{k_y}{s_y^*}\tilde{E}_z - \frac{k_z}{s_z^*}\tilde{E}_y \tag{8.4d}$$

$$\omega\mu_0\tilde{H}_y = \frac{k_z}{s_z^*}\tilde{E}_x - \frac{k_x}{s_x^*}\tilde{E}_z \tag{8.4e}$$

$$\omega\mu_0\tilde{H}_z = \frac{k_x}{s_x^*}\tilde{E}_y - \frac{k_y}{s_y^*}\tilde{E}_x, \tag{8.4f}$$

where we denote

$$\tilde{E}_x = \tilde{E}_{xy} + \tilde{E}_{xz},\ \tilde{E}_y = \tilde{E}_{yx} + \tilde{E}_{yz},\ \tilde{E}_z = \tilde{E}_{zx} + \tilde{E}_{zy}, \tag{8.5}$$

$$\tilde{H}_x = \tilde{H}_{xy} + \tilde{H}_{xz},\ \tilde{H}_y = \tilde{H}_{yx} + \tilde{H}_{yz},\ \tilde{H}_z = \tilde{H}_{zx} + \tilde{H}_{zy}. \tag{8.6}$$

Note that (8.4) can be rewritten in vector form as

$$\epsilon_0\omega\tilde{\mathbf{E}} = -\mathbf{k}_s \times \tilde{\mathbf{H}},\ \ \mu_0\omega\tilde{\mathbf{H}} = \mathbf{k}_s^* \times \tilde{\mathbf{E}}, \tag{8.7}$$

where we denote

$$\tilde{\mathbf{E}} = (\tilde{E}_x, \tilde{E}_y, \tilde{E}_z)', \quad \tilde{\mathbf{H}} = (\tilde{H}_x, \tilde{H}_y, \tilde{H}_z)'.$$

and

$$\mathbf{k}_s = (k_x/s_x, k_y/s_y, k_z/s_z)',\ \ \mathbf{k}_s^* = (k_x/s_x^*, k_y/s_y^*, k_z/s_z^*)'.$$

It is interesting to note that if applying (8.2) to the Maxwell's equations in vaccum

$$\epsilon_0\frac{\partial\mathbf{E}}{\partial t} = \nabla \times \mathbf{H}, \quad -\mu_0\frac{\partial\mathbf{H}}{\partial t} = \nabla \times \mathbf{E}, \tag{8.8}$$

we have

$$\epsilon_0 \omega \tilde{\mathbf{E}} = -\mathbf{k} \times \tilde{\mathbf{H}}, \quad \mu_0 \omega \tilde{\mathbf{H}} = \mathbf{k} \times \tilde{\mathbf{E}}, \tag{8.9}$$

which is a special case of the PML equations (8.7). This fact is not surprising, since (8.8) is a special case of (8.1) with $\sigma_i = \sigma_i^* = 0,\ i = x, y, z$.

To make the PML work properly, the electric and magnetic conductivities have to be choosen carefully. In the PML region matched to a vacuum, the transverse conductivities equal zero and the longitudinal conductivities satisfy the impedance matching condition

$$\frac{\sigma_i}{\epsilon_0} = \frac{\sigma_i^*}{\mu_0}, \quad i = x, y, z. \tag{8.10}$$

For example, consider an inner vacuum domain surrounded by an absorbing PML medium. On the six walls of the computational domain, the transverse conductivities of the PML media are set to zero in order to cancel the reflection from the vacuum-PML interfaces. For example, $(0, 0, 0, 0, \sigma_z, \sigma_z^*)$ should be used in the upper and lower walls (i.e., the interfaces normal to the z-direction). Along the 12 interface edges, the longitudinal conductivities are equal to zero, and the transverse conductivities are equal to those of the adjacent side media. Hence, no reflection is produced theoretically from side-edge interfaces. In the eight corner regions, the conductivities are chosen to match those of the adjacent edges, i.e., the transverse conductivities match at interfaces between edge layers and corner layers, which makes zero reflection from all the edge-corner interfaces. Detailed specifications of conductivities in the edge and corner regions of PML are depicted in Fig. 8.1 (cf. Fig. 3 of [35]). Many experiments ([276]) show that the PML conductivities (either σ_i or σ_i^*) can be simply chosen as a polynomial:

$$\sigma_\rho(\rho) = \sigma_*(\frac{\rho}{\delta})^n, \tag{8.11}$$

where $n \geq 2$ is the polynomial degree, δ is the PML thickness, ρ is the distance from the interface, σ_* is the conductivity on the outer side of the PML (at $\rho = \delta$). Note that given a reflection goal $R(0)$, σ_* is often chosen as

$$\sigma_* = -\frac{(n+1)\epsilon_0 C_v}{2\delta} \ln R(0). \tag{8.12}$$

Recall that $C_v = 1/\sqrt{\epsilon_0 \mu_0}$ denotes the wave propagation speed in vacuum.

Finally, we want to mention that in 2-D cases, the above 3-D PML equations (8.1) are reduced to a set of four equations by splitting only one component. More specifically, in the TE_z (transverse electric to z) case, only the magnetic component is split, which results the PML equations for the TE_z case as:

$$\epsilon_0 \frac{\partial E_x}{\partial t} + \sigma_y E_x = \frac{\partial (H_{zx} + H_{zy})}{\partial y} \tag{8.13a}$$

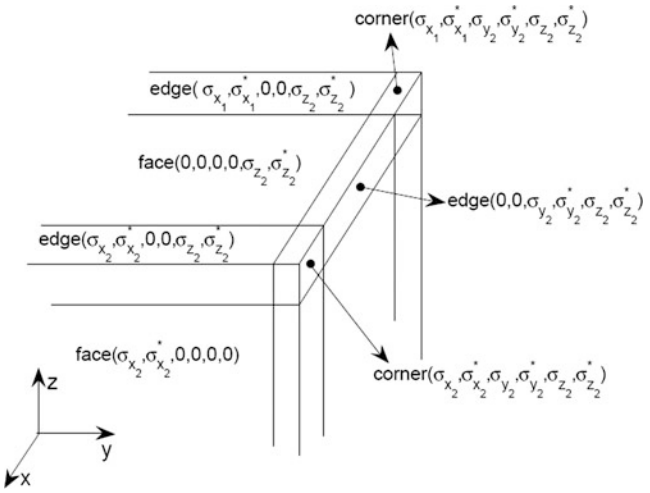

Fig. 8.1 Illustration of the 3-D PML

$$\epsilon_0 \frac{\partial E_y}{\partial t} + \sigma_x E_y = -\frac{\partial (H_{zx} + H_{zy})}{\partial x} \tag{8.13b}$$

$$\mu_0 \frac{\partial H_{zx}}{\partial t} + \sigma_x^* H_{zx} = -\frac{\partial E_y}{\partial x} \tag{8.13c}$$

$$\mu_0 \frac{\partial H_{zy}}{\partial t} + \sigma_y^* H_{zy} = \frac{\partial E_x}{\partial y}. \tag{8.13d}$$

Similarly in the TM case, only the electric component is split. For example, the PML equations for TM_z (transverse magnetic to z) case are:

$$\mu_0 \frac{\partial H_x}{\partial t} + \sigma_y^* H_x = -\frac{\partial (E_{zx} + E_{zy})}{\partial y} \tag{8.14a}$$

$$\mu_0 \frac{\partial H_y}{\partial t} + \sigma_x^* H_y = \frac{\partial (E_{zx} + E_{zy})}{\partial x} \tag{8.14b}$$

$$\epsilon_0 \frac{\partial E_{zx}}{\partial t} + \sigma_x E_{zx} = \frac{\partial H_y}{\partial x} \tag{8.14c}$$

$$\epsilon_0 \frac{\partial E_{zy}}{\partial t} + \sigma_y E_{zy} = -\frac{\partial H_x}{\partial y}. \tag{8.14d}$$

Detailed specifications of conductivities in the face and corner regions of PML are depicted in Fig. 8.2.

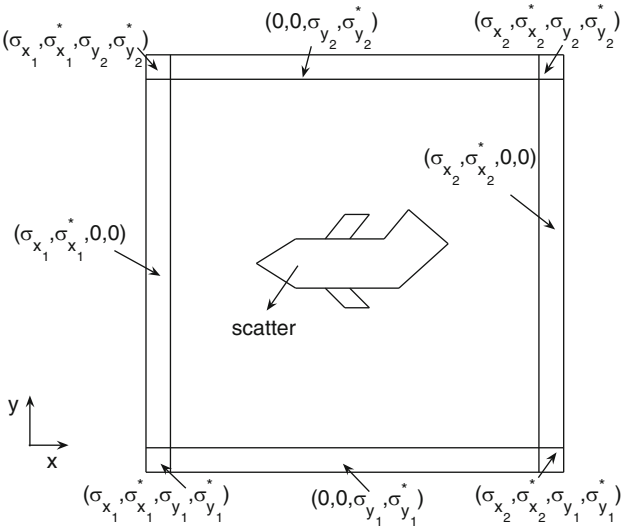

Fig. 8.2 Illustration of the 2-D PML

8.1.1.1 The PML Medium and Stretched Coordinates

Soon after Berenger's PML propoosal, researchers [74, 213, 246] found that the Berenger PML equations can be derived using a stretched coordinate approach.

Substituting the plane wave solution (8.2) into (8.1) and adding every two subequations together, we obtain the following six equations which are equivalent to (8.1):

$$j\omega\epsilon_0\tilde{E}_x = \frac{1}{s_y}\frac{\partial\tilde{H}_z}{\partial y} - \frac{1}{s_z}\frac{\partial\tilde{H}_y}{\partial z}, \tag{8.15a}$$

$$j\omega\epsilon_0\tilde{E}_y = \frac{1}{s_z}\frac{\partial\tilde{H}_x}{\partial z} - \frac{1}{s_x}\frac{\partial\tilde{H}_z}{\partial x}, \tag{8.15b}$$

$$j\omega\epsilon_0\tilde{E}_z = \frac{1}{s_x}\frac{\partial\tilde{H}_y}{\partial x} - \frac{1}{s_y}\frac{\partial\tilde{H}_x}{\partial y}, \tag{8.15c}$$

$$j\omega\mu_0\tilde{H}_x = -\frac{1}{s^*_y}\frac{\partial\tilde{E}_z}{\partial y} + \frac{1}{s^*_z}\frac{\partial\tilde{E}_y}{\partial z}, \tag{8.15d}$$

$$j\omega\mu_0\tilde{H}_y = -\frac{1}{s^*_z}\frac{\partial\tilde{E}_x}{\partial z} + \frac{1}{s^*_x}\frac{\partial\tilde{E}_z}{\partial x}, \tag{8.15e}$$

$$j\omega\mu_0\tilde{H}_z = -\frac{1}{s^*_x}\frac{\partial\tilde{E}_y}{\partial x} + \frac{1}{s^*_y}\frac{\partial\tilde{E}_x}{\partial y}, \tag{8.15f}$$

where the stretching parameters s_i and s^*_i are the same as (8.3).

Assuming that coefficients s_x, s_y, s_z vary with x, y, z, respectively, and introducing the following change of variables

$$dx' = s_x(x)dx, \quad dy' = s_y(y)dy, \quad dz' = s_z(z)dz, \tag{8.16}$$

we can rewrite the first three equations of (8.15) as:

$$j\omega\epsilon_0\tilde{E}_x = \frac{\partial\tilde{H}_z}{\partial y'} - \frac{\partial\tilde{H}_y}{\partial z'} \tag{8.17a}$$

$$j\omega\epsilon_0\tilde{E}_y = \frac{\partial\tilde{H}_x}{\partial z'} - \frac{\partial\tilde{H}_z}{\partial x'} \tag{8.17b}$$

$$j\omega\epsilon_0\tilde{E}_z = \frac{\partial\tilde{H}_y}{\partial x'} - \frac{\partial\tilde{H}_x}{\partial y'}. \tag{8.17c}$$

Changed into time domain with the stretched coordinates (8.16) and (8.17) becomes $\epsilon_0\frac{\partial\mathbf{E}}{\partial t} = \nabla' \times \mathbf{H}$, which has the same form as Ampere equations in vacuum.

By the same technique, the last three equations of (8.15) can be rewritten as:

$$j\omega\mu_0\tilde{H}_x = -\frac{\partial\tilde{E}_z}{\partial y''} + \frac{\partial\tilde{E}_y}{\partial z''} \tag{8.18a}$$

$$j\omega\mu_0\tilde{H}_y = -\frac{\partial\tilde{E}_x}{\partial z''} + \frac{\partial\tilde{E}_z}{\partial x''} \tag{8.18b}$$

$$j\omega\mu_0\tilde{H}_z = -\frac{\partial\tilde{E}_y}{\partial x''} + \frac{\partial\tilde{E}_x}{\partial y''}. \tag{8.18c}$$

In time domain, (8.18) can be rewritten as $\mu_0\frac{\partial\mathbf{H}}{\partial t} = -\nabla'' \times \mathbf{E}$ in the stretched coordinates $dx'' = s_x^*(x)dx, \quad dy'' = s_y^*(y)dy, \quad dz'' = s_z^*(z)dz$, and have exactly the same form as the Faraday equations in vacuum.

Hence, the original Berenger's split PML can be recast in a nonsplit form, which makes manipulating the PML equations easy and simplifies the understanding of the behavior of the PML. Furthermore, Berenger's split PML also offers an easy way to map the PML into other coordinate systems such as cylindrical and spherical coordinates [278].

8.1.2 The Convolutional PML

From the frequency domain equations (8.4), we can obtain the so-called convolutional PML, which was introduced by Roden and Gedney [248]. One important feature of the convolutional PML equations is that it is easy to generalize to

any physical medium, be it inhomogeneous, lossy, anisotropic, dispersive, or even nonlinear.

Transforming (8.15a) into time domain, we obtain [248]:

$$\epsilon_0 \frac{\partial E_x}{\partial t} = \overline{s_y}(t) * \frac{\partial H_z}{\partial y} - \overline{s_z}(t) * \frac{\partial H_y}{\partial z}, \tag{8.19}$$

where $*$ denotes the convolution product, and $\overline{s_y}(t)$ and $\overline{s_z}(t)$ are the inverse Laplace transforms of $\frac{1}{s_y(\omega)}$ and $\frac{1}{s_z(\omega)}$, respectively. For example, when the stretching factors $s_y(\omega)$ and $s_z(\omega)$ are given by (8.3), we have

$$\overline{s_i}(t) = \delta(t) + \xi_i(t), \quad i = y, z, \tag{8.20}$$

where $\delta(t)$ is the Dirac function, and

$$\xi_i(t) = -\frac{\sigma_i}{\epsilon_0} e^{-\frac{\sigma_i}{\epsilon_0} t} u(t), \quad i = y, z, \tag{8.21}$$

where $u(t)$ is the unit step function. Hence (8.19) can be rewritten as

$$\epsilon_0 \frac{\partial E_x}{\partial t} = \frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} + \xi_y(t) * \frac{\partial H_z}{\partial y} - \xi_z(t) * \frac{\partial H_y}{\partial z}. \tag{8.22}$$

Other five equations similar to (8.22) can be obtained from (8.4b)–(8.4f). Hence the convolutional PML is equivalent to the standard Maxwell's equations in vacuum, plus 12 convolutional terms.

8.1.3 The Uniaxial PML

Note that the split PML equations (8.1) differ from the standard Maxwell's equations, hence the split PML is often termed as non-Maxwellian in the literature. It renders implementation difficult and shows long term unstability [1,29,30], which prompted the development of other PMLs. The uniaxial PML [125,249] is one type of unsplit PMLs, and it can be derived from (8.4).

Let us introduce the following change of variables:

$$\hat{E}_x = s_x \tilde{E}_x, \quad \hat{E}_y = s_y \tilde{E}_y, \quad \hat{E}_z = s_z \tilde{E}_z, \tag{8.23}$$

$$\hat{H}_x = s_x^* \tilde{H}_x, \quad \hat{H}_y = s_y^* \tilde{H}_y, \quad \hat{H}_z = s_z^* \tilde{H}_z, \tag{8.24}$$

using which we can rewrite (8.4a) as

$$\omega\epsilon_0\frac{1}{s_x}\hat{E}_x = -\frac{1}{s_y s_z^*}k_y\hat{H}_z + \frac{1}{s_z s_y^*}k_z\hat{H}_y. \tag{8.25}$$

Assume that the matching condition (8.10) holds, i.e., $s_x = s_x^*, s_y = s_y^*, s_z = s_z^*$, then (8.25) can be rewritten as

$$\omega\epsilon_0\frac{s_y s_z}{s_x}\hat{E}_x = -k_y\hat{H}_z + k_z\hat{H}_y. \tag{8.26}$$

Similar equations can be obtained for the rest five equations of (8.4). The final system for the uniaxial PML in frequency domain can be written as [125]:

$$\omega\epsilon_0\tilde{\epsilon_s}\hat{\mathbf{E}} = -\mathbf{k}\times\hat{\mathbf{H}}, \quad \omega\mu_0\tilde{\mu_s}\hat{\mathbf{H}} = \mathbf{k}\times\hat{\mathbf{E}}, \tag{8.27}$$

where the tensors $\tilde{\epsilon_s}$ and $\tilde{\mu_s}$ are

$$\tilde{\epsilon_s} = \tilde{\mu_s} = \mathrm{diag}(\frac{s_y s_z}{s_x}, \frac{s_z s_x}{s_y}, \frac{s_x s_y}{s_z}). \tag{8.28}$$

To simplify the derivation for the uniaxial PML in time domain, below we drop the hat for all fields in (8.27). The first subequation of (8.27) can be written as

$$j\omega\epsilon_0\frac{s_y s_z}{s_x}E_x = \frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z}, \tag{8.29}$$

which can be rewritten into a system of two equations:

$$j\omega s_y D_x = \frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z}, \quad D_x = \epsilon_0\frac{s_z}{s_x}E_x. \tag{8.30}$$

Recall the explicit expressions (8.3) for $s_i, i = x, y, z$, we can transform (8.30) into time domain as:

$$\frac{\partial D_x}{\partial t} + \frac{\sigma_y}{\epsilon_0}D_x = \frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z}, \tag{8.31}$$

$$\epsilon_0\frac{\partial E_x}{\partial t} + \sigma_z E_x = \frac{\partial D_x}{\partial t} + \frac{\sigma_x}{\epsilon_0}D_x. \tag{8.32}$$

Similarly, from the second and third subequations of (8.27), we have

$$\frac{\partial D_y}{\partial t} + \frac{\sigma_z}{\epsilon_0}D_y = \frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x}, \tag{8.33}$$

$$\epsilon_0\frac{\partial E_y}{\partial t} + \sigma_x E_y = \frac{\partial D_y}{\partial t} + \frac{\sigma_y}{\epsilon_0}D_y. \tag{8.34}$$

and

$$\frac{\partial D_z}{\partial t} + \frac{\sigma_x}{\epsilon_0} D_z = \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y}, \tag{8.35}$$

$$\epsilon_0 \frac{\partial E_z}{\partial t} + \sigma_y E_z = \frac{\partial D_z}{\partial t} + \frac{\sigma_z}{\epsilon_0} D_z. \tag{8.36}$$

where we denote $D_y = \epsilon_0 \frac{s_x}{s_y} E_y$ and $D_z = \epsilon_0 \frac{s_y}{s_z} E_z$.

By the same technique, we can obtain another group of six time domain equations resulting from the second equation of (8.27). Note that (8.31), (8.33) and (8.35) can be written in vector form as

$$\frac{\partial \mathbf{D}}{\partial t} + \text{diag}(\frac{\sigma_y}{\epsilon_0}, \frac{\sigma_z}{\epsilon_0}, \frac{\sigma_x}{\epsilon_0})\mathbf{D} = \nabla \times \mathbf{H},$$

which is the Ampere equation in a lossy medium. Hence the uniaxial PML can be regarded as a lossy medium plus a set of six differential equations.

8.2 PMLs for Lossy Media

8.2.1 Split PML

A PML for lossy media was presented in [115]. Following [115], the PML equations for a lossy medium $(\epsilon, \mu, \sigma_0, \sigma_0^*)$ in a stretched coordinate can be written as (cf. (8.15) and (8.18)):

$$\nabla_s \times \tilde{\mathbf{H}} = j\omega\epsilon' \tilde{\mathbf{E}} \tag{8.37}$$

$$\nabla_{s^*} \times \tilde{\mathbf{E}} = -j\omega\mu' \tilde{\mathbf{H}} \tag{8.38}$$

where

$$\mu' = \mu + \frac{\sigma_0^*}{j\omega}, \quad \epsilon' = \epsilon + \frac{\sigma_0}{j\omega}, \tag{8.39}$$

and $\nabla_s = \hat{\mathbf{x}} \frac{1}{s_x} \frac{\partial}{\partial x} + \hat{\mathbf{y}} \frac{1}{s_y} \frac{\partial}{\partial y} + \hat{\mathbf{z}} \frac{1}{s_z} \frac{\partial}{\partial z}$. The stretching parameters are chosen as

$$s_i(i) = 1 + \frac{\sigma_i(i)}{j\omega\epsilon}, \quad s_i^*(i) = 1 + \frac{\sigma_i^*(i)}{j\omega\mu}, \quad i = x, y, z. \tag{8.40}$$

We can write the x component of (8.37) as

$$j\omega\epsilon(1+\frac{\sigma_0}{j\omega\epsilon})\tilde{E}_x=\frac{1}{s_y}\frac{\partial\tilde{H}_z}{\partial y}-\frac{1}{s_z}\frac{\partial\tilde{H}_y}{\partial z}, \tag{8.41}$$

which can be split into two subequations:

$$j\omega\epsilon(1+\frac{\sigma_0}{j\omega\epsilon})\tilde{E}_{xy}=\frac{1}{s_y}\frac{\partial\tilde{H}_z}{\partial y} \tag{8.42a}$$

$$j\omega\epsilon(1+\frac{\sigma_0}{j\omega\epsilon})\tilde{E}_{xz}=-\frac{1}{s_z}\frac{\partial\tilde{H}_y}{\partial z} \tag{8.42b}$$

where $\tilde{E}_{xy}$ and $\tilde{E}_{xz}$ are the two split components of $\tilde{E}_x$, i.e., $\tilde{E}_x=\tilde{E}_{xy}+\tilde{E}_{xz}$.

Substituting the definition (8.40) for s_y into (8.42a), we obtain

$$(j\omega\epsilon+\sigma_0+\sigma_y+\frac{\sigma_0\sigma_y}{j\omega\epsilon})\tilde{E}_{xy}=\frac{\partial\tilde{H}_z}{\partial y},$$

which can be rewritten in time domain as follows:

$$\epsilon\frac{\partial E_{xy}}{\partial t}+(\sigma_0+\sigma_y)E_{xy}+\frac{\sigma_0\sigma_y}{\epsilon}\int_{-\infty}^{t}E_{xy}dt'=\frac{\partial H_z}{\partial y}. \tag{8.43}$$

Similarly, from (8.37) we can obtain the rest five PML equations in frequency domain:

$$(j\omega\epsilon+\sigma_0+\sigma_z+\frac{\sigma_0\sigma_z}{j\omega\epsilon})\tilde{E}_{xz}=-\frac{\partial\tilde{H}_y}{\partial z}$$

$$(j\omega\epsilon+\sigma_0+\sigma_x+\frac{\sigma_0\sigma_x}{j\omega\epsilon})\tilde{E}_{yx}=-\frac{\partial\tilde{H}_z}{\partial x}$$

$$(j\omega\epsilon+\sigma_0+\sigma_z+\frac{\sigma_0\sigma_z}{j\omega\epsilon})\tilde{E}_{yz}=\frac{\partial\tilde{H}_x}{\partial z}$$

$$(j\omega\epsilon+\sigma_0+\sigma_x+\frac{\sigma_0\sigma_x}{j\omega\epsilon})\tilde{E}_{zx}=\frac{\partial\tilde{H}_y}{\partial x}$$

$$(j\omega\epsilon+\sigma_0+\sigma_y+\frac{\sigma_0\sigma_y}{j\omega\epsilon})\tilde{E}_{zy}=-\frac{\partial\tilde{H}_x}{\partial y}.$$

On the other hand, the x component of (8.38) yields

$$-j\omega(\mu+\frac{\sigma_0^*}{j\omega})\tilde{H}_x=\frac{1}{s_y^*}\frac{\partial\tilde{E}_z}{\partial y}-\frac{1}{s_z^*}\frac{\partial\tilde{E}_y}{\partial z}, \tag{8.44}$$

which can be split into two subequations:

$$-j\omega(\mu+\frac{\sigma_0^*}{j\omega})\tilde{H}_{xy}=\frac{1}{s_y^*}\frac{\partial\tilde{E}_z}{\partial y} \tag{8.45a}$$

$$-j\omega(\mu+\frac{\sigma_0^*}{j\omega})\tilde{H}_{xz}=-\frac{1}{s_z^*}\frac{\partial\tilde{E}_y}{\partial z} \tag{8.45b}$$

where $\tilde{H}_{xy}$ and $\tilde{H}_{xz}$ are the two components of $\tilde{H}_x$.

Substituting the stretching parameters into (8.45a) and (8.45b), respectively, we obtain

$$(j\omega\mu+\sigma_0^*+\sigma_y^*+\frac{\sigma_0^*\sigma_y^*}{j\omega})\tilde{H}_{xy}=-\frac{\partial\tilde{E}_z}{\partial y}$$

$$(j\omega\mu+\sigma_0^*+\sigma_z^*+\frac{\sigma_0^*\sigma_z^*}{j\omega})\tilde{H}_{xz}=\frac{\partial\tilde{E}_y}{\partial z}.$$

Similarly, the y and z components of (8.38) lead to the following four PML equations in frequency domain:

$$(j\omega\mu+\sigma_0^*+\sigma_z^*+\frac{\sigma_0^*\sigma_z^*}{j\omega})\tilde{H}_{yz}=-\frac{\partial\tilde{E}_x}{\partial z}$$

$$(j\omega\mu+\sigma_0^*+\sigma_x^*+\frac{\sigma_0^*\sigma_x^*}{j\omega})\tilde{H}_{yx}=\frac{\partial\tilde{E}_z}{\partial x}$$

$$(j\omega\mu+\sigma_0^*+\sigma_x^*+\frac{\sigma_0^*\sigma_x^*}{j\omega})\tilde{H}_{zx}=-\frac{\partial\tilde{E}_y}{\partial x}$$

$$(j\omega\mu+\sigma_0^*+\sigma_y^*+\frac{\sigma_0^*\sigma_y^*}{j\omega})\tilde{H}_{zy}=\frac{\partial\tilde{E}_x}{\partial y}.$$

Note that the PML parameters should satisfy the impedance matching conditions

$$\frac{\sigma_i}{\epsilon}=\frac{\sigma_i^*}{\mu},\quad i=x,y,z,$$

but σ_0 and σ_0^* do not need to satisfy $\frac{\sigma_0}{\epsilon}=\frac{\sigma_0^*}{\mu}$. Furthermore, if $\sigma_0=\sigma_0^*=0$, the PML equations obtained here reduce to the original Berenger PML. The corresponding time domain PML equations can be obtained similar to (8.43).

8.2.2 The Convolutional PML

The PML equation (8.41) can be rewritten in time domain as

$$\epsilon \frac{\partial E_x}{\partial t} + \sigma_0 E_x = \overline{s_y}(t) * \frac{\partial H_z}{\partial y} - \overline{s_z}(t) * \frac{\partial H_y}{\partial z}, \tag{8.46}$$

where (p. 335 of [248])

$$\overline{s_i}(t) = \frac{\delta(t)}{k_i} + \xi_i(t), \quad \xi_i(t) = -\frac{\sigma_i}{\epsilon k_i^2} e^{-\frac{1}{\epsilon}(\frac{\sigma_i}{k_i}+\alpha_i)t} u(t), \tag{8.47}$$

which is the inverse Laplace transform of the complex stretching factor proposed by Kuzuoglu and Mittra [173]:

$$s_i = k_i + \frac{\sigma_i}{\alpha_i + j\omega\epsilon}, \quad i = x, y, z, \tag{8.48}$$

where parameters $\sigma_i, \alpha_i > 0$ and $k_i \geq 1$.

Substituting (8.47) into (8.46) yields

$$\epsilon \frac{\partial E_x}{\partial t} + \sigma_0 E_x = \frac{1}{k_y}\frac{\partial H_z}{\partial y} - \frac{1}{k_z}\frac{\partial H_y}{\partial z} + \xi_y(t) * \frac{\partial H_z}{\partial y} - \xi_z(t) * \frac{\partial H_y}{\partial z}. \tag{8.49}$$

Similar equations can be obtained for the evolution of components E_y and E_z. The three equations for H components are the same as the convolutional PML matched to a vacuum.

8.2.3 The Uniaxial PML

A uniaxial PML for isotropic lossy media was presented in [125]. In frequency domain, the PML equations are

$$\omega\epsilon_0(1 + \frac{\sigma}{j\omega\epsilon_0})\tilde{\epsilon}_s\tilde{\mathbf{E}} = -\mathbf{k} \times \tilde{\mathbf{H}}, \quad \omega\mu_0\tilde{\mu}_s\tilde{\mathbf{H}} = \mathbf{k} \times \tilde{\mathbf{E}}, \tag{8.50}$$

where the tensors $\tilde{\epsilon}_s$ and $\tilde{\mu}_s$ are still defined by (8.28).

The x component of (8.50) can be written as

$$j\omega\epsilon_0(1 + \frac{\sigma}{j\omega\epsilon_0})\frac{s_y s_z}{s_x}\tilde{E}_x = \frac{\partial \tilde{H}_z}{\partial y} - \frac{\partial \tilde{H}_y}{\partial z}, \tag{8.51}$$

Introducing two new variables

$$\tilde{D}'_x = s_y \tilde{D}_x, \quad \tilde{D}_x = \epsilon_0 \frac{s_z}{s_x} \tilde{E}_x, \tag{8.52}$$

we can transform (8.51) into time domain equation as:

$$\frac{\partial D'_x}{\partial t} + \frac{\sigma}{\epsilon_0} D'_x = \frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z}, \tag{8.53}$$

plus the following two equations for the new variables:

$$\frac{\partial D_x}{\partial t} + \frac{\sigma_y}{\epsilon_0} D_x = \frac{\partial D'_x}{\partial t} \tag{8.54a}$$

$$\epsilon_0 \frac{\partial E_x}{\partial t} + \sigma_z E_x = \frac{\partial D_x}{\partial t} + \frac{\sigma_x}{\epsilon_0} D_x. \tag{8.54b}$$

Similarly, we can obtain two other groups of equations for components E_y and E_z. The magnetic field equation in the lossy uniaxial PML is exactly the same as that in the uniaxial PML matched to a vacuum.

8.2.4 *Time Derivative Lorentz Material Model*

In 1999, Ziolkowski [310] presented the time-derivative Lorentz material (TDLM) model as absorbing boundary conditions. Following the notation of [310], we assume that the PML fills a cubical simulation domain, the face regions have absorbing layers with only one normal direction; the edge regions are the joins of two face regions; and the corners are the overlapping parts of three face regions.

The corner region is reflectionless if both the relative permittivity and permeability tensor there are chosen to be

$$\Lambda_{xyz}(\omega) = \text{diag}\left(\frac{a_y(\omega)a_z(\omega)}{a_x(\omega)}, \frac{a_z(\omega)a_x(\omega)}{a_y(\omega)}, \frac{a_x(\omega)a_y(\omega)}{a_z(\omega)}\right), \tag{8.55}$$

where the coefficients

$$a_i(\omega) = 1 + \chi_i(\omega), \quad \chi_i(\omega) = \frac{\sigma_i}{j\omega}, \quad i = x, y, z.$$

Here $\sigma_i \geq 0$ represents the damping variation along the i-direction, where $i = x, y, z$.

Hence we have

$$\frac{a_y(\omega)a_z(\omega)}{a_x(\omega)} - 1 = \frac{[\chi_y(\omega) + \chi_z(\omega) - \chi_x(\omega)] + \chi_y(\omega)\chi_z(\omega)}{1 + \chi_x(\omega)}$$
$$= \frac{(j\omega)[(\sigma_y + \sigma_z) - \sigma_x] + \sigma_y\sigma_z}{-\omega^2 + j\omega\sigma_x} = \frac{P_x}{\epsilon_0 E_x}, \tag{8.56a}$$
$$\frac{a_z(\omega)a_x(\omega)}{a_y(\omega)} - 1 = \frac{[\chi_z(\omega) + \chi_x(\omega) - \chi_y(\omega)] + \chi_z(\omega)\chi_x(\omega)}{1 + \chi_y(\omega)}$$
$$= \frac{(j\omega)[(\sigma_z + \sigma_x) - \sigma_y] + \sigma_z\sigma_x}{-\omega^2 + j\omega\sigma_y} = \frac{P_y}{\epsilon_0 E_y}, \tag{8.56b}$$
$$\frac{a_x(\omega)a_y(\omega)}{a_z(\omega)} - 1 = \frac{[\chi_x(\omega) + \chi_y(\omega) - \chi_z(\omega)] + \chi_x(\omega)\chi_y(\omega)}{1 + \chi_z(\omega)}$$
$$= \frac{(j\omega)[(\sigma_x + \sigma_y) - \sigma_z] + \sigma_x\sigma_y}{-\omega^2 + j\omega\sigma_z} = \frac{P_z}{\epsilon_0 E_z}. \tag{8.56c}$$

Here P_i denotes the polarization component in the i-direction, where $i = x, y, z$.

Thus, the corresponding time domain equations for the polarizations in the PML corner region are

$$\frac{\partial^2 P_x}{\partial t^2} + \sigma_x \frac{\partial P_x}{\partial t} = \epsilon_0[(\sigma_y + \sigma_z) - \sigma_x]\frac{\partial E_x}{\partial t} + \epsilon_0\sigma_y\sigma_z E_x, \tag{8.57a}$$
$$\frac{\partial^2 P_y}{\partial t^2} + \sigma_y \frac{\partial P_y}{\partial t} = \epsilon_0[(\sigma_z + \sigma_x) - \sigma_y]\frac{\partial E_y}{\partial t} + \epsilon_0\sigma_z\sigma_x E_y, \tag{8.57b}$$
$$\frac{\partial^2 P_z}{\partial t^2} + \sigma_z \frac{\partial P_z}{\partial t} = \epsilon_0[(\sigma_x + \sigma_y) - \sigma_z]\frac{\partial E_z}{\partial t} + \epsilon_0\sigma_x\sigma_y E_z. \tag{8.57c}$$

Let us choose the x polarization current

$$J_x = \frac{\partial P_x}{\partial t} - \epsilon_0[(\sigma_y + \sigma_z) - \sigma_x]E_x. \tag{8.58}$$

Then using (8.57a), we obtain

$$\frac{\partial J_x}{\partial t} = \frac{\partial^2 P_x}{\partial t^2} - \epsilon_0[(\sigma_y + \sigma_z) - \sigma_x]\frac{\partial E_x}{\partial t} = -\sigma_x[J_x + \epsilon_0(\sigma_y + \sigma_z - \sigma_x)E_x] + \epsilon_0\sigma_y\sigma_z E_x,$$

i.e.,

$$\frac{\partial J_x}{\partial t} + \sigma_x J_x = \epsilon_0[-\sigma_x(\sigma_y + \sigma_z - \sigma_x) + \sigma_y\sigma_z]E_x = \epsilon_0(\sigma_y - \sigma_x)(\sigma_z - \sigma_x)E_x. \tag{8.59}$$

Substituting (8.58) into the Maxwell's equation

$$\frac{\partial E_x}{\partial t} = \frac{1}{\epsilon_0}(\frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z}) - \frac{1}{\epsilon_0}\frac{\partial P_x}{\partial t}$$

yields

$$\frac{\partial E_x}{\partial t} + (\sigma_y + \sigma_z - \sigma_x)E_x = \frac{1}{\epsilon_0}(\frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z}) - \frac{1}{\epsilon_0}J_x. \tag{8.60}$$

We can treat y and z directions similarly. In summary, we have

$$\frac{\partial J_x}{\partial t} + \sigma_x J_x = \epsilon_0(\sigma_y - \sigma_x)(\sigma_z - \sigma_x)E_x, \tag{8.61a}$$

$$\frac{\partial E_x}{\partial t} + (\sigma_y + \sigma_z - \sigma_x)E_x = \frac{1}{\epsilon_0}(\frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z}) - \frac{1}{\epsilon_0}J_x, \tag{8.61b}$$

$$\frac{\partial J_y}{\partial t} + \sigma_y J_y = \epsilon_0(\sigma_x - \sigma_y)(\sigma_z - \sigma_y)E_y, \tag{8.61c}$$

$$\frac{\partial E_y}{\partial t} + (\sigma_z + \sigma_x - \sigma_y)E_y = \frac{1}{\epsilon_0}(\frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x}) - \frac{1}{\epsilon_0}J_y, \tag{8.61d}$$

$$\frac{\partial J_z}{\partial t} + \sigma_z J_z = \epsilon_0(\sigma_x - \sigma_z)(\sigma_y - \sigma_z)E_z, \tag{8.61e}$$

$$\frac{\partial E_z}{\partial t} + (\sigma_x + \sigma_y - \sigma_z)E_z = \frac{1}{\epsilon_0}(\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y}) - \frac{1}{\epsilon_0}J_z. \tag{8.61f}$$

Similarly, for the magnetization M_x in the x direction we have

$$\frac{\partial^2 M_x}{\partial t^2} + \sigma_x\frac{\partial M_x}{\partial t} = \mu_0[(\sigma_y + \sigma_z) - \sigma_x]\frac{\partial H_x}{\partial t} + \mu_0(\sigma_y\sigma_z)H_x.$$

Choosing the magnetization current K_x in the x direction as:

$$K_x = \frac{\partial M_x}{\partial t} - \mu_0[(\sigma_y + \sigma_z) - \sigma_x]H_x,$$

we obtain

$$\frac{\partial K_x}{\partial t} + \sigma_x K_x = \mu_0(\sigma_y - \sigma_x)(\sigma_z - \sigma_x)H_x. \tag{8.62}$$

Then coupling (8.62) with the Maxwell's equation

$$-\nabla \times \mathbf{E} = \frac{\partial}{\partial t}(\mu_0\mathbf{H}) + \frac{\partial \mathbf{M}}{\partial t},$$

we finally have

$$\frac{\partial H_x}{\partial t} + (\sigma_y + \sigma_z - \sigma_x)H_x = -\frac{1}{\mu_0}(\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z}) - \frac{1}{\mu_0}K_x.$$

Similar equations can be derived in the y and z directions.
In all, the complete PML governing equations for the corner region are

$$\frac{\partial \mathbf{E}}{\partial t} + D_1\mathbf{E} = \frac{1}{\epsilon_0}\nabla \times \mathbf{H} - \frac{1}{\epsilon_0}\mathbf{J} \tag{8.63}$$

$$\frac{\partial \mathbf{J}}{\partial t} + D_2\mathbf{J} = \epsilon_0 D_3\mathbf{E} \tag{8.64}$$

$$\frac{\partial \mathbf{H}}{\partial t} + D_1\mathbf{H} = -\frac{1}{\mu_0}\nabla \times \mathbf{E} - \frac{1}{\mu_0}\mathbf{K} \tag{8.65}$$

$$\frac{\partial \mathbf{K}}{\partial t} + D_2\mathbf{K} = \mu_0 D_3\mathbf{H} \tag{8.66}$$

where $D_i, i = 1, 2, 3$, are 3×3 diagonal matrices shown below:

$$D_1 = \text{diag}(\sigma_y + \sigma_z - \sigma_x, \sigma_z + \sigma_x - \sigma_y, \sigma_x + \sigma_y - \sigma_z),$$
$$D_2 = \text{diag}(\sigma_x, \sigma_y, \sigma_z),$$
$$D_3 = \text{diag}((\sigma_x - \sigma_y)(\sigma_x - \sigma_z), (\sigma_y - \sigma_x)(\sigma_y - \sigma_z), (\sigma_z - \sigma_x)(\sigma_z - \sigma_y)).$$

Usually, quadratic profiles are chosen for the damping functions σ_x, σ_y and σ_z.

This TDLM PML model has a very nice feature: the governing equations for the corner region automatically reduces to those equations for the face and edge regions when the corresponding material coefficients become zero. For example, setting $\sigma_y = 0$ and the y-component of $\mathbf{J}$ zero in Eqs. (8.63)–(8.66) yields the PML equations in the xz edge region; setting $\sigma_y = \sigma_x = 0$ and $\mathbf{J} = (0, 0, \mathbf{J}_z)'$ (i.e., only z-component of $\mathbf{J}$ is nonzero) in Eqs. (8.63)–(8.66) gives the PML equations in the z-directed face region. Hence the set of PML equations (8.63)–(8.66) covers all PML regions. Furthermore, the set of PML equations (8.63)–(8.66) automatically reduces to the standard Maxwell's equations on a bounded domain by setting $D_1 = D_2 = D_3 = 0$ and interpreting $\mathbf{J}$ and $\mathbf{K}$ as given current sources. Hence the analysis of this PML model is very interesting, since the results derived from the TDLM PML model automatically cover the Maxwell's equations in free space. Some finite element schemes for solving this PML model were developed in [152].

8.3 PMLs for Dispersive Media and Metamaterials

The absorbing boundary condition is required to truncate the computational domain without reflection in simulating wave propagation in metamaterials. However, the standard PML is inherently unstable when it is extended to truncate the boundary of metamaterials without modification [88, 95, 102]. In this section, we present some PMLs developed for modeling wave propagation in dispersive media and metamaterials.

8.3.1 Complex Frequency-Shifted Technique

The complex frequency-shifted PML (CFS-PML) was introduced in [173], and its main idea is to shift poles off the real axis. Many experiments have shown that CFS-PML is quite general when applied to various dispersive media, and is also very efficient in attenuating evanescent waves and reducing late-time reflections [36, 248]. The material for this subsection is essentially from [241].

In frequency domain, the PML Maxwell's equations can be written as:

$$\nabla \times \tilde{\mathbf{H}} = j\omega\epsilon_0[\frac{\sigma(\mathbf{r})}{j\omega\epsilon_0} + \epsilon_r(\mathbf{r},\omega)]A \cdot \tilde{\mathbf{E}}, \tag{8.67}$$

$$\nabla \times \tilde{\mathbf{E}} = -j\omega\mu_0\mu_r(\mathbf{r},\omega)A \cdot \tilde{\mathbf{H}}, \tag{8.68}$$

where $A = \text{diag}\{\frac{\xi_x}{\xi_y\xi_z}, \frac{\xi_y}{\xi_z\xi_x}, \frac{\xi_z}{\xi_x\xi_y}\}$ is the material tensor. For dispersive media such as Debye, Drude, and Lorentz types, we define a general stretching parameter

$$\xi_i = 1/(k_i + \frac{\sigma_i}{\gamma + j\omega}), \quad i = x, y, z. \tag{8.69}$$

Here parameters k_i and σ_i provide additional attenuation to both propagating and evanescent waves.

8.3.1.1 Debye Media

First we consider the case of Debye medium of order N, whose relative permittivity in the frequency domain is defined as

$$\epsilon_r(\omega) = \epsilon_{r,\infty} + \sum_{p=1}^{N} \frac{\epsilon_{r,sp} - \epsilon_{r,\infty}}{1 + j\omega\tau_p}, \tag{8.70}$$

where $\epsilon_{r,\infty}$ and $\epsilon_{r,sp}$ are the relative permittivities at infinite and zero frequencies for the p-th pole, respectively, and τ_p is the relaxation time of the p-th pole.

Note that the Ampere's law (8.67) can be written as

$$\nabla \times \tilde{\mathbf{H}} = j\omega\epsilon_0\epsilon_{r,\infty}A \cdot \tilde{\mathbf{E}} + \sigma A \cdot \tilde{\mathbf{E}} + j\omega \sum_{p=1}^{N} \tilde{\mathbf{Q}}_{e,p}, \tag{8.71}$$

where $\tilde{\mathbf{Q}}_{e,p}$ (subscripts e and p stand for the electric dispersion and pole p, respectively) is defined as

$$\tilde{\mathbf{Q}}_{e,p} = \frac{\epsilon_0(\epsilon_{r,sp} - \epsilon_{r,\infty})}{1 + j\omega\tau_p} A \cdot \tilde{\mathbf{E}}. \tag{8.72}$$

We define a new variable $\tilde{\mathbf{R}} = A \cdot \tilde{\mathbf{E}}$, and rewrite (8.71) as

$$\nabla \times \tilde{\mathbf{H}} = j\omega\epsilon_0\epsilon_{r,\infty}\tilde{\mathbf{R}} + \sigma\tilde{\mathbf{R}} + j\omega \sum_{p=1}^{N} \tilde{\mathbf{Q}}_{e,p}, \tag{8.73}$$

which in time domain is equivalent to

$$\nabla \times \mathbf{H} = \epsilon_0\epsilon_{r,\infty} \frac{d\mathbf{R}}{dt} + \sigma\mathbf{R} + \sum_{p=1}^{N} \frac{d\mathbf{Q}_{e,p}}{dt}. \tag{8.74}$$

Similarly, we can transform (8.72) into time domain as

$$\mathbf{Q}_{e,p} + \tau_p \frac{d\mathbf{Q}_{e,p}}{dt} = \epsilon_0(\epsilon_{r,sp} - \epsilon_{r,\infty})\mathbf{R}. \tag{8.75}$$

By the definition of $\tilde{\mathbf{R}}$, we have

$$\tilde{R}_x = \frac{\xi_x}{\xi_y\xi_z}\tilde{E}_x, \quad \tilde{R}_y = \frac{\xi_y}{\xi_z\xi_x}\tilde{E}_y, \quad \tilde{R}_z = \frac{\xi_z}{\xi_x\xi_y}\tilde{E}_z. \tag{8.76}$$

We introduce a new variable $\tilde{\mathbf{S}}$, whose components are

$$\tilde{S}_x = \frac{\xi_x}{\xi_y}\tilde{E}_x, \quad \tilde{S}_y = \frac{\xi_y}{\xi_z}\tilde{E}_y, \quad \tilde{S}_z = \frac{\xi_z}{\xi_x}\tilde{E}_z. \tag{8.77}$$

Transforming (8.76) and (8.77) into time domain yields

$$k_x \frac{dS_x}{dt} + (k_x\gamma + \sigma_x)S_x = k_y \frac{dE_x}{dt} + (k_y\gamma + \sigma_y)E_x, \tag{8.78}$$

$$\frac{dR_x}{dt} + \gamma R_x = k_z \frac{dS_x}{dt} + (k_z\gamma + \sigma_z)S_x. \tag{8.79}$$

Similar equations can be obtained for other components.

To consider the magnetic components, we assume that $\mu_r = 1$ and $\gamma = 0$ in (8.69). Hence, the x-component of Faraday's law (8.68) can be written as

$$(\nabla \times \tilde{\mathbf{E}})_x = -j\omega \frac{\tilde{B}_x}{\xi_z}, \quad \tilde{B}_x = \mu_0 \frac{\xi_x}{\xi_y}\tilde{H}_x, \tag{8.80}$$

which in time domain become

$$(\nabla \times \mathbf{E})_x = -k_z \frac{dB_x}{dt} - \sigma_z B_x, \tag{8.81}$$

$$\sigma_x B_x + k_x \frac{dB_x}{dt} = \mu_0 (k_y \frac{dH_x}{dt} + \sigma_y H_x). \tag{8.82}$$

Similar equations can be obtained for other components.

In [241], Prokopidis developed a FDTD method to solve the Debye PML model in the following order: $R \rightarrow Q_{e,p} \rightarrow S \rightarrow E \rightarrow B \rightarrow H$.

8.3.1.2 Drude Media

The Drude media can be described by the complex permittivity:

$$\epsilon_r(\omega) = 1 + \frac{\omega_p^2}{j\omega\nu - \omega^2}, \tag{8.83}$$

where ω_p is the plasma frequency and ν is the collision frequency.

For Drude media, the corresponding Ampere's law (8.67) can be written as

$$\nabla \times \tilde{\mathbf{H}} = j\omega\epsilon_0 A \cdot \tilde{\mathbf{E}} + \sigma A \cdot \tilde{\mathbf{E}} + \tilde{\mathbf{Q}}, \quad \tilde{\mathbf{Q}} = \frac{\epsilon_0 \omega_p^2}{j\omega + \nu} A \cdot \tilde{\mathbf{E}}. \tag{8.84}$$

By introducing the new variable $\tilde{\mathbf{R}} = A \cdot \tilde{\mathbf{E}}$, we can transform (8.84) into time domain as follows:

$$\nabla \times \mathbf{H} = \epsilon_0 \frac{d\mathbf{R}}{dt} + \sigma \mathbf{R} + \mathbf{Q}, \tag{8.85}$$

$$\frac{d\mathbf{Q}}{dt} + \nu \mathbf{Q} = \epsilon_0 \omega_p^2 \mathbf{R}. \tag{8.86}$$

The rest governing equations are the same as those described above for the Debye model.

8.3.1.3 Lorentz Media

For the Lorentz medium of order N, the relative permittivity in the frequency domain is described by

$$\epsilon_r(\omega) = \epsilon_{r,\infty} + (\epsilon_{r,s} - \epsilon_{r,\infty}) \sum_{p=1}^{N} \frac{G_p \omega_p^2}{\omega_p^2 + 2j\omega\nu_p - \omega^2}, \tag{8.87}$$

where $G_p \geq 0$ and $\sum_{p=1}^{N} G_p = 1$.

If we define a new variable

$$\tilde{\mathbf{Q}}_p = \epsilon_0(\epsilon_{r,s} - \epsilon_{r,\infty})\frac{G_p\omega_p^2}{\omega_p^2 + 2j\omega\nu_p - \omega^2} A \cdot \tilde{\mathbf{E}}, \tag{8.88}$$

then Ampere's law (8.67) can be written as

$$\nabla \times \tilde{\mathbf{H}} = \sigma A \cdot \tilde{\mathbf{E}} + j\omega\epsilon_0\epsilon_{r,\infty} A \cdot \tilde{\mathbf{E}} + j\omega \sum_{p=1}^{N} \tilde{\mathbf{Q}}_p, \tag{8.89}$$

which in time domain becomes

$$\nabla \times \mathbf{H} = \sigma\mathbf{R} + \epsilon_0\epsilon_{r,\infty}\frac{d\mathbf{R}}{dt} + \sum_{p=1}^{N} \frac{d\mathbf{Q}_p}{dt}. \tag{8.90}$$

Transforming (8.88) into time domain, we have

$$\frac{d^2\mathbf{Q}_p}{dt^2} + 2\nu_p\frac{d\mathbf{Q}_p}{dt} + \omega_p^2\mathbf{Q}_p = \epsilon_0(\epsilon_{r,s} - \epsilon_{r,\infty})G_p\omega_p^2\mathbf{R}. \tag{8.91}$$

The rest PML equations are the same as those for the Debye model.

8.3.2 *Complex-Coordinate Stretching*

Here we introduce a modified PML for metamaterials [262] obtained by the complex-coordinate stretching technique [74]. For simplicity, below we present the derivation for 2-D TMz Maxwell's equations in metamaterials described by the Lorentz medium model:

$$\epsilon(\omega) = \epsilon_0(1 + \frac{\omega_{pe}^2}{\omega_{0e}^2 - \omega^2 + j\Gamma_e\omega}), \tag{8.92}$$

$$\mu(\omega) = \mu_0(1 + \frac{\omega_{pm}^2}{\omega_{0m}^2 - \omega^2 + j\Gamma_m\omega}). \tag{8.93}$$

The rest of this subsection is mainly based on [262].

8.3.2.1 Derivation of TMz Maxwell's Equations in Metamaterials

Since only three nonzero fields H_x, H_y, E_z exist in TMz case, from Maxwell's equations

$$\frac{\partial \mathbf{B}}{\partial t} = -\nabla \times \mathbf{E}, \quad \frac{\partial \mathbf{D}}{\partial t} = \nabla \times \mathbf{H},$$

we can obtain the TMz Maxwell's equations in frequency domain:

$$j\omega B_x = -\frac{\partial E_z}{\partial y} \tag{8.94}$$

$$j\omega B_y = \frac{\partial E_z}{\partial x} \tag{8.95}$$

$$j\omega D_z = \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} \tag{8.96}$$

where the constitutive equations are

$$D_z = \epsilon(\omega) E_z, \quad B_x = \mu(\omega) H_x, \quad B_y = \mu(\omega) H_y.$$

Substituting (8.93) into (8.94), we obtain

$$j\omega H_x + K_x = -\frac{1}{\mu_0}\frac{\partial E_z}{\partial y}, \tag{8.97}$$

where $K_x = \frac{j\omega\omega_{pm}^2}{\omega_{0m}^2 - \omega^2 + j\Gamma_m\omega} H_x$, which can be rewritten as

$$(\frac{\omega_{0m}^2}{j\omega} + j\omega + \Gamma_m) K_x = \omega_{pm}^2 H_x,$$

or equivalent to

$$j\omega K_x + \Gamma_m K_x = \omega_{pm}^2 H_x - \omega_{0m}^2 F_x, \tag{8.98}$$

where F_x is a new auxiliary variable defined as

$$F_x = \frac{1}{j\omega} K_x. \tag{8.99}$$

By the same technique, from (8.95) we have

$$j\omega H_y + K_y = \frac{1}{\mu_0}\frac{\partial E_z}{\partial x} \tag{8.100}$$

$$j\omega K_y + \Gamma_m K_y = \omega_{pm}^2 H_y - \omega_{0m}^2 F_y \tag{8.101}$$

$$j\omega F_y = K_y. \tag{8.102}$$

Similarly, substituting (8.92) into (8.96), we obtain

$$j\omega E_z + J_z = \frac{1}{\epsilon_0}\left(\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y}\right), \tag{8.103}$$

where $J_z = \frac{j\omega\omega_{pe}^2}{\omega_{0e}^2 - \omega^2 + j\Gamma_e\omega} E_z$, which can be rewritten as

$$\left(\frac{\omega_{0e}^2}{j\omega} + j\omega + \Gamma_e\right) J_z = \omega_{pe}^2 E_z,$$

or

$$j\omega J_z + \Gamma_e J_z = \omega_{pe}^2 E_z - \omega_{0e}^2 R_z, \tag{8.104}$$

where the auxiliary variable R_z is defined as

$$R_z = \frac{1}{j\omega} J_z. \tag{8.105}$$

Changing the above equations into time domain, we have the TMz Maxwell's equations in metamaterials:

$$\frac{\partial H_x}{\partial t} = -\frac{1}{\mu_0}\frac{\partial E_z}{\partial y} - K_x, \tag{8.106}$$

$$\frac{\partial H_y}{\partial t} = \frac{1}{\mu_0}\frac{\partial E_z}{\partial x} - K_y, \tag{8.107}$$

$$\frac{\partial E_z}{\partial t} = \frac{1}{\epsilon_0}\left(\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y}\right) - J_z, \tag{8.108}$$

supplemented by the following auxiliary constitutive equations:

$$\frac{\partial K_x}{\partial t} + \Gamma_m K_x = \omega_{pm}^2 H_x - \omega_{0m}^2 F_x \tag{8.109}$$

$$\frac{\partial F_x}{\partial t} = K_x \tag{8.110}$$

$$\frac{\partial K_y}{\partial t} + \Gamma_m K_y = \omega_{pm}^2 H_y - \omega_{0m}^2 F_y \tag{8.111}$$

$$\frac{\partial F_y}{\partial t} = K_y \tag{8.112}$$

$$\frac{\partial J_z}{\partial t} + \Gamma_e J_z = \omega_{pe}^2 E_z - \omega_{0e}^2 R_z \tag{8.113}$$

$$\frac{\partial R_z}{\partial t} = J_z. \tag{8.114}$$

8.3.2.2 The PML Equations for Metamaterials

Now we want to derive a stable non-split PML model for metamaterials. Following [262], we introduce the following complex-coordinate stretching variables:

$$\frac{\partial}{\partial x} \Rightarrow 1/[1+\frac{\sigma_x}{j\omega(1+\frac{\omega_{pe}^2}{\omega_{0e}^2-\omega^2+j\Gamma_e\omega})}]\frac{\partial}{\partial x}, \tag{8.115}$$

$$\frac{\partial}{\partial y} \Rightarrow 1/[1+\frac{\sigma_y}{j\omega(1+\frac{\omega_{pe}^2}{\omega_{0e}^2-\omega^2+j\Gamma_e\omega})}]\frac{\partial}{\partial y}, \tag{8.116}$$

and define the magnetic and electric field variables for the metamaterial PML region:

$$\tilde{H}_x = 1/[1+\frac{\sigma_y}{j\omega(1+\frac{\omega_{pe}^2}{\omega_{0e}^2-\omega^2+j\Gamma_e\omega})}]H_x \tag{8.117}$$

$$\tilde{H}_y = 1/[1+\frac{\sigma_x}{j\omega(1+\frac{\omega_{pe}^2}{\omega_{0e}^2-\omega^2+j\Gamma_e\omega})}]H_y \tag{8.118}$$

$$\tilde{E}_{z1} = 1/[1+\frac{\sigma_y}{j\omega(1+\frac{\omega_{pe}^2}{\omega_{0e}^2-\omega^2+j\Gamma_e\omega})}]E_z \tag{8.119}$$

$$\tilde{E}_{z2} = 1/[1+\frac{\sigma_x}{j\omega(1+\frac{\omega_{pe}^2}{\omega_{0e}^2-\omega^2+j\Gamma_e\omega})}]E_z. \tag{8.120}$$

With these definitions, we can easily obtain the unsplit PML equations for metamaterials:

$$\frac{\partial H_x}{\partial t} = -\frac{1}{\mu_0}\frac{\partial \tilde{E}_{z1}}{\partial y} - K_x \tag{8.121}$$

$$\frac{\partial H_y}{\partial t} = \frac{1}{\mu_0}\frac{\partial \tilde{E}_{z2}}{\partial x} - K_y \tag{8.122}$$

$$\frac{\partial E_z}{\partial t} = \frac{1}{\epsilon_0}(\frac{\partial \tilde{H}_y}{\partial x} - \frac{\partial \tilde{H}_x}{\partial y}) - J_z, \tag{8.123}$$

plus many constitutive equations we shall derive below.

Note that (8.117) is same as $[1+\frac{\sigma_y}{j\omega(1+\frac{\omega_{pe}^2}{\omega_{0e}^2-\omega^2+j\Gamma_e\omega})}]\tilde{H}_x = H_x$, which can be written as

$$\tilde{H}_x + \sigma_y P_x = H_x. \tag{8.124}$$

If we introduce a new variable $P_x = \frac{1}{j\omega(1+\frac{\omega_{pe}^2}{\omega_{0e}^2-\omega^2+j\Gamma_e\omega})}\tilde{H}_x$, which is same as

$$j\omega(1+\frac{\omega_{pe}^2}{\omega_{0e}^2-\omega^2+j\Gamma_e\omega})P_x = \tilde{H}_x,$$

or equivalently

$$j\omega P_x + \omega_{pe}^2 Q_x = \tilde{H}_x, \tag{8.125}$$

if we introduce another new variable $Q_x = \frac{j\omega}{\omega_{0e}^2-\omega^2+j\Gamma_e\omega}P_x$, which is same as

$$(\frac{\omega_{0e}^2}{j\omega} + j\omega + \Gamma_e)Q_x = P_x,$$

or

$$(j\omega + \Gamma_e)Q_x = P_x - \omega_{0e}^2 U_x, \;\; U_x = \frac{1}{j\omega}Q_x. \tag{8.126}$$

We can write (8.125) and (8.126) in time domain as:

$$\frac{\partial P_x}{\partial t} + \omega_{pe}^2 Q_x = \tilde{H}_x \tag{8.127}$$

$$\frac{\partial Q_x}{\partial t} + \Gamma_e Q_x = P_x - \omega_{0e}^2 U_x \tag{8.128}$$

$$\frac{\partial U_x}{\partial t} = Q_x. \tag{8.129}$$

By similar techniques, we can obtain the rest auxiliary time domain equations:

$$\tilde{H}_y = H_y - \sigma_x P_y \tag{8.130}$$

$$\frac{\partial P_y}{\partial t} = \tilde{H}_y - \omega_{pe}^2 Q_y \tag{8.131}$$

$$\frac{\partial Q_y}{\partial t} = P_y - \omega_{0e}^2 U_y - \Gamma_e Q_y \tag{8.132}$$

$$\frac{\partial U_y}{\partial t} = Q_y \tag{8.133}$$

$$\tilde{E}_{z1} = E_z - \sigma_y D_{z1} \tag{8.134}$$

$$\frac{\partial D_{z1}}{\partial t} = \tilde{E}_{z1} - \omega_{pe}^2 B_{z1} \tag{8.135}$$

$$\frac{\partial B_{z1}}{\partial t} = D_{z1} - \omega_{0e}^2 C_{z1} - \Gamma_e B_{z1} \tag{8.136}$$

$$\frac{\partial C_{z1}}{\partial t} = B_{z1} \tag{8.137}$$

$$\tilde{E}_{z2} = E_z - \sigma_x D_{z2} \tag{8.138}$$

$$\frac{\partial D_{z2}}{\partial t} = \tilde{E}_{z2} - \omega_{pe}^2 B_{z2} \tag{8.139}$$

$$\frac{\partial B_{z2}}{\partial t} = D_{z2} - \omega_{0e}^2 C_{z2} - \Gamma_e B_{z2} \tag{8.140}$$

$$\frac{\partial C_{z2}}{\partial t} = B_{z2}. \tag{8.141}$$

Note that in the above non-split metamaterial PML model, those terms involving temporal and spatial derivatives are exactly the same as those from the standard Maxwell's equations, which makes the PML implementation quite simple. Furthermore, when $\sigma_x = \sigma_y = 0$, the metamaterial PML equations reduce to the standard Maxwell's equations.

8.4 Bibliographical Remarks

In this chapter, we reviewed many PML models developed since Berenger introduced the PML concept in 1994. Considering the technicality of mathematical analysis of PMLs, we didn't cover the theoretical analysis of those PML models. Interested readers can consult Sect. 13.5 of Monk's book [217] for works published before 2002. More recent works on finite element analysis of PMLs can be found in [26, 49, 69] and references cited therein. Some interesting topics worth further exploration are rigorous mathematical analysis and finite element applications of those PML models coupled to metamaterials.

Chapter 9
Simulations of Wave Propagation in Metamaterials

In this chapter, we present some interesting simulations of wave propagation in metamaterials. We start in Sect. 9.1 with a perfectly matched layer model, which allows us to reduce the simulation on an infinite domain to be realized on a bounded domain. Here we present a simulation demonstrating the negative refraction index phenomenon for metamaterials. In Sects. 9.2 and 9.3, we present invisibility cloak simulations using metamaterials in frequency domain and time domain, respectively. In Sect. 9.4, we present an interesting application of metamaterials for solar cell design. In Sect. 9.5, we end this chapter by presenting some open mathematical problems related to metamaterials.

9.1 Interesting Phenomena of Wave Propagation in Metamaterials

9.1.1 Demonstration of a PML Model

First, we want to demonstrate the role of a perfectly matched layer (PML) model developed by Ziolkowski [310] in 1999 (see Chap. 8). Following the notation of [310], we assume that the PML is a cubical domain. The complete PML governing equations on the corner region are given by (cf. Sect. 8.2.4):

$$\frac{\partial \mathbf{E}}{\partial t} + D_1\mathbf{E} = \frac{1}{\epsilon_0}\nabla \times \mathbf{H} - \frac{1}{\epsilon_0}\mathbf{J}, \tag{9.1}$$

$$\frac{\partial \mathbf{J}}{\partial t} + D_2\mathbf{J} = \epsilon_0 D_3\mathbf{E}, \tag{9.2}$$

$$\frac{\partial \mathbf{H}}{\partial t} + D_1\mathbf{H} = -\frac{1}{\mu_0}\nabla \times \mathbf{E} - \frac{1}{\mu_0}\mathbf{K}, \tag{9.3}$$

J. Li and Y. Huang, *Time-Domain Finite Element Methods for Maxwell's Equations in Metamaterials*, Springer Series in Computational Mathematics 43,
DOI 10.1007/978-3-642-33789-5_9, © Springer-Verlag Berlin Heidelberg 2013

$$\frac{\partial \mathbf{K}}{\partial t} + D_2\mathbf{K} = \mu_0 D_3 \mathbf{H}. \tag{9.4}$$

Recall that the 3×3 diagonal matrices

$$\begin{aligned} D_1 &= diag(\sigma_y + \sigma_z - \sigma_x, \sigma_z + \sigma_x - \sigma_y, \sigma_x + \sigma_y - \sigma_z), \\ D_2 &= diag(\sigma_x, \sigma_y, \sigma_z), \\ D_3 &= diag((\sigma_x - \sigma_y)(\sigma_x - \sigma_z), (\sigma_y - \sigma_x)(\sigma_y - \sigma_z), (\sigma_z - \sigma_x)(\sigma_z - \sigma_y)), \end{aligned}$$

where σ_x, σ_y and σ_z are nonnegative functions and represent the damping variations along the x, y and z directions, respectively. Usually, quadratic profiles are chosen for σ_x, σ_y and σ_z [284, 310].

Note that the model (9.1)–(9.4) is the same as (5.12) of Turkel and Yefet [284] (assuming $\epsilon_0 = \mu_0 = 1$) and is well-posed mathematically because it is a symmetric hyperbolic system plus lower order terms [284, p. 545].

Since the PML model (9.1)–(9.4) is very similar to the metamaterial Drude model solved by the nodal discontinuous Galerkin method discussed in Sect. 4.4, we can easily solve the PML model (9.1)–(9.4) by the nodal discontinuous Galerkin method. Details can be found in the original paper [185].

For simplicity, here we just solve a 2-D transverse magnetic PML model, which can be obtained from (9.1) to (9.4):

$$\frac{\partial H_x}{\partial t} = -\frac{\partial E_z}{\partial y} - K_x + (\sigma_x - \sigma_y) H_x, \tag{9.5}$$

$$\frac{\partial H_y}{\partial t} = \frac{\partial E_z}{\partial x} - K_y - (\sigma_x - \sigma_y) H_y, \tag{9.6}$$

$$\frac{\partial E_z}{\partial t} = \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} - J_z - (\sigma_x + \sigma_y) E_z, \tag{9.7}$$

$$\frac{\partial J_z}{\partial t} = \sigma_x \sigma_y E_z, \tag{9.8}$$

$$\frac{\partial K_x}{\partial t} = -\sigma_x K_x + (\sigma_x - \sigma_y)\sigma_x H_x, \tag{9.9}$$

$$\frac{\partial K_y}{\partial t} = -\sigma_y K_y - (\sigma_x - \sigma_y)\sigma_y H_y, \tag{9.10}$$

where the subscripts 'x, y' and 'z' denote the corresponding components.

For this 2-D PML model, we assume that the physical domain $\Omega = (-1, 1)^2$ is surrounded by a perfectly matching layer of thickness of 0.2, which makes the real computational domain $(-1.2, 1.2)^2$. The initial electric source is given as

$$E_z(x, y, 0) = \begin{cases} \cos^6(\frac{\pi r}{2r_0}) & \text{if } r \le r_0 \\ 0 & \text{if } r \ge r_0 \end{cases} \tag{9.11}$$

where $r_0 = 0.5, r = \sqrt{x^2 + y^2}$. Furthermore, the damping function σ_x is chosen as:

$$\sigma_x(x) = \begin{cases} \sigma_0(x-1)^2 & \text{if } x \geq 1 \\ \sigma_0(x+1)^2 & \text{if } x \leq -1 \\ 0 & \text{elsewhere,} \end{cases}$$

where σ_0 is a damping constant. The damping function σ_y can be similarly defined using y variable.

Here we present a test result obtained with time step $\tau = 10^{-3}$, damping constant $\sigma_0 = 1$, discontinuous quadratic basis function on a triangular mesh with 2,748 vertices and 5,317 elements, and the simulation time $t \in (0, 1{,}500\,\tau)$ such that the wave front has propagated out of the simulation domain when $t = 1{,}500\,\tau$.

Some snapshots at various time steps are presented in Fig. 9.1, which shows that the PML performs very well since there is no wave reflected back at the interfaces between the PML layer and the free space.

9.1.2 The Multiscale Phenomena for Metamaterials

Here we solve a coupled model problem on a complex domain where a circle $(x - 0.5)^2 + y^2 = 0.5^2$ is located inside a rectangle $[-1.5, 1.5]^2$. The circle region is occupied by a Drude type metamaterial, which is governed by the non-dimensionalized 2-D transverse magnetic metamaterial modeling equations (4.59)–(4.64) with sources $g_x = g_y = f = 0$, i.e.,

$$\begin{aligned}
\frac{\partial H_x}{\partial t} &= -\frac{\partial E_z}{\partial y} - K_x, \\
\frac{\partial H_y}{\partial t} &= \frac{\partial E_z}{\partial x} - K_y, \\
\frac{\partial E_z}{\partial t} &= \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} - J_z, \\
\frac{\partial J_z}{\partial t} &= \omega_e^2 E_z - \Gamma_e J_z, \\
\frac{\partial K_x}{\partial t} &= \omega_m^2 H_x - \Gamma_m K_x, \\
\frac{\partial K_y}{\partial t} &= \omega_m^2 H_y - \Gamma_m K_y.
\end{aligned}$$

Outside of the circle but within the rectangle $[-1, 1]^2$ is filled by air, which is modeled by the 2-D transverse magnetic modeling equations:

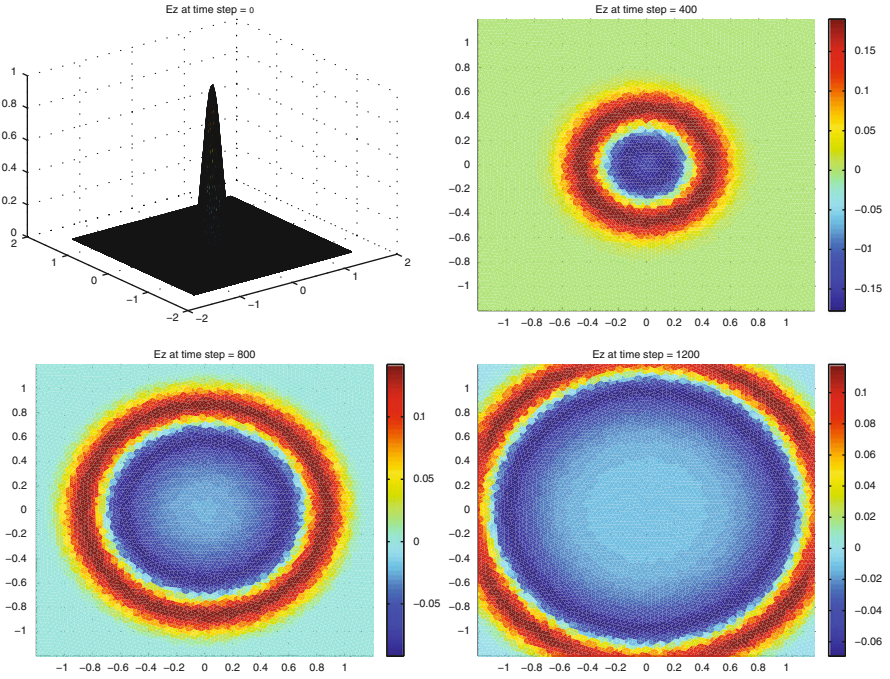

Fig. 9.1 Demonstration of PML effects. (*Top Left*) Surface plot of E_z at $t = 0$; contour plots of E_z at various time steps: $t = 400\tau$ (*Top Right*); $t = 800\tau$ (*Bottom Left*); $t = 1{,}200\,\tau$ (*Bottom Right*)

$$\frac{\partial H_x}{\partial t} = -\frac{\partial E_z}{\partial y} \tag{9.12}$$

$$\frac{\partial H_y}{\partial t} = \frac{\partial E_z}{\partial x} \tag{9.13}$$

$$\frac{\partial E_z}{\partial t} = \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y}. \tag{9.14}$$

The remaining area outside of $[-1, 1]^2$ is modeled by the PML equations (9.1)–(9.4).

This coupled model problem is solved by the nodal discontinuous Galerkin method. The initial source wave has the same form as (9.11) but centered at $(-0.5, 0)$. This problem has been solved using various parameters in [185]. After many numerical tests, it is found that the metamaterial model (4.59)–(4.64) has very different wave propagation phenomena, which depend on the relative size of those physical parameters.

Below we present an exemplary result solved with $\Gamma_e = \Gamma_m = 1, \tau = 10^{-3}$ on a triangular mesh (see Fig. 9.2) with 1,713 nodes and 3,304 elements. The basis function is second order. The problem is solved with a varying $\omega_e = \omega_m$. Numerical

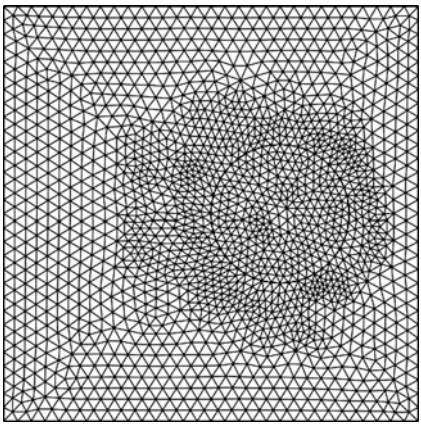

Fig. 9.2 The mesh and model setup for the coupled problem

results show that when $\omega_e < 1$, the wave can propagate through the metamaterial region without much damage of the initial electric field E_z. When ω_e becomes larger than 1, the wave gets damped as it moves into the metamaterial region. When ω_e is in the range of $[10, 20]$, the wave not only gets damped but also reflects from the metamaterial region. When ω_e is larger than 50, the wave propagates into the metamaterial region and damps very badly without much reflection. The exemplary results shown in Fig. 9.3 are obtained with $\tau = 10^{-3}$ running for 1,000 time steps, and $\omega_e = 0.2, 5, 20, 100$. Figure 9.3 demonstrates again that modeling wave propagation in metamaterials is quite challenging due to the inherited multiscale characteristics.

9.1.3 Demonstration of Backward Wave Propagation

In Sect. 1.1.1, we mentioned that since the refractive index of metamaterial is negative, the phase velocity is antiparallel to the energy flow direction, which fact leads to the so-called backward wave propagation pheonomenon in metamaterials.

To demonstrate this pheonomenon, Ziolkowski [311] designed some interesting examples to model electromagnetic wave propagation in metamaterials. Following [311], we consider the 2-D transverse magnetic model:

$$\epsilon_0 \frac{\partial E_y}{\partial t} = \frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x} - J_y, \tag{9.15}$$

$$\mu_0 \frac{\partial H_z}{\partial t} = -\frac{\partial E_y}{\partial x} - K_z, \tag{9.16}$$

$$\mu_0 \frac{\partial H_x}{\partial t} = \frac{\partial E_y}{\partial z} - K_x, \tag{9.17}$$

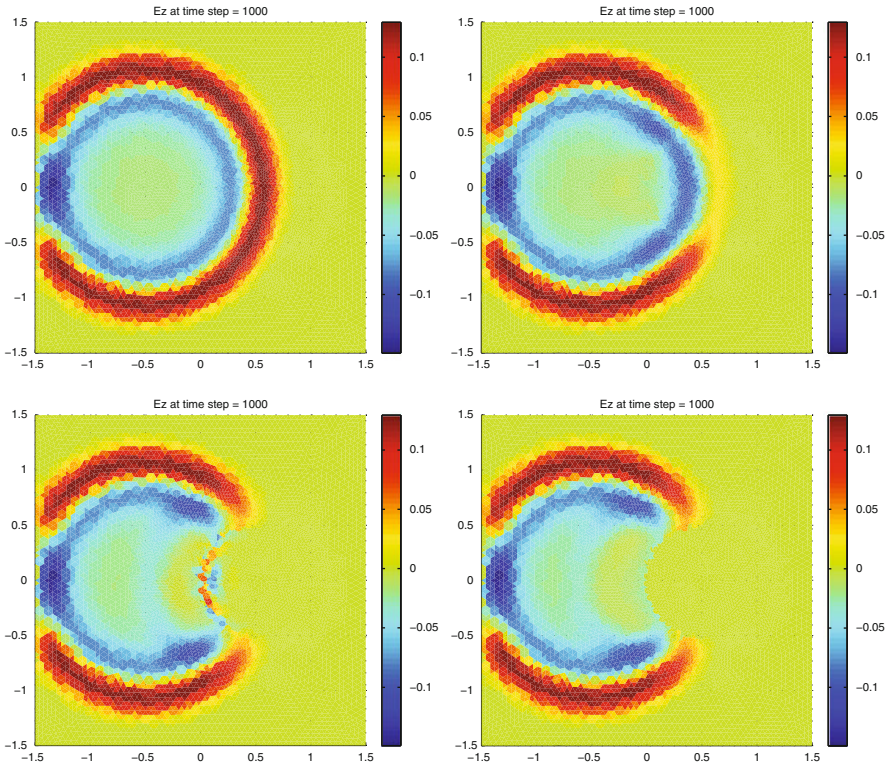

Fig. 9.3 The electric fields obtained with fixed $\Gamma_e = \Gamma_m = 1$, and varying $\omega_e = \omega_m$. (*Top Left*): $\omega_e = 0.2$; (*Top Right*): $\omega_e = 5$; (*Bottom Left*): $\omega_e = 20$; (*Bottom Right*): $\omega_e = 100$

$$\frac{1}{\epsilon_0 \omega_{pe}^2} \frac{\partial J_y}{\partial t} + \frac{\Gamma_e}{\epsilon_0 \omega_{pe}^2} J_y = E_y, \tag{9.18}$$

$$\frac{1}{\mu_0 \omega_{pm}^2} \frac{\partial K_z}{\partial t} + \frac{\Gamma_m}{\mu_0 \omega_{pm}^2} K_z = H_z, \tag{9.19}$$

$$\frac{1}{\mu_0 \omega_{pm}^2} \frac{\partial K_x}{\partial t} + \frac{\Gamma_m}{\mu_0 \omega_{pm}^2} K_x = H_x, \tag{9.20}$$

where H_z, H_x and E_y are the field components in the z, x and y directions, respectively. Note that this model is obtained using the Drude model introduced in Chap. 1 (cf. (1.12) and (1.13)).

First, we model a normal incidence wave beam interacting with a metamaterial slab with refractive index $n \approx -1$, which can be achieved by choosing

$$\Gamma_e = \Gamma_m = \Gamma = 10^8\, s^{-1}, \quad \omega_{pe} = \omega_{pm} = \omega_p = 2\pi\sqrt{2} f_0, \quad f_0 = 30\text{GHz}.$$

Note that in this case, the refractive index

$$n(\omega) = \sqrt{\frac{\epsilon(\omega)\mu(\omega)}{\epsilon_0\mu_0}} = 1 - \frac{\omega_p^2}{\omega(\omega - i\Gamma)} \approx -1,$$

where μ_0 and ϵ_0 are the vacuum permeability and permittivity, respectively.

Following Ziolkowski [311], the incident wave is chosen to be varied in space as $\exp(-x^2/\text{waist}^2)$ and in time as

$$f(t) = \begin{cases} 0 & \text{for } t < 0, \\ g_{on}(t)\sin(\omega_0 t) & \text{for } 0 < t < mT_p, \\ \sin(\omega_0 t) & \text{for } mT_p < t < (m+k)T_p, \\ g_{off}(t)\sin(\omega_0 t) & \text{for } (m+k)T_p < t < (2m+k)T_p, \\ 0 & \text{for } (2m+k)T_p < t, \end{cases} \tag{9.21}$$

where we denote $T_p = 1/f_0$, and

$$\begin{aligned} g_{on}(t) &= 10x_{on}^3(t) - 15x_{on}^4(t) + 6x_{on}^5(t), \quad x_{on}(t) = t/mT_p, \\ g_{off}(t) &= 1 - [10x_{off}^3(t) - 15x_{off}^4(t) + 6x_{on}^5(t), \\ x_{off}(t) &= [t - (m+k)T_p]/mT_p. \end{aligned}$$

In the first example, the simulation domain is chosen as 830×640 cells (z vs. x), where the cell thickness $dx = 10^{-4}$ m. The wave source is located at the center in the x-direction and 40 cells above the bottom of the simulation domain. A metamaterial slab is put at 200 cells above the beam source, and its thickness is 200 cells. The Bérenger PML of eight cells in thickness is used around the simulation domain. The remaining area is modelled by the Maxwell's equations in free space:

$$\begin{aligned} \epsilon_0 \frac{\partial E_y}{\partial t} &= \frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x}, \\ \mu_0 \frac{\partial H_z}{\partial t} &= -\frac{\partial E_y}{\partial x}, \\ \mu_0 \frac{\partial H_x}{\partial t} &= \frac{\partial E_y}{\partial z}, \end{aligned}$$

which can be obtained by choosing $J_y = K_z = K_x = 0$ in (9.15)–(9.20).

The electric field intensity (obtained with time step $dt = 0.1$ ps, the beam waist being 50 cells, and the incident wave (9.21) choosing $m = 2, k = 100$) is plotted in Fig. 9.4, where the left one is the field at 5,000 time steps, which clearly shows that the wave propagates backward inside the metamaterial slab. Another interesting property of metamaterial is that the electromagnetic wave propagates very slowly inside metamaterial. To see this clearly, we plot the electric field

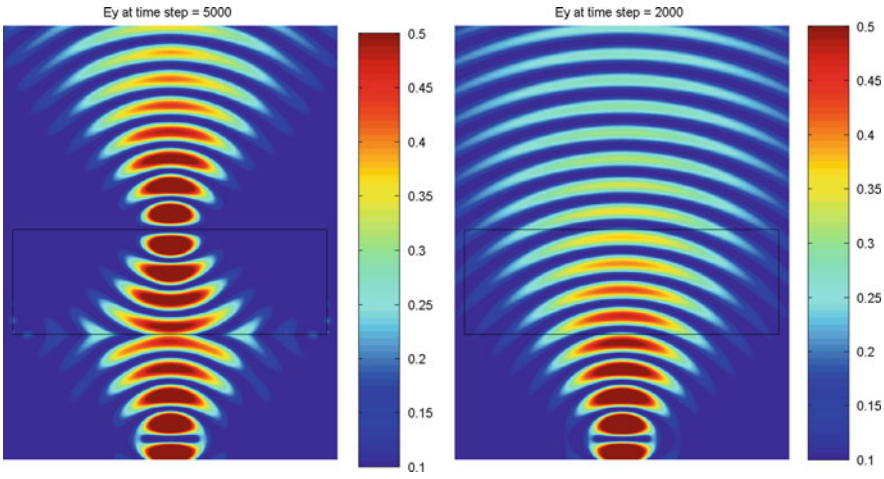

Fig. 9.4 The electric field intensity distribution: (*Left*) passing through a metamaterial slab with $n \approx -1$; (*Right*) passing through free space only (i.e., $n = 1$) (Reprinted from Li et al. [191]. Copyright (2008), with permission from Elsevier)

intensity obtained at 2,000 time steps for a wave propagating in free space (just replacing the metamaterial slab by vacuum) in Fig. 9.4 (Right). This figure shows that the wave reaches the boundary on the other side in only 2,000 time steps, while it takes about 5,000 time steps when a metamaterial slab is present.

Figure 9.4 (Left) also shows that the metamaterial slab has a nice refocusing property, which usually can be achieved by convex lenses. This property has prompted many researchers to work on the so-called perfect lens [234]. Encouraged by this phenomenon, we can further simulate a wave beam interacting with many metamaterial slabs of $n \approx -1$. An example of three slabs is given in [191], where the simulation domain is $1,500 \times 500$ cells (z vs. x), the bottom metamaterial slab is located 240 cells above the bottom side, the distance between each slab is 400 cells. Each slab is 460 cells in width and 200 cells in thickness. The source beam is located at the same place as the previous example. The obtained electric field intensity distribution at different times are presented in Fig. 9.5, which clearly show that the source beam can be transmitted further away via multiple metamaterial slabs. This phenomenon opens the potential applications in nano-waveguides.

The last example modified from [157] is used to demonstrate the backward wave propagation phenomenon and the Snell's law using a triangular metamaterial slab. The physical domain is chosen to be $[0, 0.06] \times [0, 0.064]$ m. The incident source wave is located at $x = 0.004$ m and imposed as a scalar component. A triangle metamaterial slab is determined by vertices $(0.014, 0.02)$, $(0.014, 0.062)$ and $(0.044, 0.062)$. Outside this slab is vacuum. For this example, a hybrid mesh is used, where a triangular mesh is used for the metamaterial slab and its neighboring elements, and a rectangular mesh is used in the vacuum region and PML region. A leap-frog mixed finite element method is used for this example, where the lowest-

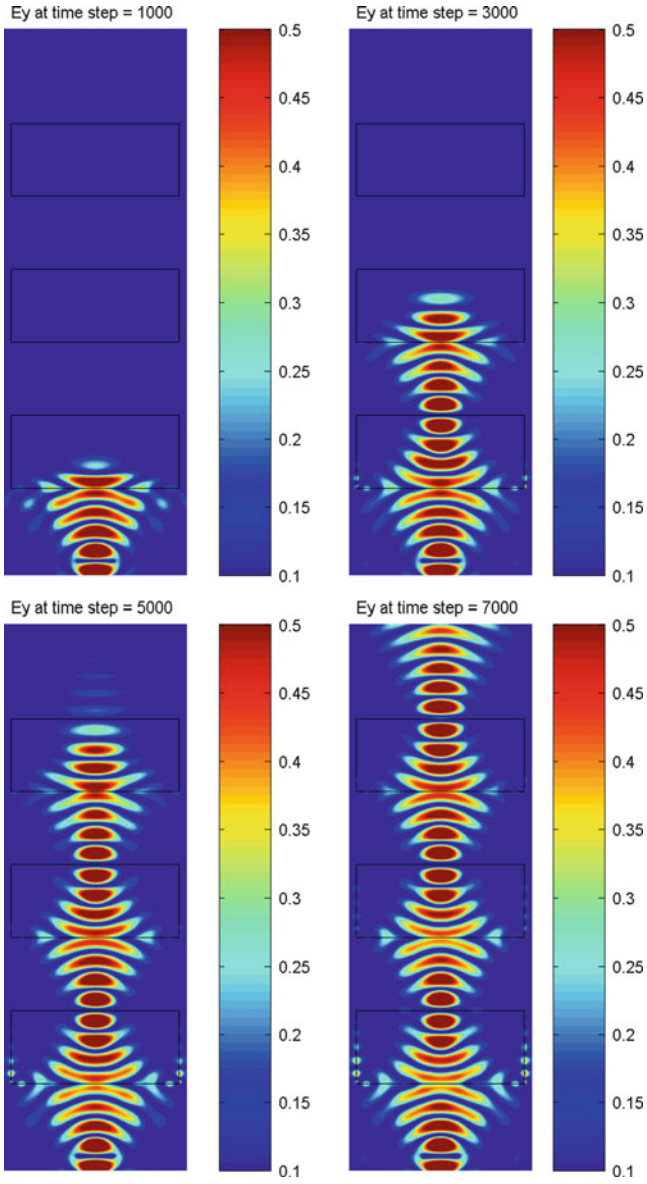

Fig. 9.5 Electric field intensity distribution interacting with three metamaterial slabs of $n \approx -1$ (Reprinted from Li et al. [191]. Copyright (2008), with permission from Elsevier)

order triangular edge element and rectangular edge element are used. Details about the algorithm and implementation can be found in the original paper [157].

An exemplary mesh is shown in Fig. 9.6, which is quite coarse for illustration purpose. The results presented in Fig. 9.7 uses a mesh by uniformly refining Fig. 9.6

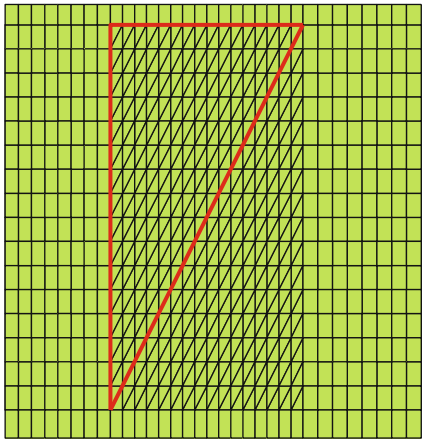

Fig. 9.6 An exemplary mixed mesh used for the triangular metamaterial slab

four times, i.e., the real mesh has 131,072 and 81,920 triangular and rectangular elements, respectively. Hence the total number of degrees of freedom for **E** is 361,248. In this case, the time step $\tau = 10^{-13}$. The calculated E_y components at various time steps are plotted in Fig. 9.7, which shows clearly that the wave propagates backward inside the metamaterial slab. After the wave exits the metamaterial region, the wave bends according to the Snell's law (1.1).

9.2 Metamaterial Electromagnetic Cloak

In recent years, inspired by the pioneering work of Pendry et al. [237] and Leonhardt [180] in 2006, there is a growing interest in the study of using metamaterials to construct invisibility cloaks of different shapes. More details and references on cloaking can be found in recent reviews [73, 132, 135]. One of the major avenues towards electromagnetic and acoustic cloaking is the so-called transformation optics [180, 237], which uses the coordinate transformation to design the material parameters to steer the light around the cloaked regions. In this section, we present some cloaking results obtained via Maxwell's equations, although cloaking can be achieved through solving other types of equations (e.g.,[9, 166, 167])

9.2.1 Form Invariant Property for Maxwell's Equations

Modeling of electromagnetic phenomena at a fixed frequency ω is governed by the full Maxwell's equations (assuming a time harmonic variation of $\exp(j\omega t)$):

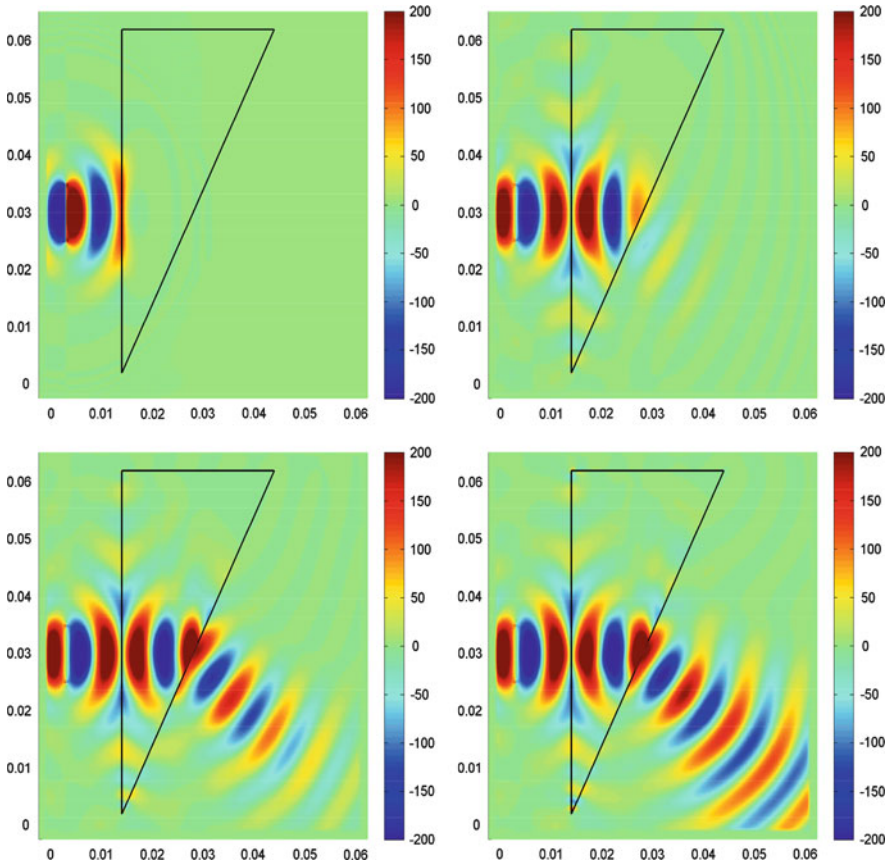

Fig. 9.7 Example 4. Electric fields E_y at various time steps: (*Top Left*) 800 steps; (*Top Right*) 2,000 steps; (*Bottom Left*) 3,000 steps; (*Bottom Right*) 4,000 steps

$$\nabla \times \mathbf{E} + j\omega\mu\mathbf{H} = 0, \quad \nabla \times \mathbf{H} - j\omega\epsilon\mathbf{E} = 0, \tag{9.22}$$

where $\mathbf{E}(\mathbf{x})$ and $\mathbf{H}(\mathbf{x})$ are the electric and magnetic fields in the frequency domain, and ϵ and μ are the permittivity and permeability of the material.

A very important property for Maxwell's equations is that Maxwell's equations are form invariant under coordinate transformations (cf. [214]). More specifically, we have

Theorem 9.1. *Under a coordinate transformation $\boldsymbol{x}' = \boldsymbol{x}'(\boldsymbol{x})$, the equations (9.22) keep the same form in the transformed coordinate system:*

$$\nabla' \times \boldsymbol{E}' + j\omega\mu'\boldsymbol{H}' = 0, \quad \nabla' \times \boldsymbol{H}' - j\omega\epsilon'\boldsymbol{E}' = 0, \tag{9.23}$$

where all new variables are given by

$$\boldsymbol{E}'(\boldsymbol{x}') = A^{-T}\boldsymbol{E}(\boldsymbol{x}),\ \boldsymbol{H}'(\boldsymbol{x}') = A^{-T}\boldsymbol{H}(\boldsymbol{x}),\ A = (a_{ij}),\ a_{ij} = \frac{\partial x_i'}{\partial x_j}, \tag{9.24}$$

and

$$\mu'(\boldsymbol{x}') = A\mu(\boldsymbol{x})A^T/det(A), \quad \epsilon'(\boldsymbol{x}') = A\epsilon(\boldsymbol{x})A^T/det(A). \tag{9.25}$$

Proof. From (9.24), (9.25) and (9.22), we have

$$j\omega\mu'\mathbf{H}' = j\omega A\mu\mathbf{H}/det(A) = -A\nabla\times\mathbf{E}/det(A).$$

Hence to prove the first identity of (9.23), we just need to show that

$$A\nabla\times\mathbf{E} = det(A)\cdot\nabla'\times\mathbf{E}'. \tag{9.26}$$

Before we prove (9.26), let us recall the 3-D Levi-Civita symbol ϵ_{ijk}, which is 1 if (i, j, k) is an even permutation of $(1, 2, 3)$, -1 if it is an odd permutation, and 0 if any index is repeated. Hence by using the Einstein notation (i.e., omitting the summation symbols), we have

$$det(A) = \epsilon_{ijk}\frac{\partial x_1'}{\partial x_i}\frac{\partial x_2'}{\partial x_j}\frac{\partial x_3'}{\partial x_k}, \tag{9.27}$$

and the ith component of $\nabla\times\mathbf{E}$:

$$(\nabla\times\mathbf{E})_i = \epsilon_{ijk}\frac{\partial E_k}{\partial x_j},$$

from which and $\mathbf{E} = A^T\mathbf{E}'$, we obtain

$$\begin{aligned}(A\nabla\times\mathbf{E})_i &= \frac{\partial x_i'}{\partial x_m}\epsilon_{mjk}\frac{\partial E_k}{\partial x_j} = \frac{\partial x_i'}{\partial x_m}\epsilon_{mjk}\frac{\partial}{\partial x_j}(\frac{\partial x_l'}{\partial x_k}E_l')\\ &= \frac{\partial x_i'}{\partial x_m}\epsilon_{mjk}(\frac{\partial^2 x_l'}{\partial x_j\partial x_k}E_l' + \frac{\partial x_l'}{\partial x_k}\frac{\partial E_l'}{\partial x_j})\\ &= \frac{\partial x_i'}{\partial x_m}\epsilon_{mjk}\frac{\partial x_l'}{\partial x_k}\frac{\partial E_l'}{\partial x_j}\\ &= \frac{\partial x_i'}{\partial x_m}\epsilon_{mjk}\frac{\partial x_l'}{\partial x_k}\frac{\partial E_l'}{\partial x_p'}\frac{\partial x_p'}{\partial x_j},\end{aligned} \tag{9.28}$$

where in the above we used the fact that the first term is zero by swapping the indices j and k.

On the other hand, we have

$$det(A) \cdot (\nabla' \times \mathbf{E}')_i = det(A) \cdot \epsilon_{ipl} \frac{\partial E'_l}{\partial x'_p}. \tag{9.29}$$

Comparing (9.28) with (9.29), we can see that proof of (9.26) boils down to proof of the following

$$\frac{\partial x'_i}{\partial x_m} \epsilon_{mjk} \frac{\partial x'_l}{\partial x_k} \frac{\partial x'_p}{\partial x_j} = det(A) \cdot \epsilon_{ipl},$$

which is true by checking different i, p, l. For example, $i = 1, p = 2, l = 3$ is just (9.27). □

9.2.2 Design of Cylindrical and Square Cloaks

In this subsection, we present detailed derivation of the metamaterial's permittivity and permeability which lead to cloaking pheonomena. The contents of this subsection are mainly based on [188].

9.2.2.1 Cylindrical Cloak

Following [237], to hide an object inside the cylindrical region $r \leq R_1$, a special electromagnetic metamaterial can be designed in the cloaking region $R_1 < r < R_2$ through the so-called transformation optics technique. The idea is to take all fields in the region $r < R_2$ and compress them into the region $R_1 < r < R_2$. This can be accomplished by the following simple coordinate transformation:

$$r'(r, \theta) = \frac{R_2 - R_1}{R_2} r + R_1, \quad 0 \leq r \leq R_2, \tag{9.30}$$

$$\theta'(r, \theta) = \theta. \tag{9.31}$$

To carry out a cloaking simulation in Cartesian coordinates, we have to transform the material parameters given in polar coordinates to Cartesian coordinates. It is known that a point (x_1, x_2) in the Cartesian coordinate system corresponds to a point (r, θ) in polar coordinate system through the relations:

$$r = \sqrt{x_1^2 + x_2^2}, \quad \theta = \tan^{-1} \frac{x_2}{x_1}, \tag{9.32}$$

and

$$x_1 = r \cos \theta, \quad x_2 = r \sin \theta, \tag{9.33}$$

which leads to

$$\frac{\partial r}{\partial x_1} = \frac{x_1}{r} = \cos\theta, \quad \frac{\partial r}{\partial x_2} = \frac{x_2}{r} = \sin\theta, \tag{9.34}$$

$$\frac{\partial \theta}{\partial x_1} = -\frac{x_2}{r^2} = -\frac{\sin\theta}{r}, \quad \frac{\partial \theta}{\partial x_2} = \frac{x_1}{r^2} = \frac{\cos\theta}{r}. \tag{9.35}$$

By Theorem 9.1, the electromagnetic permittivity and permeability in the transformed space are given by (9.25), which needs the information of $\frac{\partial x_i'}{\partial x_j}$. For the transformation (9.30) and (9.31), by chain rule we can obtain

$$\begin{aligned}(i)\ \frac{\partial x_1'}{\partial x_1} &= \frac{\partial x_1'}{\partial r'}\frac{\partial r'}{\partial r}\frac{\partial r}{\partial x_1} + \frac{\partial x_1'}{\partial \theta'}\frac{\partial \theta'}{\partial \theta}\frac{\partial \theta}{\partial x_1} \\ &= \cos\theta \cdot \frac{R_2 - R_1}{R_2} \cdot \cos\theta - r' \sin\theta \cdot (-\frac{\sin\theta}{r}) \\ &= \frac{R_2 - R_1}{R_2} + \frac{R_1}{r}\sin^2\theta,\end{aligned}$$

$$\begin{aligned}(ii)\ \frac{\partial x_1'}{\partial x_2} &= \frac{\partial x_1'}{\partial r'}\frac{\partial r'}{\partial r}\frac{\partial r}{\partial x_2} + \frac{\partial x_1'}{\partial \theta'}\frac{\partial \theta'}{\partial \theta}\frac{\partial \theta}{\partial x_2} \\ &= \cos\theta \cdot \frac{R_2 - R_1}{R_2} \cdot \sin\theta - r' \sin\theta \cdot (\frac{\cos\theta}{r}) \\ &= -\frac{R_1}{r}\sin\theta\cos\theta,\end{aligned}$$

$$\begin{aligned}(iii)\ \frac{\partial x_2'}{\partial x_1} &= \frac{\partial x_2'}{\partial r'}\frac{\partial r'}{\partial r}\frac{\partial r}{\partial x_1} + \frac{\partial x_2'}{\partial \theta'}\frac{\partial \theta'}{\partial \theta}\frac{\partial \theta}{\partial x_1} \\ &= \sin\theta \cdot \frac{R_2 - R_1}{R_2} \cdot \cos\theta + r' \cos\theta \cdot (-\frac{\sin\theta}{r}) \\ &= -\frac{R_1}{r}\sin\theta\cos\theta,\end{aligned}$$

and

$$\begin{aligned}(iv)\ \frac{\partial x_2'}{\partial x_2} &= \frac{\partial x_2'}{\partial r'}\frac{\partial r'}{\partial r}\frac{\partial r}{\partial x_2} + \frac{\partial x_2'}{\partial \theta'}\frac{\partial \theta'}{\partial \theta}\frac{\partial \theta}{\partial x_2} \\ &= \sin\theta \cdot \frac{R_2 - R_1}{R_2} \cdot \sin\theta + r' \cos\theta \cdot (\frac{\cos\theta}{r}) \\ &= \frac{R_2 - R_1}{R_2} + \frac{R_1}{r}\cos^2\theta.\end{aligned}$$

Hence by Theorem 9.1, the transformation matrix A can be obtained as

$$A = \begin{pmatrix} \frac{R_2-R_1}{R_2} + \frac{R_1}{r}\sin^2\theta & -\frac{R_1}{r}\sin\theta\cos\theta \\ symmetric & \frac{R_2-R_1}{R_2} + \frac{R_1}{r}\cos^2\theta \end{pmatrix}, \tag{9.36}$$

whose determinant is

$$\det(A) = \frac{R_2-R_1}{R_2}\left(\frac{R_2-R_1}{R_2} + \frac{R_1}{r}\right) = \left(\frac{R_2-R_1}{R_2}\right)^2 \cdot \frac{r'}{r'-R_1}. \tag{9.37}$$

Substituting (9.36) and (9.37) into (9.25) with relative permittivity $\epsilon_r = 1$ in the original space, we obtain the relative permittivity in the transformed space

$$\epsilon' = \begin{pmatrix} \epsilon'_{xx} & \epsilon'_{xy} \\ \epsilon'_{yx} & \epsilon'_{yy} \end{pmatrix} = AA^T/\det(A)$$

$$= \begin{pmatrix} \left(\frac{R_2-R_1}{R_2}\right)^2 + \frac{R_1}{r}\left(2\frac{R_2-R_1}{R_2} + \frac{R_1}{r}\right)\sin^2\theta & -\frac{R_1}{r}\left(2\frac{R_2-R_1}{R_2} + \frac{R_1}{r}\right)\sin\theta\cos\theta \\ symmetric & \left(\frac{R_2-R_1}{R_2}\right)^2 + \frac{R_1}{r}\left(2\frac{R_2-R_1}{R_2} + \frac{R_1}{r}\right)\cos^2\theta \end{pmatrix}/\det(A),$$

i.e., the material parameters in Cartesian coordinates become as follows:

$$\epsilon'_{xx} = \left[\left(\frac{R_2-R_1}{R_2}\right)^2 + \frac{R_1}{r}\left(2\frac{R_2-R_1}{R_2} + \frac{R_1}{r}\right)\sin^2\theta\right]/\det(A),$$

$$\epsilon'_{xy} = \epsilon'_{yx} = \left[-\frac{R_1}{r}\left(2\frac{R_2-R_1}{R_2} + \frac{R_1}{r}\right)\sin\theta\cos\theta\right]/\det(A),$$

$$\epsilon'_{yy} = \left[\left(\frac{R_2-R_1}{R_2}\right)^2 + \frac{R_1}{r}\left(2\frac{R_2-R_1}{R_2} + \frac{R_1}{r}\right)\cos^2\theta\right]/\det(A),$$

and $\epsilon'_z = 1/\det(A)$. The permeability μ' has the same form as permittivity ϵ'.

9.2.2.2 Square Cloak

The transformation optics idea can be used for designing a square-shaped cloak. In this case, the fields inside a square with width $2S_2$ are compressed into a square annulus with inner square width $2S_1$ and outer square width $2S_2$. This task can be accomplished through four mappings.

The right triangle in the original space is mapped into the right-subdomain in the transformed space (see Fig. 9.8) by the coordinate transformation [244]

$$x'_1(x_1, x_2) = x_1\frac{S_2 - S_1}{S_2} + S_1, \tag{9.38}$$

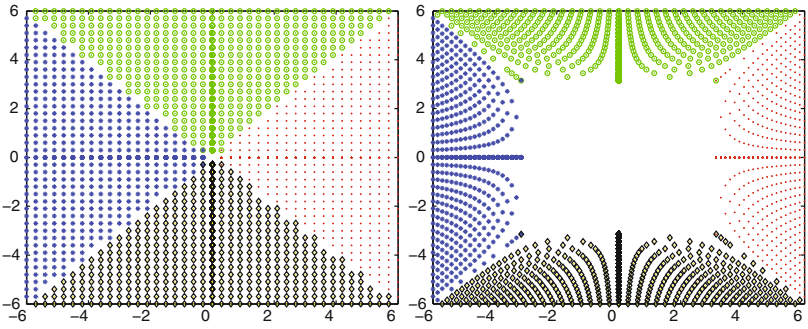

Fig. 9.8 (*Left*) The original *square* formed by four *triangles*; (*Right*) The transformed *square* annulus

$$x_2'(x_1, x_2) = x_2\left(\frac{S_2 - S_1}{S_2} + \frac{S_1}{x_1}\right). \tag{9.39}$$

It is easy to prove that the transformation matrix in this case is

$$A_r = \begin{pmatrix} \frac{S_2-S_1}{S_2} & 0 \\ -\frac{x_2 S_1}{x_1^2} & \frac{S_2-S_1}{S_2} + \frac{S_1}{x_1} \end{pmatrix}, \tag{9.40}$$

which has determinant

$$\det(A_r) = \frac{S_2 - S_1}{S_2}\left(\frac{S_2 - S_1}{S_2} + \frac{S_1}{x_1}\right). \tag{9.41}$$

Mapping the unit permittivity tensor $\epsilon = I$ by (9.25), we obtain

$$\epsilon_r' = A_r A_r^T / \det(A_r) = \begin{pmatrix} (\frac{S_2-S_1}{S_2})^2 & -\frac{x_2 S_1}{x_1^2} \cdot \frac{S_2-S_1}{S_2} \\ \textit{symmetric} & (\frac{x_2 S_1}{x_1^2})^2 + (\frac{S_2-S_1}{S_2} + \frac{S_1}{x_1})^2 \end{pmatrix} / \det(A_r). \tag{9.42}$$

Corresponding formulas for the upper, left and bottom sub-domains of the cloak can be similarly obtained by applying rotation matrix $R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ to the right sub-domain with rotation angles $\theta = \pi/2, \pi$ and $3\pi/2$, respectively.

More specifically, for the upper subdomain, we have

$$\epsilon_u' = R(\frac{\pi}{2})\epsilon_r' R(\frac{\pi}{2})^T = \begin{pmatrix} (\frac{S_2-S_1}{S_2} + \frac{S_1}{x_1})^2 + (\frac{x_2 S_1}{x_1^2})^2 & \frac{S_2-S_1}{S_2} \cdot \frac{x_2 S_1}{x_1^2} \\ \textit{symmetric} & (\frac{S_2-S_1}{S_2})^2 \end{pmatrix} / \det(A_r).$$

For the left subdomain, we have

$$\epsilon_l' = R(\pi)\epsilon_r' R(\pi)^T = \begin{pmatrix} (\frac{S_2-S_1}{S_2})^2 & -\frac{x_2 S_1}{x_1^2}\cdot\frac{S_2-S_1}{S_2} \\ symmetric & (\frac{x_2 S_1}{x_1^2})^2 + (\frac{S_2-S_1}{S_2} + \frac{S_1}{x_1})^2 \end{pmatrix} / \det(A_r).$$

For the bottom subdomain, we have

$$\epsilon_b' = R(\frac{3\pi}{2})\epsilon_r' R(\frac{3\pi}{2})^T = \begin{pmatrix} (\frac{S_2-S_1}{S_2} + \frac{S_1}{x_1})^2 + (\frac{x_2 S_1}{x_1^2})^2 & \frac{S_2-S_1}{S_2}\cdot\frac{x_2 S_1}{x_1^2} \\ symmetric & (\frac{S_2-S_1}{S_2})^2 \end{pmatrix} / \det(A_r).$$

9.2.3 Cloak Simulation in the Frequency Domain

Before we move to the time domain cloak simulation in the next section, here we present some 2-D cloaking simulations in the frequency domain. Without loss of generality, we consider the 2-D transverse magnetic model. Reducing (9.22) with $\epsilon = \epsilon_0\epsilon_r$ and $\mu = \mu_0\mu_r$ into just one equation involving the scalar variable E_z, we obtain

$$\nabla \times (\mu_r^{-1}\nabla \times E_z) - k_0^2 \epsilon_r E_z = 0, \tag{9.43}$$

where μ_r and ϵ_r are the relative permeability and permittivity, and k_0 denotes the wave number of free space $k_0 = \omega\sqrt{\epsilon_0\mu_0} = \frac{\omega}{C_v}$. As before, $C_v = 1/\sqrt{\epsilon_0\mu_0}$ denotes the light speed in free space.

The simulations given below are performed by COMSOL Multiphysics package, where quadratic triangular elements and the direct solver SPOOLES are used.

9.2.3.1 Cylindrical Cloak

For this test, the cylinder cloak shell is chosen to have $R_1 = 0.15$ m and $R_2 = 0.3$ m, and located inside the square $[-1.0, 1.0]^2$. A PML with 0.5 m thickness is imposed on both ends of this square in the x-direction, and the periodic boundary is imposed on the top and bottom boundaries. The incident plane waves of 1–4 GHz are excited on the interface $x = -1$.

First, a coarse mesh with 14,624 elements and 7,417 nodes is used for the simulation. In this case, the total number of DOFs is 25,584. The obtained electric field distributions for incident waves of several frequencies are presented in Fig. 9.9, which show that the 1–3 GHz plane wave patterns are restored quite well after the waves propagate out of the cloaked area. Hence this structure demonstrates good cloaking effect for 1–3 GHz plane waves. However, the cloaking phenomenon is not clear for the 4 GHz incident wave.

Then the same problem is solved again with a finer mesh obtained by uniformly refining the previous mesh twice, in which case the mesh has 58,496 elements, 29,457 nodes and 101,728 DOFs. With this finer mesh, the cloaking effect can be seen quite clearly for all 1–4 GHz incident waves as demonstrated in Fig. 9.10.

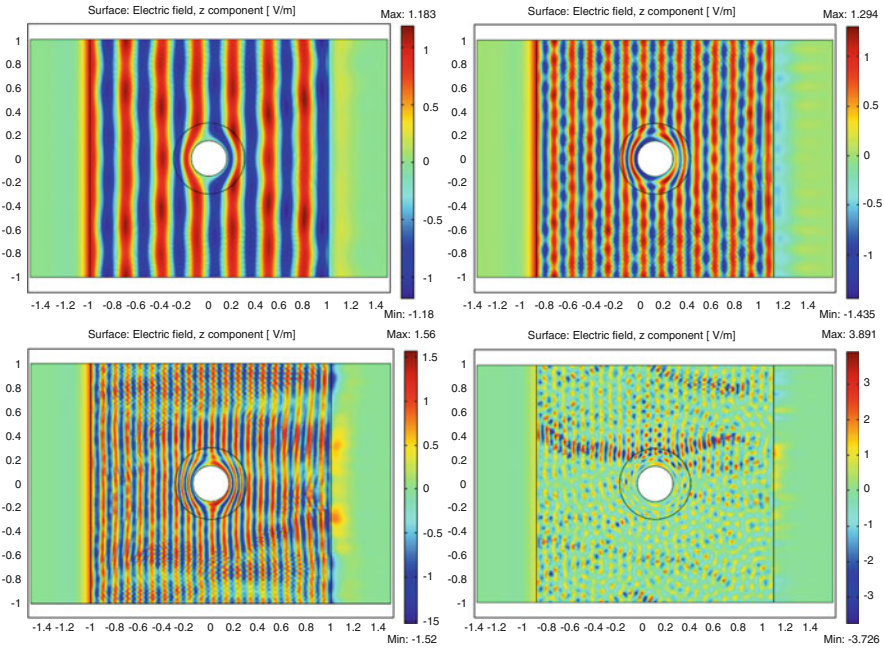

Fig. 9.9 The real part of the electric-field phasor obtained for the cylinder cloak with incident waves of different frequencies: (*Top Left*) 1 GHz; (*Top Right*) 2 GHz; (*Bottom Left*) 3 GHz; (*Bottom Right*) 4 GHz

9.2.3.2 Square Cloak

The square cloak has the same geometry as the cylindrical case, except that the circular shell is replaced by a rectangular shell. This problem is solved first using a mesh with 7,200 elements, 3,720 nodes and 11,352 DOFs, and the cloaking effect can be seen only for waves with 1 and 2 GHz frequencies. Then the problem is solved again using a finer mesh refined uniformly from the previous one, and the cloaking effect can be seen for waves with 3 and 4 GHz frequencies. In Fig. 9.11, the obtained electric field distributions for 3 and 4 GHz waves are presented for both the coarse and fine meshes. This example shows that modeling wave propagation in metamaterials is quite challenging, since the right physical phenomena can be observed only when the mesh is fine enough.

9.3 Time Domain Cloak Simulation

Compared to many frequency domain cloak simulations, not much attention has been paid to the time-domain modeling of cloaks. Since 2008, some papers have

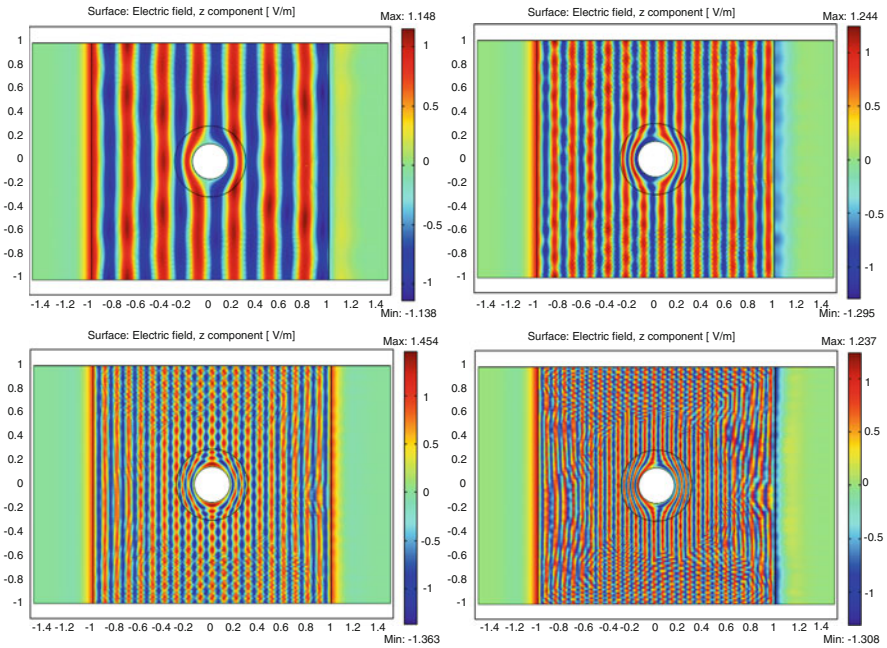

Fig. 9.10 The real part of the electric-field phasor obtained for the cylinder cloak with a fine mesh and various incident waves: (*Top Left*) 1 GHz; (*Top Right*) 2 GHz; (*Bottom Left*) 3 GHz; (*Bottom Right*) 4 GHz

been published on time-domain simulation of 2-D cloaking structures (see [197, 303, 304] and references therein). The recently designed broadband cloaks by Liu et al. in 2009 [206] make the time-domain simulation more appealing and necessary. Inspired by the work of [304], Li et al. [194] developed the first time-domain finite element (FETD) scheme for cloak simulation. This section is mainly based on [194].

9.3.1 The Governing Equations

Following [304], the time-domain cloak modeling is based on equations:

$$\frac{\partial \boldsymbol{B}}{\partial t} = -\nabla \times \boldsymbol{E}, \tag{9.44}$$

$$\frac{\partial \boldsymbol{D}}{\partial t} = \nabla \times \boldsymbol{H}, \tag{9.45}$$

and the constitutive relations

$$\boldsymbol{D} = \epsilon \boldsymbol{E}, \tag{9.46}$$

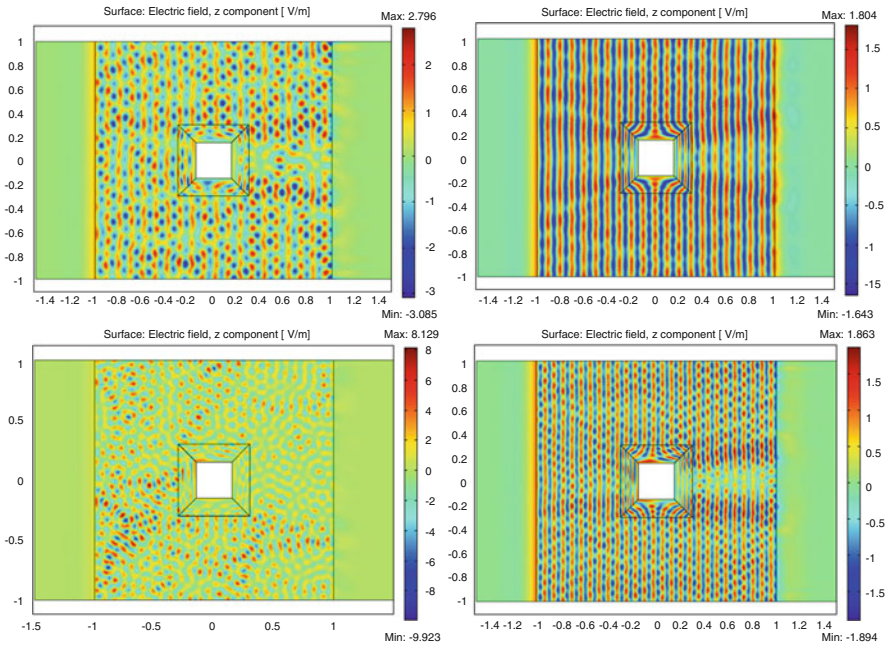

Fig. 9.11 The real part of the electric-field phasor obtained for the *square* cloak under two meshes: (*Top Row*) 3 GHz wave; (*Bottom Row*) 4 GHz wave

$$\boldsymbol{B} = \mu \boldsymbol{H}, \tag{9.47}$$

where as usual $\boldsymbol{E}$ and $\boldsymbol{H}$ are the electric and magnetic fields respectively, $\boldsymbol{D}$ and $\boldsymbol{B}$ are the electric displacement and magnetic induction respectively, ε and μ are cloak permittivity and permeability, respectively. For the cylindrical cloak, the ideal material parameters in the polar coordinate system were first given by Pendry et al. [237]:

$$\epsilon_r = \mu_r = \frac{r - R_1}{r}, \quad \epsilon_\phi = \mu_\phi = \frac{r}{r - R_1}, \quad \epsilon_z = \mu_z = \left(\frac{R_2}{R_2 - R_1}\right)^2 \frac{r - R_1}{r}, \tag{9.48}$$

where R_1 and R_2 are the inner and outer radius of the cloak. From (9.48), it can be seen that the cloaking metamaterial's permittivity and permeability are nonhomogeneous and highly anisotropic.

Following [304], here we consider the 2-D case with $\mathbf{E}$ being a vector, and $\mathbf{H}$ being a scalar, i.e., we can write $\mathbf{E} = (E_x, E_y)'$ and $H = H_z$, where the subindex x, y or z denotes the component in each direction. To carry out the simulation in Cartesian coordinates, we have to transform the material parameters (9.48) into Cartesian coordinates. It is easy to see that

$$\begin{bmatrix} \epsilon_{xx} & \epsilon_{xy} \\ \epsilon_{yx} & \epsilon_{yy} \end{bmatrix} = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix} \cdot \begin{bmatrix} \epsilon_r & 0 \\ 0 & \epsilon_\phi \end{bmatrix} \cdot \begin{bmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{bmatrix}$$
$$= \begin{bmatrix} \epsilon_r \cos^2\phi + \epsilon_\phi \sin^2\phi & (\epsilon_r - \epsilon_\phi)\sin\phi\cos\phi \\ (\epsilon_r - \epsilon_\phi)\sin\phi\cos\phi & \epsilon_r \sin^2\phi + \epsilon_\phi \cos^2\phi \end{bmatrix},$$

which, along with $\epsilon_0 \begin{bmatrix} E_x \\ E_y \end{bmatrix} = \begin{bmatrix} \epsilon_{xx} & \epsilon_{xy} \\ \epsilon_{yx} & \epsilon_{yy} \end{bmatrix}^{-1} \begin{bmatrix} D_x \\ D_y \end{bmatrix}$, yields

$$\epsilon_0\epsilon_r\epsilon_\phi \boldsymbol{E} = \begin{bmatrix} \epsilon_r \sin^2\phi + \epsilon_\phi \cos^2\phi & (\epsilon_\phi - \epsilon_r)\sin\phi\cos\phi \\ (\epsilon_\phi - \epsilon_r)\sin\phi\cos\phi & \epsilon_r \cos^2\phi + \epsilon_\phi \sin^2\phi \end{bmatrix} \boldsymbol{D}. \tag{9.49}$$

To obtain the cloak phenomenon, the material parameters have to be constructed from dispersive medium models. For simplicity, we consider the Drude model for the permittivity:

$$\varepsilon_r(\omega) = 1 - \frac{\omega_p^2}{\omega^2 - j\omega\gamma}, \tag{9.50}$$

where $\gamma \geq 0$ and $\omega_p > 0$ are the collision and plasma frequencies, respectively. Substituting (9.50) into (9.49) and using the following rules

$$j\omega \to \frac{\partial}{\partial t}, \quad \omega^2 \to -\frac{\partial^2}{\partial t^2}, \tag{9.51}$$

we have (cf. [304]):

$$\varepsilon_0\varepsilon_\phi \left(\frac{\partial^2}{\partial t^2} + \gamma\frac{\partial}{\partial t} + w_p^2\right)\boldsymbol{E}$$
$$= \left(\frac{\partial^2}{\partial t^2} + \gamma\frac{\partial}{\partial t} + w_p^2\right) M_A \boldsymbol{D} + \varepsilon_\phi \left(\frac{\partial^2}{\partial t^2} + \gamma\frac{\partial}{\partial t}\right) M_B \boldsymbol{D}, \tag{9.52}$$

where the vector $\boldsymbol{D} = (D_x, D_y)'$ and

$$M_A = \begin{bmatrix} \sin^2\phi & -\sin\phi\cos\phi \\ -\sin\phi\cos\phi & \cos^2\phi \end{bmatrix}, \quad M_B = \begin{bmatrix} \cos^2\phi & \sin\phi\cos\phi \\ \sin\phi\cos\phi & \sin^2\phi \end{bmatrix}.$$

Similarly, the permeability is described by the Drude model [304]:

$$\mu_z(\omega) = A\left(1 - \frac{\omega_{pm}^2}{\omega^2 - j\omega\gamma_m}\right), \tag{9.53}$$

where $A = \frac{R_2}{R_2 - R_1}$, and $\omega_{pm} > 0$ and $\gamma_m \geq 0$ are the magnetic plasma and collision frequencies, respectively. Substituting (9.53) into (9.47), we obtain

$$B_z = \mu_o \mu_z H_z = \mu_0 A \left(1 - \frac{\omega_{pm}^2}{\omega^2 - j\omega\gamma_m}\right) H_z.$$

Then using rules (9.51), we have

$$\left(\frac{\partial^2}{\partial t^2} + \gamma_m \frac{\partial}{\partial t}\right) B_z = \mu_0 A \left(\frac{\partial^2}{\partial t^2} + \gamma_m \frac{\partial}{\partial t} + \omega_{pm}^2\right) H_z. \tag{9.54}$$

To carry out the cloak simulation, we use Bérenger's PML [34] to reduce the infinite domain to a bounded one by absorbing those waves leaving the computational domain without introducing reflections. The two dimensional Bérenger PML governing equations can be written as:

$$\varepsilon_0 \frac{\partial E_x}{\partial t} + \sigma_y E_x = \frac{\partial \left(H_{zx} + H_{zy}\right)}{\partial y}, \tag{9.55}$$

$$\varepsilon_0 \frac{\partial E_y}{\partial t} + \sigma_x E_y = -\frac{\partial \left(H_{zx} + H_{zy}\right)}{\partial x}, \tag{9.56}$$

$$\mu_0 \frac{\partial H_{zx}}{\partial t} + \sigma_{mx} H_{zx} = -\frac{\partial E_y}{\partial x}, \tag{9.57}$$

$$\mu_0 \frac{\partial H_{zy}}{\partial t} + \sigma_{my} H_{zy} = \frac{\partial E_x}{\partial y}, \tag{9.58}$$

where the parameters $\sigma_i, \sigma_{mi}, i = x, y$, are the homogeneous electric and magnetic conductivities in the x and y directions, respectively.

For implementation purpose, (9.55) and (9.56) is written in the vector form:

$$\varepsilon_0 \frac{\partial \boldsymbol{E}}{\partial t} + \begin{pmatrix} \sigma_y & 0 \\ 0 & \sigma_x \end{pmatrix} \boldsymbol{E} = \nabla \times H, \tag{9.59}$$

where the 2-D vector curl operator $\nabla \times H = \begin{pmatrix} \frac{\partial H}{\partial y} \\ -\frac{\partial H}{\partial x} \end{pmatrix}$ for $H = H_{zx} + H_{zy}$.

9.3.2 An Explicit Finite Element Scheme

To design the scheme, we partition Ω by a family of regular meshes T_h with maximum mesh size h. To accommodate the problem easily, a hybrid mesh is used: triangles in both cloaking and free space regions; rectangles in the PML region, cf. Fig. 9.12b below. The basis functions used are the lowest-order Raviart-Thomas-Nédélec's mixed spaces $\boldsymbol{U}_h$ and $\boldsymbol{V}_h$: For a rectangular mesh T_h, we choose

$$\boldsymbol{U}_h = \{\psi_h \in L^2(\Omega):\ \ \psi_h|_K \in Q_{0,0},\ \forall\ K \in T_h\},$$
$$\boldsymbol{V}_h = \{\phi_h \in H(curl;\Omega):\ \ \phi_h|_K \in Q_{0,1} \times Q_{1,0},\ \forall\ K \in T_h\};$$

while on a triangular mesh,

$$\boldsymbol{U}_h = \{\psi_h \in L^2(\Omega):\ \ \psi_h|_K \text{ is a piecewise constant},\ \forall\ K \in T_h\},$$
$$\boldsymbol{V}_h = \{\phi_h \in H(curl;\Omega):\ \ \phi_h|_K = \text{span}\{\lambda_i \nabla\lambda_j - \lambda_j \nabla\lambda_i\},\ i, j = 1, 2, 3, \forall\ K \in T_h\}.$$

To accommodate the perfect conducting boundary condition $\mathbf{n}\times\mathbf{E} = \mathbf{0}$, we introduce the subspace of $\boldsymbol{V}_h$:

$$\boldsymbol{V}_h^0 = \{\phi_h \in \boldsymbol{V}_h,\ \ \mathbf{n} \times \phi_h = \mathbf{0}\ \text{ on } \partial\Omega\}.$$

Following [194], a leap-frog type scheme can be constructed for the modeling equations in the cloak region: For $n = 1, 2, \ldots$, find $\boldsymbol{D}_h^{n+\frac{1}{2}},\ \boldsymbol{E}_h^{n+\frac{1}{2}} \in \boldsymbol{V}_h^0$, $B_h^{n+1},\ H_h^{n+1} \in U_h$ such that

$$\left(\delta_\tau \boldsymbol{D}_h^{n+\frac{1}{2}}, \phi_h\right) - \left(H_h^n, \nabla \times \phi_h\right) = 0, \tag{9.60}$$

$$\begin{aligned}
&\left(\varepsilon_0\varepsilon_\phi \delta_\tau^2 \boldsymbol{E}_h^{n+\frac{1}{2}}, \widetilde{\phi}_h\right) + \left(\gamma\varepsilon_0\varepsilon_\phi \delta_{2\tau} \boldsymbol{E}_h^{n+\frac{1}{2}}, \widetilde{\phi}_h\right) + \left(\omega_p^2 \varepsilon_0\varepsilon_\phi \overline{\boldsymbol{E}}_h^{n-\frac{1}{2}}, \widetilde{\phi}_h\right) \\
&= \left(\left(M_A + \varepsilon_\phi M_B\right) \delta_\tau^2 \boldsymbol{D}_h^{n+\frac{1}{2}}, \widetilde{\phi}_h\right) + \left(\omega_p^2 M_A \overline{\boldsymbol{D}}_h^{n-\frac{1}{2}}, \widetilde{\phi}_h\right) \\
&\quad + \left(\gamma \left(M_A + \varepsilon_\phi M_B\right) \delta_{2\tau} \boldsymbol{D}_h^{n+\frac{1}{2}}, \widetilde{\phi}_h\right),
\end{aligned} \tag{9.61}$$

$$\left(\delta_\tau B_h^{n+1}, \psi_h\right) + \left(\nabla \times \boldsymbol{E}_h^{n+\frac{1}{2}}, \psi_h\right) = 0, \tag{9.62}$$

$$\begin{aligned}
&\left(\mu_0 A \delta_\tau^2 H_h^{n+1}, \widetilde{\psi}_h\right) + \left(\mu_0 A \gamma_m \delta_{2\tau} H_h^{n+1}, \widetilde{\psi}_h\right) + \left(\mu_0 A \omega_{pm}^2 \overline{H}_h^n, \widetilde{\psi}_h\right) \\
&= \left(\delta_\tau^2 B_h^{n+1}, \widetilde{\psi}_h\right) + \left(\gamma_m \delta_{2\tau} B_h^{n+1}, \widetilde{\psi}_h\right),
\end{aligned} \tag{9.63}$$

hold true for any $\phi_h, \widetilde{\phi}_h \in \boldsymbol{V}_h^0$, $\psi_h, \widetilde{\psi}_h \in U_h$. Here and below we denote the difference operators: For any $\boldsymbol{u}^n = \boldsymbol{u}(\cdot, t_n)$,

$$\delta_\tau \boldsymbol{u}^n = \frac{\boldsymbol{u}^n - \boldsymbol{u}^{n-1}}{\tau},\ \ \delta_\tau^2 \boldsymbol{u}^n = \frac{\boldsymbol{u}^n - 2\boldsymbol{u}^{n-1} + \boldsymbol{u}^{n-2}}{\tau^2},$$
$$\delta_{2\tau} \boldsymbol{u}^n = \frac{\boldsymbol{u}^n - \boldsymbol{u}^{n-2}}{2\tau},\ \ \overline{\boldsymbol{u}}^{n-1} = \frac{\boldsymbol{u}^n + 2\boldsymbol{u}^{n-1} + \boldsymbol{u}^{n-2}}{4},\ \ \widehat{\boldsymbol{u}}^n = \frac{\boldsymbol{u}^n + \boldsymbol{u}^{n-1}}{2}.$$

To couple (9.63) well with the PML equations (9.57), (9.58), and (9.63) is split into

$$\begin{aligned}&\left(\mu_0 A\delta_\tau^2 H_{zx,h}^{n+1},\widetilde{\psi}_h\right)+\left(\mu_0 A\gamma_m\delta_{2\tau}H_{zx,h}^{n+1},\widetilde{\psi}_h\right)+\left(\mu_0 A\omega_{pm}^2\overline{H}_{zx,h}^{n},\widetilde{\psi}_h\right)\\&=\frac{1}{2}\left(\delta_\tau^2 B_h^{n+1},\widetilde{\psi}_h\right)+\frac{1}{2}\left(\gamma_m\delta_{2\tau}B_h^{n+1},\widetilde{\psi}_h\right),\end{aligned}\tag{9.64}$$

$$\begin{aligned}&\left(\mu_0 A\delta_\tau^2 H_{zy,h}^{n+1},\widetilde{\psi}_h\right)+\left(\mu_0 A\gamma_m\delta_{2\tau}H_{zy,h}^{n+1},\widetilde{\psi}_h\right)+\left(\mu_0 A\omega_{pm}^2\overline{H}_{zy,h}^{n},\widetilde{\psi}_h\right)\\&=\frac{1}{2}\left(\delta_\tau^2 B_h^{n+1},\widetilde{\psi}_h\right)+\frac{1}{2}\left(\gamma_m\delta_{2\tau}B_h^{n+1},\widetilde{\psi}_h\right).\end{aligned}\tag{9.65}$$

Similarly, a leap-frog type scheme can be constructed for solving the Eqs. (9.59), (9.57), and (9.58) in the PML region: find $\boldsymbol{E}_h^{n+\frac{1}{2}}\in\boldsymbol{V}_h^0$, $H_{zx,h}^{n+1}$, $H_{zy,h}^{n+1}\in U_h$ such that

$$\varepsilon_0\left(\delta_\tau\boldsymbol{E}_h^{n+\frac{1}{2}},\widetilde{\boldsymbol{\phi}}_h\right)+\left(\begin{pmatrix}\sigma_y & 0\\ 0 & \sigma_x\end{pmatrix}\widehat{\boldsymbol{E}}_h^{n+\frac{1}{2}},\widetilde{\boldsymbol{\phi}}_h\right)=\left(H_{zx,h}^{n}+H_{zy,h}^{n},\nabla\times\widetilde{\boldsymbol{\phi}}_h\right),\tag{9.66}$$

$$\mu_0\left(\delta_\tau H_{zx,h}^{n+1},\psi_{1,h}\right)+\left(\sigma_{mx}\widehat{H}_{zx,h}^{n+1},\psi_{1,h}\right)=-\left(\frac{\partial}{\partial x}E_{y,h}^{n+\frac{1}{2}},\psi_{1,h}\right),\tag{9.67}$$

$$\mu_0\left(\delta_\tau H_{zy,h}^{n+1},\psi_{2,h}\right)+\left(\sigma_{my}\widehat{H}_{zy,h}^{n+1},\psi_{2,h}\right)=\left(\frac{\partial}{\partial y}E_{x,h}^{n+\frac{1}{2}},\psi_{2,h}\right),\tag{9.68}$$

hold true for any $\widetilde{\boldsymbol{\phi}}_h\in\boldsymbol{V}_h^0$, $\psi_{1,h},\psi_{2,h}\in U_h$.

In summary, the above developed mixed finite element time-domain algorithm for modeling the invisible cloak can be implemented as follows: first, construct a proper mesh $\mathscr{T}_h$ of Ω, choose a proper time step size τ and proper initial conditions; then at each time step n, perform the **FETD Algorithm**:

1. Solve (9.60) for $\boldsymbol{D}_h^{n+\frac{1}{2}}$ on $\mathscr{T}_h$.
2. Solve (9.61) and (9.66) for $\boldsymbol{E}_h^{n+\frac{1}{2}}$ on $\mathscr{T}_h$.
3. Solve (9.62) for $\boldsymbol{B}_h^{n+1}$ on $\mathscr{T}_h$.
4. Solve (9.64) and (9.67) for $H_{zx,h}^{n+1}$ on $\mathscr{T}_h$.
5. Solve (9.65) and (9.68) for $H_{zy,h}^{n+1}$ on $\mathscr{T}_h$.
6. Calculate $H_h^{n+1}=H_{zx,h}^{n+1}+H_{zy,h}^{n+1}$, then go back to step 1 and repeat the above process. Note that in the free space region, $\boldsymbol{E}_h^{n+\frac{1}{2}}$ and H^{n+1} are updated using (9.66)–(9.68) with $\sigma_x=\sigma_y=\sigma_{mx}=\sigma_{my}=0$.

9.3.2.1 Time-Domain Cloaking Simulation Results

The cloak simulation setup is shown in Fig. 9.12a, where the cloaked object is hided inside a perfectly electrically conducting cylinder with radius R_1, and the cylinder is wrapped by a cylindrical cloak with thickness $R_2 - R_1$.

In the cloak simulation, $R_1 = 0.1\,\text{m}$, $R_2 = 0.2\,\text{m}$ and $\gamma = \gamma_m = 0$ (i.e., no loss) are used in the Drude model. A plane wave source is specified by the function $H_z = 0.1\sin(\omega t)$, where $\omega = 2\pi f$ with operating frequency $f = 2\,\text{GHz}$. The parameters $\omega_p = \omega_{pm}$ is calculated by the Drude model $\omega_p = \omega\sqrt{1-\epsilon_r}$.

As mentioned in Sect. 9.2.3, in order to see the cloaking phenomenon, the mesh has to be fine enough. In the results presented below, the corresponding mesh is obtained by uniformly refining the given one in Fig. 9.12b four times, in which case the total number of edges used are 623,808, the DOFs for $\mathbf{E}$ is 621,376, and the total numbers of triangular elements and rectangular elements are 262,144 and 114,688, respectively. Hence the DOFs for H is 376,832. The time step is chosen as $\tau = 0.1$ picosecond (ps), and the total number of time steps is 50,000, i.e., $T = 5.0$ nanosecond (ns).

To see how wave propagates in the cloak structure, several snapshots of E_y fields are plotted in Fig. 9.13, which show clearly how the wave gets distorted in the cloak region. After 50,000 time steps, the plane wave pattern is almost restored, which renders the object placed inside the cloak region invisible to external electromagnetic fields.

9.4 Solar Cell Design with Metamaterials

In this section, we present an interesting application of metamaterials in solar cell design. This section is mainly derived from [192].

9.4.1 A Brief Introduction

A solar cell is a device that can directly convert solar energy into electricity through the photovoltaic effect. Generally speaking, a solar cell works in three steps: (1) Photons in sunlight hit the solar panel and are absorbed by some semiconducting materials; (2) Electrons are knocked loose from their atoms, thus forming an electric current flowing through the material; (3) An array of solar cells converts solar energy into electricity. Therefore, the operation of a solar cell requires three basic attributes: The absorption of light; The separation of various types of charge carriers; The extraction of those carriers to an external circuit.

A solar cell's performance is measured by its efficiency, which is usually broken down into reflectance efficiency, thermodynamic efficiency, charge carrier

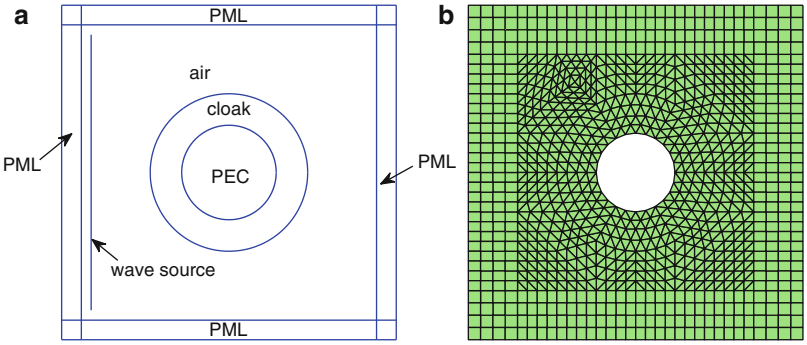

Fig. 9.12 (**a**) The cloak modeling setup; (**b**) A coarse mesh (Reprinted from Li et al. [194]. Copyright (2012), with permission from Elsevier)

separation efficiency and conductive efficiency. To reduce the cost of solar energy, high-efficiency solar cells are of interest.

Since various materials have different efficiencies and costs, creating cheap and efficient solar cells is an important research subject. Currently, many solar cells are made from bulk materials that are cut into wafers with thickness between 180–240 micrometers and are then processed like other semiconductors. The most prevalent bulk material for solar cells is crystalline silicon, which can be further classified into several categories such as monocrystalline silicon, polycrystalline silicon, and ribbon silicon.

Other solar cell materials are made as thin-films layers, organic dyes, and organic polymers. Thin-film technologies reduce the amount of materials used in solar cells. However, the majority of thin film panels have quite low conversion efficiencies and occupy large areas per watt production. Cadmium telluride, copper indium gallium selenide and amorphous silicon are three thin-film technologies often used as outdoor photovoltaic solar power production. Silicon remains the most popular material used in both bulk and thin-film forms. Silicon thin-film cells are mainly deposited by chemical vapor deposition from silane gas and hydrogen gas. Though solar cells made from various silicons such as amorphous silicon and protocrystalline silicon are cheap to produce, they still have lower energy conversion efficiency than bulk silicon.

In recent years, nanotechnology has been applied to solar cell materials, which can be made from nanocrystals and quantum dots. For example, large parallel nanowire arrays enable long absorption lengths, which can trap more light and hence improve the efficiency of the solar cell.

In the rest section we present an approach for solar cell design, which uses nanomaterials, more specficially electromagnetic metamaterials, to increase the solar cell efficiency. This approach is based on the metamaterial's striking re-focusing property (cf. Fig. 9.4 (Left)): In a planar negative-index metamaterial slab, an evanescent wave decaying away from an object grows exponentially inside

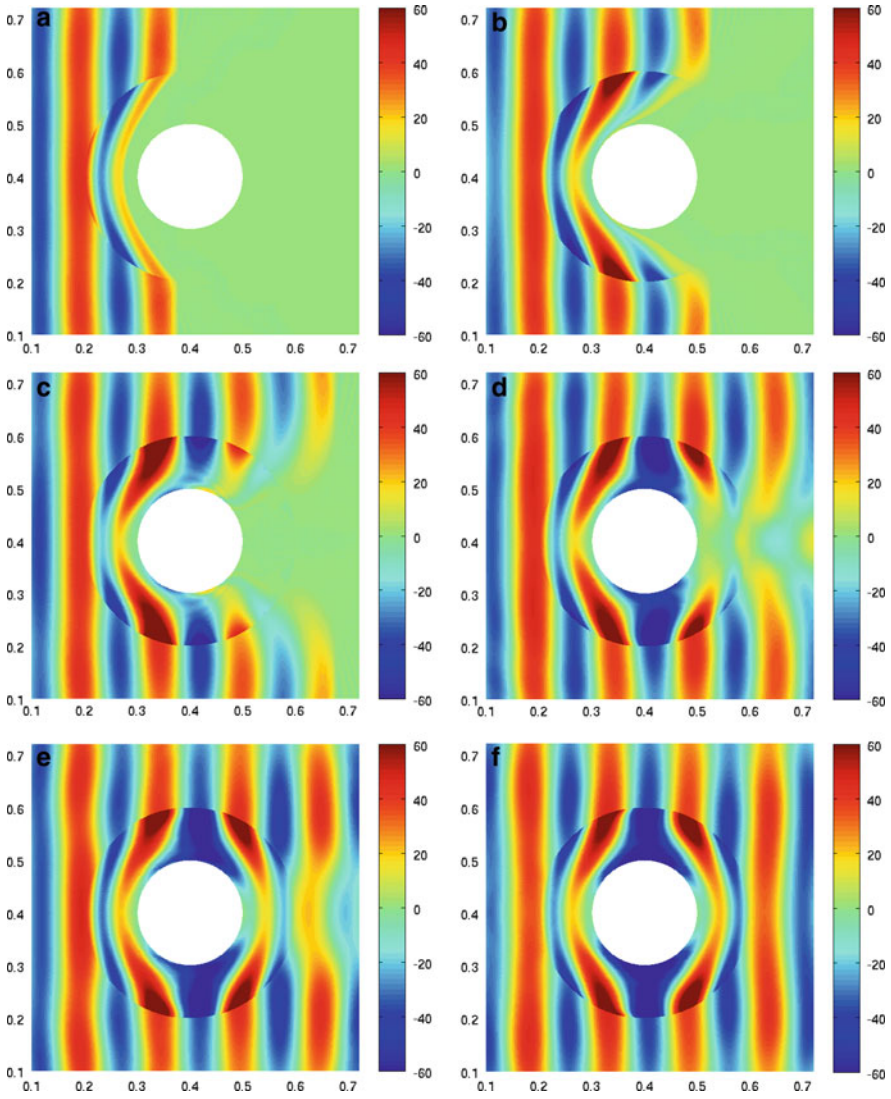

Fig. 9.13 E_y at (**a**) $t = 1.0$ ns; (**b**) $t = 1.5$ ns; (**c**) $t = 2.0$ ns; (**d**) $t = 3.0$ ns; (**e**) $t = 4.0$ ns; (**f**) $t = 5.0$ ns (Reprinted from Li et al. [194]. Copyright (2012), with permission from Elsevier)

the slab, and refocuses the source wave at the exit interface if the slab thickness equals the distance from the wave source to the slab's front interface. This property shows that metamaterials can be efficient subwavelength absorbers [14]. Ultra-thin metamaterial slabs have been shown [92] to sustain their high absorptivity for a wide range of incident angles. This property is highly desirable for developing efficient thermalphotovoltaics [175] and photovoltaics [101]. In photovoltaic applications,

the efficiency of solar cells can be enhanced by the strong field resonance inside the absorbing metamaterial.

9.4.2 The Mathematical Formulation

Modeling of solar cells boils down to solving Maxwell's equations:

$$\nabla \times \tilde{\mathbf{H}} = \frac{\partial \tilde{\mathbf{D}}}{\partial t}, \tag{9.69}$$

$$\nabla \times \tilde{\mathbf{E}} = -\frac{\partial \tilde{\mathbf{B}}}{\partial t}, \tag{9.70}$$

where $\tilde{\mathbf{E}}(\mathbf{x}, t)$ and $\tilde{\mathbf{H}}(\mathbf{x}, t)$ are the electric and magnetic fields, and $\tilde{\mathbf{D}}(\mathbf{x}, t)$ and $\tilde{\mathbf{B}}(\mathbf{x}, t)$ are the corresponding electric and magnetic flux densities. For linear electromagnetic materials, these variables are connected through the constitutive relations:

$$\tilde{\mathbf{D}} = \epsilon_0 \epsilon_r \tilde{\mathbf{E}}, \quad \tilde{\mathbf{B}} = \mu_0 \mu_r \tilde{\mathbf{H}}, \tag{9.71}$$

where ϵ_r and μ_r are the relative permittivity and permeability, respectively.

Substituting (9.71) into (9.69), (9.70), and using the time harmonic form

$$\tilde{\mathbf{E}}(\mathbf{x}, t) = \mathbf{E}(\mathbf{x}) e^{j\omega t}, \quad \tilde{\mathbf{H}}(\mathbf{x}, t) = \mathbf{H}(\mathbf{x}) e^{j\omega t},$$

we can transform the time-dependent Maxwell's equations into the time harmonic form:

$$j\omega \epsilon_0 \epsilon_r \mathbf{E} = \nabla \times \mathbf{H}, \tag{9.72}$$

$$j\omega \mu_0 \mu_r \mathbf{H} = -\nabla \times \mathbf{E}, \tag{9.73}$$

where ω denotes the wave frequency. Note that (9.72) and (9.73) can be further reduced to a simple vector wave equation in terms of either the electric field or the magnetic field:

$$\nabla \times (\mu_r^{-1} \nabla \times \mathbf{E}) - k_0^2 \epsilon_r \mathbf{E} = 0, \tag{9.74}$$

$$\nabla \times (\epsilon_r^{-1} \nabla \times \mathbf{H}) - k_0^2 \mu_r \mathbf{H} = 0. \tag{9.75}$$

Here $k_0 = \frac{\omega}{C_v} = \omega \sqrt{\epsilon_0 \mu_0}$ denotes the wave number of free space.

Under the assumption that the material is non-magnetic (hence $\mu_r = 1$), we can use the refractive index $n = \sqrt{\epsilon_r \mu_r}$ to rewrite (9.74) and (9.75) as follows:

$$\nabla \times (\nabla \times \mathbf{E}) - k_0^2 n^2 \mathbf{E} = 0, \tag{9.76}$$

$$\nabla \times (n^{-2}\nabla \times \mathbf{H}) - k_0^2 \mathbf{H} = 0. \tag{9.77}$$

To efficiently model the 2-D solar cells, we first solve (9.77) for the unknown magnetic field $H = H_z$, then postprocess H by using (9.72) to obtain the unknown electric field $\mathbf{E} = (E_x, E_y)'$, i.e.,

$$E_x = \frac{1}{j\omega\epsilon_0\epsilon_r}\frac{\partial H}{\partial y}, \quad E_y = \frac{-1}{j\omega\epsilon_0\epsilon_r}\frac{\partial H}{\partial x}.$$

9.4.3 Numerical Simulations

9.4.3.1 A Benchmark Problem

A benchmark problem of [295] is solved in [192] by using the commercial multiphysics finite element package COMSOL. The proposed solar cell structure is uniform in the z-direction. The unit cell (illustrated in Fig. 9.14) has periodic boundary conditions in the x-direction, and contains a benzocyclobutene (BCB) layer with thickness $c = 50\,\text{nm}$ and a gold substrate with thickness $b = 100\,\text{nm}$. The gold substrate is used to absorb the radiation energy coming into the cell. Furthermore, there is a gold strip of dimension $f \times e$ embedded inside the BCB layer. In Fig. 9.14, g denotes the gap between the strip and the BCB boundary. To obtain a good absorption for the solar cell, we choose $g = 15\,\text{nm}$, $f = 18\,\text{nm}$, and $e = 256\,\text{nm}$.

The permittivity for gold is modeled by the Drude model

$$\epsilon_r(\omega) = 1 - \frac{\omega_p^2}{\omega(\omega + i\gamma)},$$

where the plasma frequency ω_p and the collision frequency γ are calculated as

$$\gamma = \frac{\omega\epsilon_2}{1-\epsilon_1}, \quad \omega_p = \sqrt{(1-\epsilon_1)(\omega^2+\gamma^2)},$$

where ϵ_1 and ϵ_2 are functions of the incident wavelength λ, obtained through polynomial fitting:

$$\begin{aligned}\epsilon_1(\lambda) &= -1.1\lambda^3 - 39\lambda^2 - 12\lambda + 12,\\ \epsilon_2(\lambda) &= 7.3\lambda^8 - 100\lambda^7 + 580\lambda^6 - 1{,}900\lambda^5 + 3{,}700\lambda^4\\ &\quad -4{,}400\lambda^3 + 3{,}200\lambda^2 - 1{,}300\lambda + 210.\end{aligned}$$

For a P-polarized radiation at frequency $2.89 \cdot 10^{14}\,\text{Hz}$ (which is in the infrared region) with 0° incident angle (i.e., penetrating the solar cell vertically), the obtained

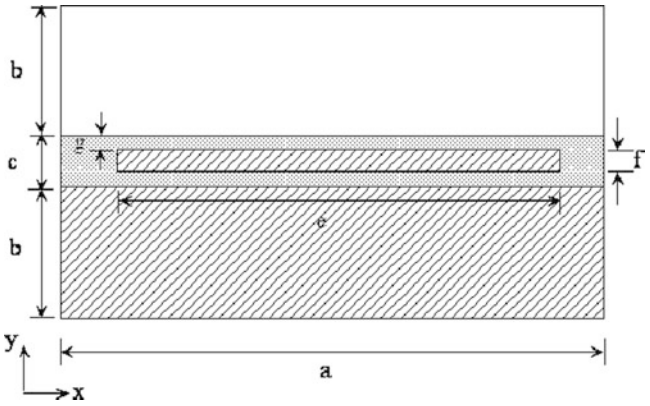

Fig. 9.14 The unit cell structure for the benchmark problem (With permission from Global Science Press [192])

electric and magnetic field magnitudes $|E_x|, |E_y|$ and $|H_z|$ are plotted in Fig. 9.15. In this simulation, the refractive index $n = 1.56$ is chosen for BCB, the port condition at both incident and exit surfaces is used, and Floquet periodic boundary condition is imposed in the x-direction. Furthermore, tangential continuity across subdomain interfaces is imposed.

Many numerical experiments are carried out by varying the wave incident angles and wave frequencies in the infrared (IR) and visible region. Figure 9.16 shows how the absorption varies with the incident angles in the IR and visible region. In COMSOL, the absorption on a fixed port is defined as $1 - |S_{11}|^2$, where S_{11} is the reflection coefficient expressed as

$$S_{11} = \sqrt{\text{Power reflected from the port}} / \sqrt{\text{Power incident on the port}}.$$

Furthermore, Fig. 9.16 shows that the absorption decreases as the incident angle increases for fixed frequencies, and the average absorption is about 60 % for all incident angles between 0° and 40°.

9.4.3.2 Results with Other Metals and Micro-structures

Considering the cost of gold and the uncommon BCB material, several combinations of materials were tested in [192] for their absorptions in order to find a cheap but efficient solar cell design. Detailed numerical experiments were performed with gold replaced by high melting point metals such as copper, nickel and tungsten, and with BCB replaced by dielectric SiO_2, semiconductor C[100] and Poly-Si. Numerical results of [192] showed that the average absorption over the visible frequencies for most combinations is about 50 %.

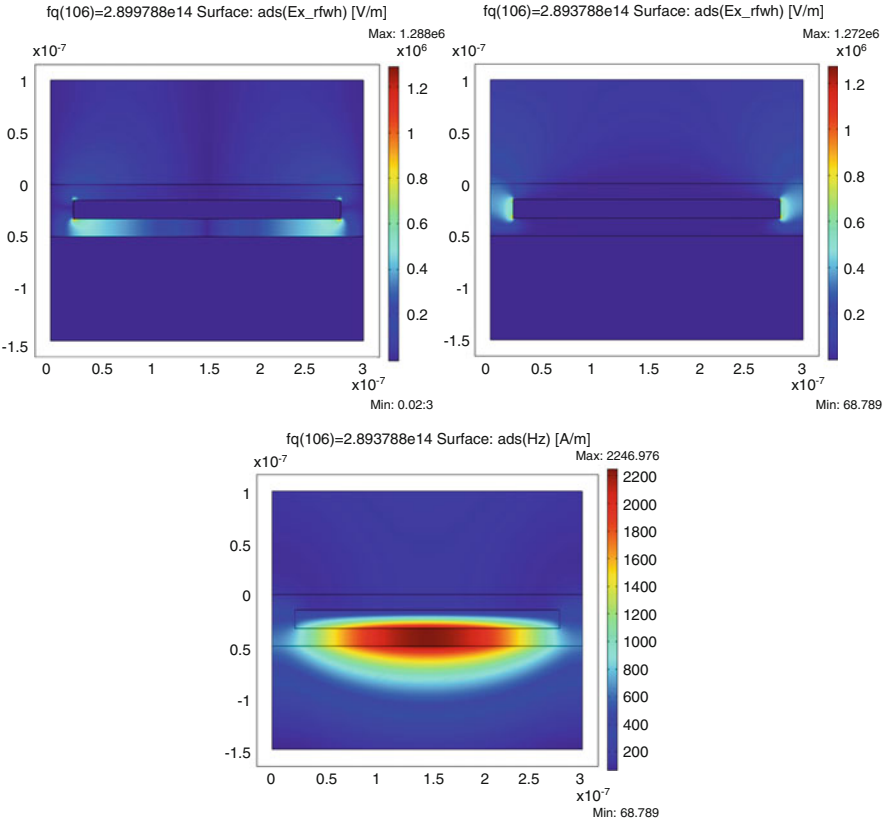

Fig. 9.15 Simulation results for the benchmark problem: (*Top Right*) Plot of $|E_x|$; (*Top Left*) Plot of $|E_y|$; (*Bottom*) Plot of $|H_z|$ (With permission from Global Science Press [192])

To reduce the usage of metals (hence the weight of the solar cell), [192] also considered using several micro-structures to replace the metallic strip. For example, in a nickel and Poly-Si combination, a micro-structure consisting of 44 equal rectangles with dimensions of 5 by 10 nm was tested with different frequency waves penetrating the solar cell vertically. An exemplary power flow obtained from this structure is presented in Fig. 9.17.

Another micro-structure consisting of 44 equal circles with 5 nm radius was tested for a nickel and Poly-Si combination in [192]. An exemplary power flow obtained from this micro-structure is presented in Fig. 9.18.

In [192], both the rectangular and circular micro-structures were shown to have about 80 % absorption for the vertically penetrating waves in the entire solar spectrum. Results of [192] suggest that efficient solar cells can be constructed using metamaterials.

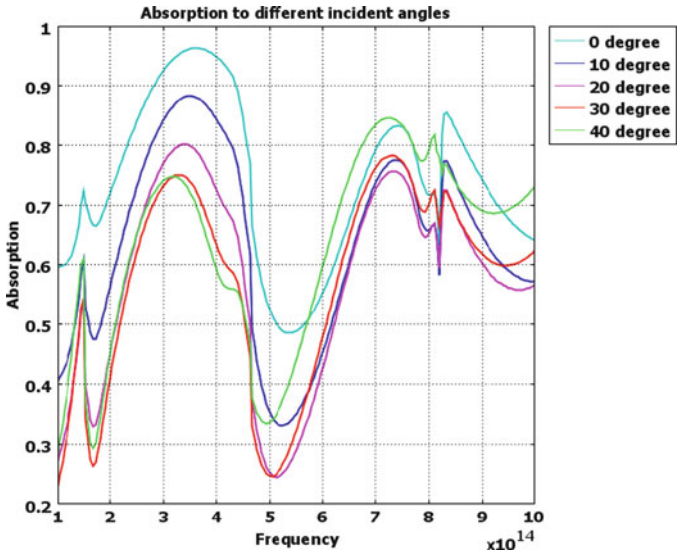

Fig. 9.16 The benchmark problem: the absorption corresponding to the infrared and visible frequencies with various incident angles (With permission from Global Science Press [192])

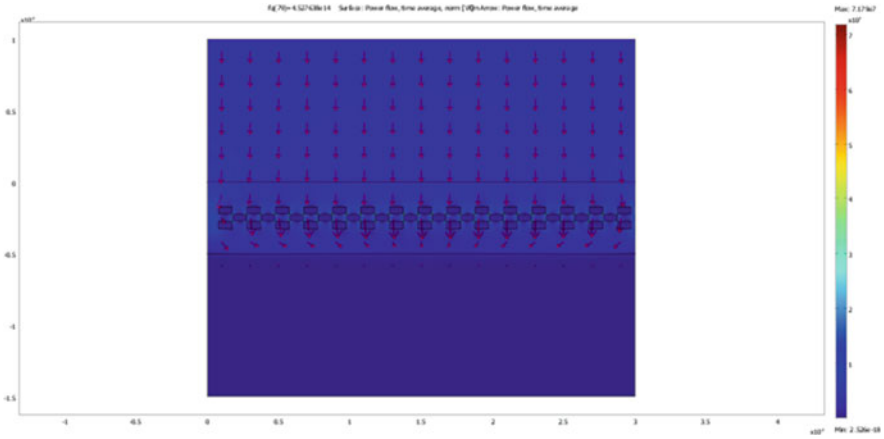

Fig. 9.17 The power flow obtained with 44 micro-rectangles (With permission from Global Science Press [192])

9.5 Problems Needing Special Attention

9.5.1 Unit Cell Design and Homogenization

The metamaterials discussed in this book are structured composites which lead to simultaneously negative permittivity and permeability. A key issue in the theory of composites [211] is the study of how their physical properties such as permittivity

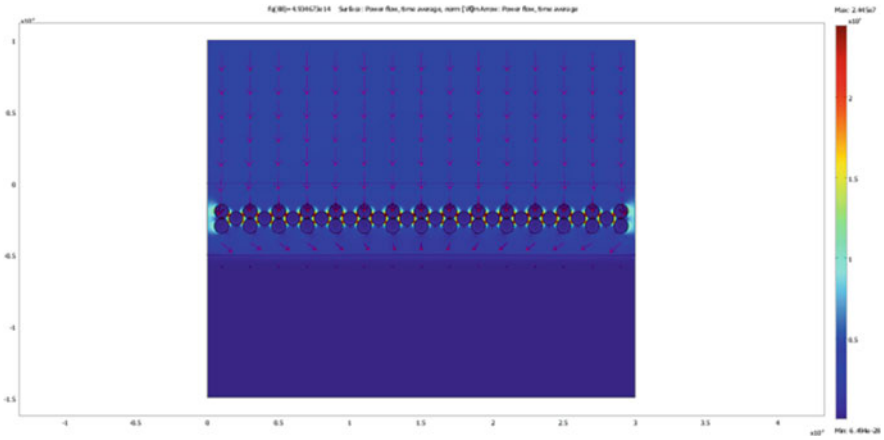

Fig. 9.18 The power flow obtained with 44 micro-circles (With permission from Global Science Press [192])

and permeability depend on the microstructure (or unit cell). When the period of the microstructure is small compared to the wavelength, the physical parameters in Maxwell's equations oscillate rapidly, which makes the numerical simulation very challenging. In this case, homogenization approach [33, 81, 250] is often used: the rapidly oscillating parameters are replaced with effective constitutive parameters. A distinguishing feature of the homogenization problem for metamaterials is that the cell size a is not vanishingly small compared to the vacuum wavelength λ_0 at a given frequency. The typical range in practice is $a \sim 0.1 - 0.3\lambda_0$. Hence, the classical homogenization procedures valid for $a \to 0$ have limited applicability for metamaterials. In the physics and engineering community, the homogenized material parameters are often calculated using S-parameter retrieval method [71, 196], the field-averaging method [270], and other averaging operations such as Maxwell-Garnett, Bruggeman and Clausius-Mossotti mixing formulas (cf. [264] and references cited therein). Rigorous mathematical treatment of the homogenization of metamaterials is still in its early stages [27, 47, 165, 231], and much more work is needed in this direction.

In the mathematics community, there are already many studies devoted to homogenization of Maxwell's equations. It is well known that homogenization results can be obtained by the classical method of asymptotic expansions in two scales for Maxwell's equations (see e.g. [33, 250]). Homogenization for the non-stationary Maxwell's system is discussed in books [161, 250]. Using the two-scale convergence method, Wellander [291] obtained convergence results for the homogenization method for the time-dependent Maxwell's equations in a heterogeneous medium obeying linear constitutive relations. Barbatis and Stratis [27] studied the periodic homogenization of Maxwell's equations for dissipative bianisotropic media in the time domain. Using the periodic unfolding method (originally introduced by Cioranescu et al. [80] in the abstract framework of elliptic equations), Bossavit,

Griso, and Miara [46] investigated the behavior of the electromagnetic field of a medium with periodic microstructures made of bianisotropic material and proved convergence results for their homogenization method. In 2006, Banks et al. [24] used the periodic unfolding method and derived homogenization results of the nonstationary Maxwell's equations in dispersive media. On the other hand, there are many more homogenization publications for time-harmonic Maxwell's equations than the time-dependent Maxwell's equations (see e.g. [60, 169, 268] and references therein).

To bring interested readers to the forefront of homogenization, below we present two examples of homogenization of time-dependent Maxwell's equations in composite materials with periodic microsctructures in 3-D. The first example is for Maxwell's equations written in one vector wave equation. The second example is for Maxwell's equations expressed as a system of first-order differential equations.

9.5.1.1 Homogenization via Multiscale Asymptotic Expansion

Consider the time-dependent Maxwell's equations with rapidly oscillatory coefficients as follows:

$$\frac{\partial^2 \mathbf{u}^\alpha(x,t)}{\partial t^2} + \operatorname{curl}(A(\frac{x}{\alpha})\operatorname{curl}\mathbf{u}^\alpha(x,t)) = \mathbf{f}(x,t), \quad \text{in } \Omega \times (0,T), \tag{9.78}$$

$$\nabla \cdot \mathbf{u}^\alpha = 0, \quad \text{in } \Omega \times (0,T), \tag{9.79}$$

$$\mathbf{n} \times \mathbf{u}^\alpha(x,t) = 0, \quad \text{on } \partial\Omega \times (0,T), \tag{9.80}$$

$$\mathbf{u}^\alpha(x,0) = \mathbf{u}_0(x), \ \ \partial_t \mathbf{u}^\alpha(x,0) = \mathbf{u}_1(x), \text{in } \Omega, \tag{9.81}$$

where $\mathbf{u}$ can be either the electric field $\mathbf{E}$ or magnetic field $\mathbf{H}$, and the matrix function $A(\frac{x}{\alpha}) = (a_{ij}(\frac{x}{\alpha})), i, j = 1, 2, 3$. Here the small number $\alpha > 0$ represents the size of the periodic microstructure of a composite material (see Fig. 9.19).

Note that when α becomes quite small, direct accurate numerical computation of the solution $\mathbf{u}^\alpha$ is very challenging or even impossible since a very fine mesh is required which leads to a prohibitive amount of memory storage and computational time. For clarity, we denote $\xi = \frac{x}{\alpha}$. In the classic multiscale asympototic method, x and ξ are called "slow" and "fast" variables, respectively. Furthermore, we denote $Q = (0,1)^3$ for the reference cell of the periodic structure. The remaining material of this subsection is essentially from [33, Sect. 11 of Chap. 1] and [60, 302].

The solution $\mathbf{u}^\alpha$ to the problem (9.78)–(9.81) can be approximated by the two-scale asymptotic expansion

$$\mathbf{u}^\alpha(x,t) = \mathbf{u}^*(x,t) + \alpha\theta_1(\xi)\operatorname{curl}\mathbf{u}^*(x,t) + \alpha^2\theta_2(\xi)\operatorname{curl}^2\mathbf{u}^*(x,t) + \cdots, \tag{9.82}$$

where $\operatorname{curl}^2 = \operatorname{curlcurl}$, $\mathbf{u}^*$ is the solution to a homogenized problem, and $\theta_1(\xi)$ and $\theta_2(\xi)$ are corrector functions. Details are given below.

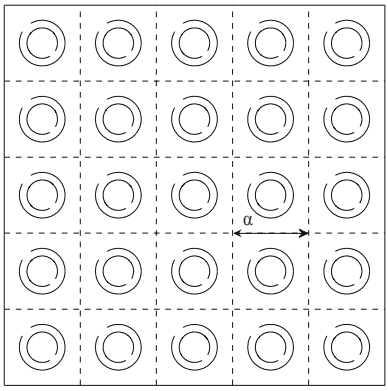

Fig. 9.19 A composite material with periodic microstructures

Substituting (9.82) into (9.78), using the fact that curl $=$ curl$_x$ $+$ curl$_\xi$, and equalizing all terms with power α^{-1}, we have: for $k = 1, 2, 3$,

$$\text{curl}_\xi(A(\xi)\text{curl}_\xi\theta_1^k(\xi) + A(\xi)e_k) = 0, \quad \text{in } Q, \tag{9.83}$$

$$\nabla_\xi \cdot \theta_1^k(\xi) = 0, \quad \text{in } Q, \tag{9.84}$$

$$\mathbf{n} \times \theta_1^k(\xi) = 0, \quad \text{on } \partial Q, \tag{9.85}$$

where e_k is the canonical basis of R^3. Using $\theta_1^k(\xi)$, we can form the matrix cell function $\theta_1(\xi) = (\theta_1^1(\xi), \theta_1^2(\xi), \theta_1^3(\xi))$.

Similarly, equalizing the terms with power α^0, we can define $\tilde{\theta}_2^k(\xi), k = 1, 2, 3$, which satisfy

$$\text{curl}_\xi(A(\xi)\text{curl}_\xi\tilde{\theta}_2^k(\xi)) = -\text{curl}_\xi(A(\xi)\theta_1^k(\xi)) + G(\xi), \quad \text{in } Q, \tag{9.86}$$

$$\mathbf{n} \times \tilde{\theta}_2^k(\xi) = 0, \quad \text{on } \partial Q, \tag{9.87}$$

where $G(\xi) = -A(\xi)\text{curl}_\xi\theta_1^k(\xi) - A(\xi)e_k + A^*e_k$, and A^* is the homogenized coefficient matrix defined as (cf. (11.44) of [33, p. 145])

$$A^* = \int_Q (A(\xi) + A(\xi)\text{curl}_\xi\theta_1(\xi))d\xi. \tag{9.88}$$

Note that if $\nabla_\xi \cdot G(\xi) \neq 0$, then no solution exists for Eq. (9.86). To avoid this issue, we can introduce scalar functions $\phi^k(\xi), k = 1, 2, 3$, which satisfy

$$-\Delta_\xi\phi^k(\xi) = \nabla_\xi \cdot G(\xi) \quad \text{in } Q, \tag{9.89}$$

$$\phi^k(\xi) = 0, \quad \text{on } \partial Q. \tag{9.90}$$

Now we can define $\theta_2^k(\xi), k = 1, 2, 3$ such that

$$\begin{aligned}
&\text{curl}_\xi(A(\xi)\text{curl}_\xi\theta_2^k(\xi)) \\
&= -\text{curl}_\xi(A(\xi)\theta_1^k(\xi)) + G(\xi) + \nabla_\xi\phi^k(\xi), \quad \text{in } Q, \qquad (9.91)\\
&\nabla_\xi \cdot \theta_2^k(\xi) = 0, \quad \text{in } Q, \qquad (9.92)\\
&\mathbf{n} \times \theta_2^k(\xi) = 0, \quad \text{on } \partial Q, \qquad (9.93)
\end{aligned}$$

from which we construct the matrix cell function $\theta_2(\xi) = (\theta_2^1(\xi), \theta_2^2(\xi), \theta_2^3(\xi))$.

It is shown [302] that $\mathbf{u}^*(x,t)$ is the solution to the following homogenized Maxwell's equations:

$$\begin{aligned}
&\frac{\partial^2\mathbf{u}^*(x,t)}{\partial t^2} + \text{curl}(A^*\text{curl}\,\mathbf{u}^*(x,t)) = \mathbf{f}(x,t), \quad \text{in } \Omega \times (0,T), \qquad (9.94)\\
&\nabla \cdot \mathbf{u}^* = 0, \quad \text{in } \Omega \times (0,T), \qquad (9.95)\\
&\mathbf{n} \times \mathbf{u}^*(x,t) = 0, \quad \text{on } \partial\Omega \times (0,T), \qquad (9.96)\\
&\mathbf{u}^*(x,0) = \mathbf{u}_0(x), \ \ \partial_t\mathbf{u}^*(x,0) = \mathbf{u}_1(x), \quad \text{in } \Omega. \qquad (9.97)
\end{aligned}$$

It is proved [250] that if $A(\xi)$ is symmetric and positive definite, and elements $a_{ij}(\xi)$ are 1-periodic in ξ, then the problem (9.78)–(9.81) has a unique solution $\mathbf{u}^\alpha(x,t) \in L^2(0,T;H(\text{curl};\Omega)) \cap C^0(0,T;(L^2(\Omega))^3)$ under the assumption that $\mathbf{f} \in L^2(0,T;(L^2(\Omega))^3), \mathbf{u}_0 \in (H^1(\Omega))^3, \mathbf{u}_1 \in (L^2(\Omega))^3$. Furthermore,

$$\mathbf{u}^\alpha(x,t) \to \mathbf{u}^*(x,t) \quad \text{in } L^\infty(0,T;(L^2(\Omega))^3) \text{ weakly } \star.$$

Under more regularity assumptions, Zhang et al. [302] proved that the two-scale asymptotic expansions for problem (9.78)–(9.81):

$$\mathbf{u}_1^\alpha(x,t) = \mathbf{u}^*(x,t) + \alpha\theta_1(\xi)\text{curl}\,\mathbf{u}^*(x,t)$$

and

$$\mathbf{u}_2^\alpha(x,t) = \mathbf{u}^*(x,t) + \alpha\theta_1(\xi)\text{curl}\,\mathbf{u}^*(x,t) + \alpha^2\theta_2(\xi)\text{curl}^2\,\mathbf{u}^*(x,t)$$

converges to $\mathbf{u}^\alpha(x,t)$ uniformly in α. More specifically, they proved

$$||\mathbf{u}^\alpha(x,t) - \mathbf{u}_k^\alpha(x,t)||_{L^\infty(0,T;H(\text{curl};\Omega))} + ||\partial_t(\mathbf{u}^\alpha(x,t) - \mathbf{u}_k^\alpha(x,t))||_{L^\infty(0,T;(L^2(\Omega))^3)} \le C\alpha$$

for $k = 1, 2$. Here C is a constant independent of the small structure size α.

Finally, we like to remark that the homogenized problem (9.94)–(9.97) is a Maxwell system with constant coefficients, and it can be solved by various numerical methods on a relatively coarse mesh. Furthermore, the corrector functions

$\theta_1(\xi)$ and $\theta_2(\xi)$ need to be solved on a unit cell only once. Hence an efficient multiscale finite element method can be developed for Maxwell's equations with rapidly oscillatory coefficients. For details, see [60, 302].

9.5.1.2 Homogenization by the Periodic Unfolding Method

The periodic unfolding method was introduced by Cioranescu et al. [80] in the abstract framework of elliptic equations, and later were extended to Maxwell's equations used for simulating the electromagnetic field in composite media with spatially periodic microstructures [24, 46], from which this subsection is mainly derived.

Consider the Maxwell's equations posted on $\Omega \times (0, T)$:

$$\frac{\partial \mathbf{D}(x,t)}{\partial t} = \text{curl}\, \mathbf{H}(x,t) - \mathbf{J}_s(x,t), \tag{9.98}$$

$$\frac{\partial \mathbf{B}(x,t)}{\partial t} = -\text{curl}\, \mathbf{E}(x,t), \tag{9.99}$$

with initial conditions

$$\mathbf{E}(x,0) = \mathbf{E}_0(x), \quad \mathbf{H}(x,0) = \mathbf{H}_0(x),$$

and the perfect conducting boundary condition

$$\mathbf{n} \times \mathbf{E} = 0 \quad \text{on } \partial\Omega \times (0,T).$$

This system is completed by the general constitutive laws

$$\mathbf{D}(x,t) = \epsilon_0 \epsilon_r(x)\mathbf{E}(x,t) + \int_0^t \{\sigma_E(x) + \nu_E(x, t-s)\}\mathbf{E}(x,s)ds, \tag{9.100}$$

$$\mathbf{B}(x,t) = \mu_0 \mu_r(x)\mathbf{H}(x,t) + \int_0^t \{\sigma_H(x) + \nu_H(x, t-s)\}\mathbf{H}(x,s)ds, \tag{9.101}$$

where ϵ_0 and μ_0 are the permittivity and permeability of free space, ϵ_r and μ_r are the relative permittivity and permeability of the media, σ_E and σ_H are the electric and magnetic conductivities, μ_E and μ_H are the electric and magnetic susceptibilities, and $\mathbf{J}_s$ is the source current density.

Let us introduce the vector of fields

$$\mathbf{u} = \begin{pmatrix} \mathbf{E} \\ \mathbf{H} \end{pmatrix} \in W^{1,1}(0, T; H^1(\Omega; R^6)), \tag{9.102}$$

and the operator

$$L\mathbf{u}(x,t) = \begin{pmatrix} \mathbf{D}(x,t) \\ \mathbf{B}(x,t) \end{pmatrix}, \tag{9.103}$$

which can be written as

$$L\mathbf{u}(x,t) = A(x)\mathbf{u}(x,t) + \int_0^t \{F(x) + G(x,t-s)\}\mathbf{u}(x,s)ds, \tag{9.104}$$

where the 6×6 matrices A, F and G are defined as

$$A(x) = \begin{pmatrix} \epsilon_0\epsilon_r(x)I_3 & 0_3 \\ 0_3 & \mu_0\mu_r(x)I_3 \end{pmatrix}, \quad F(x) = \begin{pmatrix} \sigma_E I_3 & 0_3 \\ 0_3 & \sigma_H I_3 \end{pmatrix}, \tag{9.105}$$

$$G(x,t) = \begin{pmatrix} \nu_E(x,t)I_3 & 0_3 \\ 0_3 & \nu_H(x,t)I_3 \end{pmatrix}. \tag{9.106}$$

Here I_3 and 0_3 denote the 3×3 identity and zero matrices, respectively.

Furthermore, we define the Maxwell operator M as

$$M\mathbf{u}(x,t) = M\begin{pmatrix} \mathbf{E} \\ \mathbf{H} \end{pmatrix} = \begin{pmatrix} \text{curl}\,\mathbf{H}(x,t) \\ -\text{curl}\,\mathbf{E}(x,t) \end{pmatrix} \tag{9.107}$$

and the vector $\mathbf{J}_s(t) = (\mathbf{J}_s(t), 0, 0, 0) \in R^6$. Hence we can rewrite the Maxwell's equations in the form

$$\frac{d}{dt}L\mathbf{u} = M\mathbf{u} - \mathbf{J}_s(t), \quad \text{in } \Omega \times (0,T), \tag{9.108}$$

$$\mathbf{u}(x,0) = \mathbf{u}^0(x), \quad \text{in } \Omega, \tag{9.109}$$

$$\mathbf{n} \times \mathbf{u}_1(x,t) = 0, \quad \text{on } \partial\Omega \times (0,T), \tag{9.110}$$

where $\mathbf{u}_1 = \mathbf{E}$.

Now we assume that the medium occupying the domain Ω has periodic microstructures, i.e., $\epsilon_r, \mu_r, \sigma_E, \sigma_H, \nu_E$ and ν_H are highly oscillatory functions in space, which lead to matrices A, F and G with spatially oscillatory coefficients. In this case, we have to solve the following Maxwell's equations:

$$\begin{aligned} &\frac{d}{dt}(A^\alpha(x)\mathbf{u}^\alpha(x,t) + \int_0^t \{F^\alpha(x) + G^\alpha(x,t-s)\}\mathbf{u}^\alpha(x,s)ds) \\ &\qquad = M\mathbf{u}^\alpha - \mathbf{J}_s(t), \quad \text{in } \Omega \times (0,T), \end{aligned} \tag{9.111}$$

$$\mathbf{u}^\alpha(x,0) = \mathbf{u}^0(x), \quad \text{in } \Omega, \tag{9.112}$$

$$\mathbf{n} \times \mathbf{u}_1^\alpha(x,t) = 0, \quad \text{on } \partial\Omega \times (0,T), \tag{9.113}$$

where we assume that the periodic structure is characterized by an elementary microstructure with size $\alpha > 0$, i.e., we assume that

$$A^\alpha(x) = A(x, \frac{x}{\alpha}), \quad F^\alpha(x) = F(x, \frac{x}{\alpha}), \quad G^\alpha(x) = G(x, \frac{x}{\alpha}).$$

We approximate the solution $\mathbf{u}^\alpha$ of (9.111)–(9.113) by the two-scale expansion

$$\mathbf{u}^\alpha(x,t) = \mathbf{u}^*(x) + \nabla_\xi \overline{\mathbf{u}}(x,\xi) + \cdots, \quad x \in \Omega,\ \xi \in Q, \tag{9.114}$$

where $\mathbf{u}^*$ is a solution to the homogenized problem (9.115) and (9.116) shown below, and $\overline{\mathbf{u}}$ is the first corrector term. Recall that Q denotes the reference cell $(0,1)^3$. Before we present the specific form of the homogenized problem, we have to introduce some corrector functions first.

Let us denote $H^1_{per}(Q)$ for the space of periodic functions with vanishing mean value. By Bossavit et al. [46, p. 848], corrector functions $\overline{w}^A_k \in H^1_{per}(Q; R^2)$, $\overline{w}_k \in W^{1,1}(0,T; H^1_{per}(Q;R^2))$ and $\overline{w}^0_k \in W^{2,1}(0,T; H^1_{per}(Q;R^2))$ are solutions to the following variational problems:

$$(i) \quad \int_Q A(\xi)\nabla_\xi \overline{w}^A_k \cdot \nabla_\xi \overline{v}(\xi) d\xi = -\int_Q A(\xi) e_k \cdot \nabla_\xi \overline{v}(\xi) d\xi,$$

$$(ii) \quad \int_Q \{A(\xi)\nabla_\xi \overline{w}_k(\xi,t) + \int_0^t (F(\xi) + G(\xi, t-s))\nabla_\xi \overline{w}_k(\xi,s) ds\} \cdot \nabla_\xi \overline{v}(\xi) d\xi$$
$$= -\int_Q (F(\xi) + G(\xi,t))(e_k + \nabla_\xi \overline{w}^A_k) \cdot \nabla_\xi \overline{v}(\xi) d\xi,$$

$$(iii) \quad \int_Q \{A(\xi)\nabla_\xi \overline{w}^0_k(\xi,t) + \int_0^t (F(\xi) + G(\xi, t-s))\nabla_\xi \overline{w}^0_k(\xi,s) ds\} \cdot \nabla_\xi \overline{v}(\xi) d\xi$$
$$= -\int_Q A(\xi) e_k \cdot \nabla_\xi \overline{v}(\xi) d\xi,$$

for all $\overline{v} \in H^1_{per}(Q;R^2)$ and $k = 1, 2, \cdots, 6$. Here the vector e_k is the canonical basis of R^6, and ∇_ξ is the divergence operator defined as $\nabla_\xi = (\nabla_{\xi_1}, \nabla_{\xi_2}, \nabla_{\xi_3})' \in R^{3\times 1}$. For a vector $\mathbf{v} = (v_1, v_2)'$, where v_1 and v_2 are scalar functions, we define $\nabla_\xi \mathbf{v} = (\nabla_\xi v_1, \nabla_\xi v_2)' \in R^{6\times 1}$.

It is proved [46] that $\mathbf{u}^*$ is the solution to the homogenized problem:

$$\frac{d}{dt} L^* \mathbf{u} = M\mathbf{u} - \mathbf{J}_s - \mathbf{J}^0, \quad \text{in } \Omega \times (0,T), \tag{9.115}$$

$$\mathbf{u}(x,0) = \mathbf{u}^0(x), \quad \text{in } \Omega, \tag{9.116}$$

$$\mathbf{n} \times \mathbf{u}_1(x,t) = 0, \quad \text{on } \partial\Omega \times (0,T), \tag{9.117}$$

where the operator L^* is defined as

$$L^*\mathbf{u}(x,t) = A^*\mathbf{u}(x,t) + \int_0^t \{F^* + G^*(t-s)\}\mathbf{u}(x,s)ds, \tag{9.118}$$

where the homogenized 6×6 matrices A^*, F^* and G^* are computed as follows:

$$A_k^* = \int_Q A(\xi)\{e_k + \nabla_\xi \overline{w}_k^A(\xi)\}d\xi, \tag{9.119}$$

$$F_k^* = \int_Q F(\xi)\{e_k + \nabla_\xi \overline{w}_k^A(\xi)\}d\xi, \tag{9.120}$$

$$\begin{aligned} G_k^*(t) = &\int_Q G(\xi,t)\{e_k + \nabla_\xi \overline{w}_k^A(\xi)\}d\xi + \int_Q A(\xi)\nabla_\xi \overline{w}_k(\xi,t)d\xi \\ &+ \int_Q \int_0^t \{F(\xi) + G(\xi,t-s)\}\nabla_\xi \overline{w}_k(\xi,s)dsd\xi, \end{aligned} \tag{9.121}$$

for $k = 1, 2\cdots, 6$, and A_k^*, F_k^* and G_k^* are the column vectors of A^*, F^* and G^*.

The extra source term $\mathbf{J}^0$ in (9.115) is given by

$$J_k^0(x,t) = u_k^0(x)\frac{d}{dt}\{\int_Q (A\nabla_\xi \overline{w}_k^0(t) + \int_0^t (F + G(t-s))\nabla_\xi \overline{w}_k^0(s)ds)d\xi\}, \tag{9.122}$$

for $k = 1, 2, \cdots, 6$. Here $u_k^0(x)$ are components of the decomposition $\mathbf{u}^0(x) = u_k^0(x)e_k$.

Similarly, by considering the decomposition $\mathbf{u}^*(x,t) = u_k^*(x,t)e_k$, we can obtain the corrector $\overline{\mathbf{u}} \in W^{2,1}(0,T; H^1_{per}(Q; R^2))$ as

$$\overline{\mathbf{u}}(x,\xi,t) = \overline{w}_k^A(\xi)u_k^*(x,t) + \int_0^t \overline{w}_k(\xi,t-s)u_k^*(x,s)ds + \overline{w}_k^0(\xi,t)u_k^0(x), \tag{9.123}$$

or in matrix form:

$$\overline{\mathbf{u}}(x,\xi,t) = \overline{w}^A(\xi)u^*(x,t) + \int_0^t \overline{w}(\xi,t-s)u^*(x,s)ds + \overline{w}^0(\xi,t)u^0(x),$$

where $\overline{w}^A \in R^{2\times 6}$ with columns $\overline{w}_k^A, k = 1, 2, \cdots, 6$. Similarly, $\overline{w}^0$ and $\overline{w} \in R^{2\times 6}$.

By solving corrector variational problems and the homogenized problem (9.115)–(9.117) using finite element methods on a regular mesh, we can obtain a quite accurate numerical solution to the original rapidly oscillatory coefficient problem (9.111)–(9.113).

9.5.2 A Posteriori Error Estimator

Due to the pioneering work of Babuska and Rheinboldt in the late 1970s [16], the adaptive finite element method has been well developed as evidenced by the vast literature in this area (cf. review papers [32, 64, 111, 126, 227], books [4, 20, 21, 252, 287, 297], and references cited therein). One important task in adaptive finite element method is to develop a robust and effective a posteriori error estimator, which can be used to guide where to refine or coarsen the mesh and/or how to choose the proper orders of the basis functions in different regions. As we mentioned in Sect. 6.1, though the study of a posteriori error estimator for elliptic, parabolic and second order hyperbolic problems seems mature, publications on a posteriori error estimator for Maxwell's equations are quite limited and are almost exclusively for the lowest-order edge element used for the model problem:

$$\nabla \times (\alpha \nabla \times \mathbf{u}) + \beta \mathbf{u} = \mathbf{f} \text{ in } \Omega \subset R^3, \text{ and } \mathbf{n} \times \mathbf{u} = \mathbf{0} \text{ on } \Gamma. \tag{9.124}$$

In 2009, Li [182] initiated analysis of a posteriori error estimator for time-dependent Maxwell's equations when cold plasma is involved. However, results of [182] are only for a semi-discrete scheme (cf. Sect. 6.3). Much more work is needed for edge elements with applications to Maxwell's equations when complex media such as metamaterials are involved.

Another area worth exploring is the hp-adaptive method [17,97,98,255] by varying both the mesh sizes and the orders of the basis functions. It is known that some pioneering works on hp methods have been initiated for time-harmonic Maxwell's equations (e.g., [3,97,98,273] and references cited therein). Extending them to time-domain Maxwell's equations involving metamaterials would be interesting but very challenging. Even for the free space case, the application of hp $H(curl)$ conforming finite element method to time-domain computational electromagnetics remain in its infancy [176, p. 295]. Furthermore, from a theoretical point of view, the hp finite element analysis for Maxwell's equations has just started (e.g. [91, p. 578] and [238]).

9.5.3 Concluding Remarks

The amount of literature and topics on metamaterials are so vast that our bibliography is by no means exhaustive. For example, there are increasingly more works on acoustic and elastic metamaterials, and acoustic and elastic cloaking (e.g.,[143, 214, 229, 233]). We decided to skip these subjects, since the underlying equations are totally different from the time-domain Maxwell's equations we focused on. Interested readers can refer to a very recent book edited by Craster and Guenneau [93], which is dedicated to this subject. We hope that the selected entries provide readers with a way to examine the covered topics more deeply.

Due to our limited experience in applications of metamaterials, we feel very sorry for missing any engineering and physics references. Readers can consult those published metamaterial books mentioned in Chap. 1.

Even from a mathematical modeling and scientific computing viewpoint, our book provides only an introduction to modeling wave propagation in metamaterials by using the so-called time-domain finite element methods. More robust and efficient numerical methods should be investigated in the future. To inspire more computational scientists (especially young researchers) to enter into this exciting area, below we summarize a list of interesting topics (at least to us) to be explored:

1. *Well-postedness and regularity* Though we investigated the well-postedness of some Maxwell's equations resulting from those popular metamaterial models, there are other models we haven't looked into yet. An important and challenging issue is how regular the solutions can be, since singularities can be caused by many things such as non-trivial physical domains, discontinuous material coefficients, and non-smooth source terms. Even for Maxwell's equations in free space, the analysis is quite involved (e.g., [89, 90]).
2. *Mass-lumping* For time-dependent large-scale simulations, it seems that explicit schemes such as leap-frog types are quite popular. However, inverting a mass matrix makes the algorithm not fully explicit. One way to overcome this issue is the so-called mass-lumping technique, which approximates the mass matrix by a diagonal matrix to speed up the simulation. Though some strategies have been proposed [86, 87, 108, 121], some issues remain to be resolved such as how to do mass-lumping for high-order edge elements, how to do mass-lumping on hybrid grids, and how mass-lumping affects the accuracy and dispersion error etc.
3. *Dispersion and dissipation analysis* The dispersive and dissipative errors play a very important role in wave propagation modeling. Though dispersion analysis has been carried out for Maxwell's equations in free space [87, 218, 279] and in dispersive media [25], no such analysis has been carried out for Maxwell's equations in metamaterials.
4. *Multiscale techniques* Since metamaterials are composites of periodic microstructures, it would be beneficial to develop some numerical methods coupled with multiscale techniques [105, 110, 293]. Some homogenization works on metamaterials have been carried out (e.g. [27, 47, 165, 231]), and much more work is needed in this direction.
5. *Nonconforming elements* Recent works [52, 103, 150, 242] show that it is possible to design convergent nonconforming finite element methods for solving time-harmonic Maxwell equations. Some comparisons with edge elements would be great, and applications of those nonconforming elements should be carried out to see if they can correctly simulate the wave propagation phenomena in metamaterials.
6. *Fast solvers* To improve the efficiency and robustness of linear system solvers, multigrid methods [136, 300] and domain decomposition methods [280] have been intensively investigated over the past three decades. Though there are

many publications on these subjects for edge elements used for time-harmonic Maxwell's equations (e.g. [5, 12, 129, 144, 146, 149, 281, 307]), very few papers exist for metamaterial Maxwell's equations. Recently, the so-called sweeping preconditioners proposed in [283] seem very efficient in solving time-harmonic Maxwell's equations with edge elements. Their numerical results with unstructured meshes (including a cloaking example) have demonstrated $O(N)$ complexity in 2-D and $O(N \log N)$ complexity in 3-D. It would be interesting to see how this algorithm performs for high order edge elements and time dependent problems.

7. *A posteriori error estimation* A posteriori error estimation plays a very important role in adaptive finite element methods. There is a huge amount of literature on a posteriori error estimation (cf. [4,20,21,252,287,297] and references cited therein). As we mentioned in Chap. 6, there are no more than 20 papers on a posteriori error estimation based on edge elements for Maxwell's equations. Additionally, most works are mainly on the lowest-order edge elements and just for Maxwell's equations in free space. Hence there is a great opportunity to obtain many interesting results for metamaterial models.
8. *Superconvergence* As mentioned in Chap. 5, many interesting results have been obtained for standard equations such as elliptic, parabolic and the second-order hyperbolic types (cf. [67, 201, 289]). But superconvergence results on edge elements for solving Maxwell's equations (especially when metamaterials are involved) are quite limited. So far, superconvergence has been proved and demonstrated for bilinear or trilinear edge elements [153, 198, 202, 215], and the lowest-order triangular edge elements formed as pairs of parallelograms [154]. It is still unknown wheather superconvergence exists for tetrahedral edge elements, or even higher-order triangular edge elements.
9. *Hp-adaptivity* The hp-adaptive finite element method can be thought as the most desirable method in that the mesh size and basis function order can be automatically adapted during a computer simulation. In this sense, adaptive DG methods can be put into the hp method family. It is known that the realization of an efficient hp method is very challenging. As we mentioned in Sect. 9.5.2, the hp finite element analysis and application for time-dependent Maxwell's equations still has many issues to be resolved.
10. *Frequency-domain analysis* In the book, we mainly focused on time-domain simulation of Maxwell's equations in metamaterials. It would be interesting to consider the metamaterial simulation in frequency-domain. Though many applications of frequency-domain finite element (FEFD) methods have been carried out by engineers, numerical analysis of FEFD methods seems quite limited (cf. Fernandes and Raffetto's works [117–119], and Bonnet-Ben Dhia et al. [43,44]). Further exploration in this direction should be done in the future.

References

1. Abarbanel, S., Gottlieb, D., Hesthaven, J.S.: Long time behavior of the perfectly matched layer equations in computational electromagnetics. J. Sci. Comput. **17**, 405–421 (2002)
2. Adams, R.A.: Sobolev Spaces. Academic, New York (1975)
3. Ainsworth, M., Coyle, J.: Hierarchic hp-edge element families for Maxwell's equations on hybrid quadrilateral/triangular meshes. Comput. Methods Appl. Mech. Eng. **190**, 6709–6733 (2001)
4. Ainsworth, M., Oden, J.T.: A Posteriori Error Estimation in Finite Element Analysis. Wiley-Interscience, New York (2000)
5. Alonso, A., Valli, A.: An optimal domain decomposition preconditioner for low-frequency time-harmonic Maxwell equations. Math. Comput. **68**, 607–631 (1999)
6. Alu, A., Engheta, N.: Achieving transparency with plasmonic and metamaterial coatings. Phys. Rev. E **72**, 016623 (2005)
7. Alu, A., Bilotti, E., Engheta, N., Vegni, L.: Sub-wavelength, compact, resonant patch antennas loaded with metamaterials. IEEE Trans. Antennas Propag. **AP-55**(1), 13–25 (2007)
8. Alu, A., Bilotti, E., Engheta, N., Vegni, L.: A conformal omni-directional sub-wavelength metamaterial leaky-wave antenna. IEEE Trans. Antennas Propag. **AP-55**(6), 1698–1708 (2007)
9. Ammari, H., Garnier, J., Jugnon, V., Kang, H., Lee, H., Lim, M.: Enhancement of near-cloaking. Part III: numerical simulations, statistical stability, and related questions. Contemporary Mathematics **577**, 1–24 (2012)
10. Appelo, D., Hagstrom, T., Kreiss, G.: Perfectly matched layers for hyperbolic systems: general formulation, well-posedness, and stability. SIAM J. Appl. Math. **67**, 1–23 (2006)
11. Amrouche, C., Bernardi, C., Dauge, M., Girault, V.: Vector potentials in three-dimensional non-smooth domains. Math. Methods Appl. Sci. **21**, 823–864 (1998)
12. Arnold, D.N., Falk, R.S., Winther, R.: Multigrid in H(div) and H(curl). Numer. Math. **85**, 175–195 (2000)
13. Arnold, D.N., Falk, R.S., Winther, R.: Finite element exterior calculus, homological techniques, and applications. Acta Numer. **15**, 1–155 (2006)
14. Avitzour, Y., Urzhumov, Y.A., Shvets, G.: Wide-angle infrared absorber based on negative index plasmonic metamaterial. Phys. Rev. B **79**, 045131 (2008)
15. Aydin, K., Bulu, I., Guven, K., Kafesaki, M., Soukoulis, C.M., Ozbay, E.: Investigation of magnetic resonances for different split-ring resonator parameters and designs. New J. Phys. **7**, 168 (2005)
16. Babuška, I., Rheinboldt, W.: Error estimates for adaptive finite element computations. SIAM J. Numer. Anal. **15**, 736–754 (1978)

J. Li and Y. Huang, *Time-Domain Finite Element Methods for Maxwell's Equations in Metamaterials*, Springer Series in Computational Mathematics 43,
DOI 10.1007/978-3-642-33789-5, © Springer-Verlag Berlin Heidelberg 2013

17. Babuška, I., Suri, M.: The p and $h-p$ versions of the finite element method, basic principles and properties. SIAM Rev. **36**, 578–632 (1994)
18. Baena, J.D., Jelinek, L., Marques, R., Mock, J.J., Gollub, J., Smith, D.R.: Isotropic frequency selective surfaces made of cubic resonators. Appl. Phys. Lett. **91**, 191105 (2007)
19. Banerjee, B.: An Introduction to Metamaterials and Waves in Composites. CRC, Boca Raton (2011)
20. Bangerth, W., Rannacher, R.: Adaptive Finite Element Methods for Solving Differential Equations. Birkhäuser, Basel (2003)
21. Bank, R.E.: PLTMG: A Software Package for Solving Elliptic Partial Differential Equations: Users' Guide 8.0. SIAM, Philadelphia (1998)
22. Bank, R.E., Xu, J.: Asymptotically exact a posteriori error estimators, part I: grids with superconvergence. SIAM J. Numer. Anal. **41**, 2294–2312 (2004)
23. Bank, R.E., Xu, J.: Asymptotically exact a posteriori error estimators, part II: general unstructured grids. SIAM J. Numer. Anal. **41**, 2313–2332 (2004)
24. Banks, H.T., Bokil, V.A., Cioranescu, D., Gibson, N.L., Griso, G., Miara, B.: Homogenization of periodically varying coefficients in electromagnetic materials. J. Sci. Comput. **28**, 191–221 (2006)
25. Banks, H.T., Bokil, V.A., Gibson, N.L.: Analysis of stability and dispersion in a finite element method for Debye and Lorentz media. Numer. Methods Partial Differ. Equ. **25**, 885–917 (2009)
26. Bao, G., Li, P., Wu, H.: An adaptive edge element with perfectly matched absorbing layers for wave scattering by biperiodic structures. Math. Comput. **79**, 1–34 (2010)
27. Barbatis, G., Stratis, I.G.: Homogenization of Maxwells equations in dissipative bianisotropic media. Math. Methods Appl. Sci. **26**, 1241–1253 (2003)
28. Barrett, R., Berry, M.W., Chan, T.F., Demmel, J., Donato, J., Dongarra, J., Eijkhout, V., Pozo, R., Romine, C., van der Vorst, H.: Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods. SIAM, Philadelphia (1993)
29. Becache, E., Joly, P.: On the analysis of Berenger's perfectly matched layers for Maxwell's equations. Math. Model. Numer. Anal. **36**, 87–119 (2002)
30. Becache, E., Petropoulos, P., Gedney, S.: On the long-time behavior of unsplit Perfectly Matched Layers. IEEE Trans. Antennas Propag. **54**, 1335–1342 (2004)
31. Beck, R., Hiptmair, R., Hoppe, R.H.W., Wohlmuth, B.: Residual based a posteriori error estimators for eddy current computation. M2AN Math. Model. Numer. Anal. **34**, 159–182 (2000)
32. Becker, R., Rannacher, R.: An optimal control approach to a posteriori error estimation in finite element methods. Acta Numer. **11**, 1–102 (2001)
33. Bensoussan, A., Lions, J.L., Papanicolaou, G.: Asymptotic Analysis for Periodic Structures. North-Holland, New York (1978)
34. Berenger, J.P.: A perfectly matched layer for the absorbing EM waves. J. Comput. Phys. **114**, 185–200 (1994)
35. Berenger, J.P.: Three-dimensional perfectly matched layer for the absorbtion of electromagnetic waves. J. Comput. Phys. **127**, 363–379 (1996)
36. Berenger, J.P.: Numerical reflection from FDTD-PMLs: a comparison of the split PML with the unsplit and CFS PMLs. IEEE Trans. Antennas Propag. **50**, 258–265 (2002)
37. Beruete, M., Falcone, F., Freire, M.J., Marques, R., Baena, J.D.: Electromagnetic waves in chains of complementary metamaterial elements. Appl. Phys. Lett. **88**, 083503 (2006)
38. Bilotti, E., Alu, A., Vegni, L.: Design of miniaturized patch antennas with μ-negative loading. IEEE Trans. Antennas Propag. **AP-56**(6), 1640–1647 (2008)
39. Bochev, P.B., Gunzburger, M.D.: Least-Squares Finite Element Methods. Springer, New York (2009)
40. Boffi, D., Fernandes, P., Gastaldi, L., Perudia, I.: Computational models of electromagnetic resonators: analysis of edge element approximation. SIAM J. Numer. Anal. **36**, 1264–1290 (1999)
41. Boffi, D., Costabel, M., Dauge, M., Demkowicz, L., Hiptmair, R.: Discrete compactness for the p-version of discrete differential forms. SIAM J. Numer. Anal. **49**, 135–158 (2011)

42. Bondeson, A., Rylander, T., Ingelstrom, P.: Computational Electromagnetics. Springer, New York (2010)
43. Bonnet-Ben Dhia, A.S., Ciarlet, P., Zwölf, C.M.: Two- and three-field formulations for wave transmission between media with opposite sign dielectric constants. J. Comput. Appl. Math. **204**, 408–417 (2007)
44. Bonnet-Ben Dhia, A.S., Ciarlet, P., Zwölf, C.M.: Time harmonic wave diffraction problems in materials with sign-shifting coefficients. J. Comput. Appl. Math. **234**, 1912–1919 (2010). Corrigendum **234**, 2616 (2010)
45. Bossavit, A.: Computational Electromagnetism. Academic, San Diego (1998)
46. Bossavit, A., Griso, G., Miara, B.: Modelling of periodic electromagnetic structures bianisotropic materials with memory effects. J. Math. Pures Appl. **84**, 819–850 (2005)
47. Bouchitté, G., Schweizer, B.: Homogenization of Maxwell's equations in a split ring geometry. Multiscale Model. Simul. **8**, 717–750 (2010)
48. Braess, D., Schöberl, J.: Equilibrated residual error estimator for edge elements. Math. Comput. **77**, 651–672 (2008)
49. Bramble, J.H., Pasciak, J.E.: Analysis of a finite element PML approximation for the three dimensional time-harmonic Maxwell problem. Math. Comput. **77**, 1–10 (2008)
50. Brandts, J.: Superconvergence of mixed finite element semi-discretization of two time-dependent problems. Appl. Math. **44**, 43–53 (1999)
51. Brenner, S.C., Scott, L.R.: The Mathematical Theory of Finite Element Methods. Springer, Berlin/Heidelberg (1994)
52. Brenner, S.C., Li, F., Sung, L.-Y.: A locally divergence-free nonconforming finite element method for the time-harmonic Maxwell equations. Math. Comput. **76**, 573–595 (2007)
53. Brezis, H., Functional Analysis, Sobolev Spaces and Partial Differential Equations. Springer, New York (2011)
54. Brezzi, F., Fortin, M.: Mixed and Hybrid Finite Element Methods. Springer, Berlin/Heidelberg (1991)
55. Buffa, A., Costabel, M., Schwab, C.: Boundary element methods for maxwell's equations on non-smooth domains. Numer. Math. **92**, 679–710 (2002)
56. Buffa, A., Hiptmair, R., von Petersdorff, T., Schwab, C.: Boundary element methods for maxwell's equations on Lipschitz domains. Numer. Math. **95**, 459–485 (2003)
57. Cai, W., Shalaev, V.: Optical Metamaterials: Fundamentals and Applications. Springer, New York (2009)
58. Caloz, C., Itoh, T.: Electromagnetic Metamaterials: Transmission Line Theory and Microwave Applications. Wiley, Hoboken (2005)
59. Canuto, C., Hussaini, M.Y., Quarteroni, A., Zang, T.A.: Spectral Methods: Fundamentals in Single Domains. Springer, New York (2010)
60. Cao, L., Zhang, Y., Allegretto, W., Lin, Y.: Multiscale asymptotic method for Maxwell's equations in composite materials. SIAM J. Numer. Anal. **47**, 4257–4289 (2010)
61. Capolino, F. (ed.): Metamaterials Handbook – Two Volume Slipcase Set: Theory and Phenomena of Metamaterials. CRC, Boca Raton (2009)
62. Carey, G.F., Oden, J.T.: Finite Elements: Computational Aspects. Prentice-Hall, Englewood Cliffs (1983)
63. Carstensen, C., Hu, J.: A unifying theory of a posteriori error control for nonconforming finite element methods. Numer. Math. **107**, 473–502 (2007)
64. Carstensen, C., Eigel, M., Löbhard, C., Hoppe, R.H.W.: A review of unified a posteriori finite element error control. IMA Preprint Series # 2338, University of Minnesota, Oct. 2010
65. Chen, Z.: Finite Element Methods and Their Applications. Springer, Berlin (2005)
66. Chen, H., Chen, M.: Flipping photons backward: reversed Cherenkov radiation. Materialstoday **14**, 34–41 (2011)
67. Chen, C.M., Huang, Y.: High Accuracy Theory of Finite Element Methods (in Chinese). Hunan Science Press, China (1995)
68. Chen, Q., Monk, P.: Introduction to applications of numerical analysis in time domain computational electromagnetism. In: Blowey, J., Jensen, M. (eds.) Frontiers in Numerical Analysis – Durham 2010, pp. 149–225. Springer, Berlin (2012)

69. Chen, Z., Wu, H.: An adaptive finite element method with perfectly matched layers for the wave scattering by periodic structures. SIAM J. Numer. Anal. **41**, 799–826 (2003)
70. Chen, M.-H., Cockburn, B., Reitich, F.: High-order RKDG methods for computational electromagnetics. J. Sci. Comput. **22**, 205–226 (2005)
71. Chen, X., Wu, B.-I., Kong, J.-A., Grzegorezyk, T.: Retrieval of the effective constitutive parameters of bianisotropic metamaterials. Phys. Rev. E **71**, 046610 (2005)
72. Chen, J., Xu, Y., Zou, J.: Convergence analysis of an adaptive edge element method for Maxwell's equations. Appl. Numer. Math. **59**, 2950–2969 (2009)
73. Chen, H., Chan, C.T., Sheng, P.: Transformation optics and metamaterials. Nat. Mater. **9**, 387–396 (2010)
74. Chew, W.C., Weedon, W.H.: A 3D perfectly matched medium from modified Maxwell's equations with stretched coordinates. Microw. Opt. Technol. Lett. **7**, 599–604 (1994).
75. Christiansen, S.H.: Foundations of finite element methods for wave equations of Maxwell type. In: Quak, E., Soomere, T. (eds.) Applied Wave Mathematics, pp. 335–393. Springer, Berlin (2009)
76. Chung, E.T., Engquist, B.: Convergene analysis of fully discrete finite volume methods for Maxwell's equations in nonhomogenous media. SIAM J. Numer. Anal. **43**, 303–317 (2005)
77. Chung, E.T., Du, Q., Zou, J.: Convergence analysis on a finite volume method for Maxwell's equations in non-homogeneous media. SIAM J. Numer. Anal. **41**, 37–63 (2003)
78. Ciarlet, P.G.: The Finite Element Method for Elliptic Problems. North-Holland, Amsterdam (1978)
79. Ciarlet, P. Jr., Zou, J.: Fully discrete finite element approaches for time-dependent Maxwell's equations. Numer. Math. **82**, 193–219 (1999)
80. Cioranescu, D., Damlamian, A., Griso, G.: Periodic unfolding and homogenization, C. R. Math. Acad. Sci. Paris **335**, 99–104 (2002)
81. Cioranescu, D., Donato, P.: An Introduction to Homogenization. Oxford University Press, Oxford (1999)
82. Cochez-Dhondt, S., Nicaise, S.: Robust a posteriori error estimation for the Maxwell equations. Comput. Methods Appl. Mech. Eng. **196**, 2583–2595 (2007)
83. Cockburn, B., Karniadakis, G.E., Shu, C.-W.: The development of discontinuous Galerkin methods. In: Cockburn, B., Karniadakis, G.E., Shu, C.-W. (eds.) Discontinuous Galerkin Methods: Theory, Computation and Applications, pp. 3–50. Springer, Berlin (2000)
84. Cockburn, B., Li, F., Shu, C.-W.: Locally divergence-free discontinuous Galerkin methods for the Maxwell equations. J. Comput. Phys. **194**, 588–610 (2004)
85. Cohen, G.C.: Higher-Order Numerical Methods for Transient Wave Equations. Springer, Berlin (2001)
86. Cohen, G.C., Monk, P.: Gauss point mass lumping schemes for Maxwell's equations. Numer. Methods Partial Diff. Equ. **14**, 63–88 (1998)
87. Cohen, G.C., Monk, P.: Mur-Nédélec finite element schemes for Maxwell's equations. Comput. Methods Appl. Mech. Eng. **169**, 197–217 (1999)
88. Correia, D., Jin, J.-M.: 3D-FDTD-PML analysis of left-handed metamaterials. Microw. Opt. Technol. Lett. **40**, 201–205 (2004)
89. Costabel, M., Dauge, M.: Singularities of electromagnetic fields in polyhedral domains. Arch. Ration. Mech. Anal. **151**(3), 221–276 (2000)
90. Costabel, M., Dauge, M., Nicaise, S.: Singularities of eddy current problems. M2AN Math. Model. Numer. Anal. **37**, 807–831 (2003)
91. Costabel, M., Dauge, M., Schwab, C.: Exponential convergence of hp-FEM for Maxwell equations with weighted regularization in polygonal domains. Math. Models Methods Appl. Sci. **15**, 575–622 (2005)
92. Coutts, T.J.: A review of progress in thermophotovoltaic generation of electricity. Renew. Sustain. Energy Rev. **3**, 77–184 (1999)
93. Craster, R.V., Guenneau, S. (eds.): Acoustic Metamaterials: Negative Refraction, Imaging, Lensing and Cloaking. Springer, New York (2013)

94. Cui, T.J., Smith, D., Liu, R. (eds.): Metamaterials: Theory, Design, and Applications. Springer, New York (2009)
95. Cummer, S.A.: Perfectly matched layer behavior in negative refractive index materials. IEEE Antennas Wirel. Propag. Lett. **3**, 172–175 (2004)
96. Cummer, S.A., Popa, B.-I., Schurig, D., Smith, D.R., Pendry, J.: Full-wave simulations of electromagnetic cloaking structures. Phys. Rev. E **74**, 036621 (2006)
97. Demkowicz, L.: Computing with hp-Adaptive Finite Elements I. One and Two-Dimensional Elliptic and Maxwell Problems. CRC, Boca Raton (2006)
98. Demkowicz, L., Kurtz, J., Pardo, D., Paszynski, M., Rachowicz, W., Zdunek, A.: Computing with Hp-Adaptive Finite Elements, Vol. 2: Frontiers: Three Dimensional Elliptic and Maxwell Problems with Applications. CRC, Boca Raton (2007)
99. Di Pietro, D.A., Ern, A.: Mathematical Aspects of Discontinuous Galerkin Methods. Springer, Berlin (2012)
100. Dolean, V., Fahs, H., Fezoui, L., Lanteri, S.: Locally implicit discontinuous Galerkin method for time domain electromagnetics. J. Comput. Phys. **229**, 512–526 (2010)
101. Dolling, G., Enkrich, C., Wegener, M., Soukoulis, C.M., Linden, S.: Simultaneous negative phase and group velocity of light in a metamaterial. Science **312**, 892–894 (2006)
102. Dong, X.T., Rao, X.S., Gan, Y.B., Guo, B., Yin, W.Y.: Perfectly matched layer-absorbing boundary condition for left-handed materials. IEEE Microw. Wirel. Compon. Lett. **14**, 301–303 (2004)
103. Douglas, J. Jr., Santos, J.E., Sheen, D.: A nonconforming mixed finite element method for Maxwell's equations. Math. Models Methods Appl. Sci. **10**, 593–613 (2000)
104. Duan, Z.Y., Wu, B.-I., Chen, H.-S., Xi, S., Chen, M.: Research progress in reversed Cherenkov radiation in double-negative metamaterials. Prog. Electromagn. Res. **90**, 75–87 (2009)
105. Efendiev, Y., Hou, T.Y.: Multiscale Finite Element Methods: Theory and Applications. Springer, New York (2009)
106. Eleftheriades, G.V., Balmain, K.G. (eds.): Negative Refraction Metamaterials: Fundamental Principles and Applications. Wiley, Hoboken (2005)
107. Elman, H.C., Silvester, D.J., Wathen, A.J.: Finite Elements and Fast Iterative Solvers with Applications in Incompressible Fluid Dynamics. Oxford University Press, Oxford (2005)
108. Elmkies, A., Joly, P.: Elements finis d'arete et condesation de masse pour les equations de Maxwell: le cas 3D. C. R. Acad. Sci. Paris Serie 1 **325**, 1217–1222 (1997)
109. Engheta, N., Ziolkowski, R.W. (eds.): Electromagnetic Metamaterials: Physics and Engineering Explorations. Wiley, Hoboken (2006)
110. Engquist, B., Runborg, O., Tsai, Y.-H.R. (eds.): Numerical Analysis of Multiscale Computations: Proceedings of a Winter Workshop at the Banff International Research Station 2009. Springer, New York (2011)
111. Eriksson, K., Estep, D., Hansbo, P., Johnson, C.: Introduction to adaptive methods for differential equations. Acta Numer. **4**, 105–158 (1995)
112. Ern, A., Guermond, J.-L.: Theory and Practice of Finite Elements. Springer, New York (2004)
113. Ewing, R.E., Lin, Y., Sun, T., Wang, J., Zhang, S.: Sharp L2-error estimates and superconvergence of mixed finite element methods for non-Fickian flows in porous media. SIAM J. Numer. Anal. **40**, 1538–1560 (2002)
114. Fairweather, G.: Finite Element Galerkin Methods for Differential Equations. Marcel Dekker, New York-Basel (1978)
115. Fang, J., Wu, Z.: Generalized perfectly matched layer for the absorption of propagating and evanescent waves in lossless and lossy media. IEEE Trans. Microw. Theory Tech. **44**, 2216–2222 (1996)
116. Fang, N., Lee, H., Sun, C., Zhang, X.: Sub-diffraction-limited optical imaging with a silver superlens. Science **308**, 534–537 (2005)
117. Fernandes, P., Raffetto, M.: Existence, uniqueness and finite element approximation of the solution of time-harmonic electromagnetic boundary value problems involving metamaterials. COMPEL **24**, 1450–1469 (2005)

118. Fernandes, P., Raffetto, M.: Well posedness and finite element approximability of time-harmonic electromagnetic boundary value problems involving bianisotropic materials and metamaterials. Math. Models Methods Appl. Sci. **19**, 2299–2335 (2009)
119. Fernandes, P., Raffetto, M.: Realistic and correct models of impressed sources for time-harmonic electromagnetic boundary value problems involving metamaterials. Preprint, Oct. 2011
120. Fezoui, L., Lanteri, S., Lohrengel, S., Piperno, S.: Convergence and stability of a discontinuous Galerkin time-domain methods for the 3D heterogeneous Maxwell equations on unstructured meshes. Model. Math. Anal. Numer. **39**(6), 1149–1176 (2005)
121. Fisher, A., Rieben, R.N., Rodrigue, G.H., White, D.A.: A generalized mass lumping technique for vector finite-element solutions of the time-dependent Maxwell equations. IEEE Trans. Antennas Propag. **53**(9), 2900–2910 (2005)
122. Frantzeskakis, D.J., Ioannidis, A., Roach, G.F., Stratis, I.G., Yannacopoulos, A.N.: On the error of the optical response approximation in chiral media. Appl. Anal. **82**, 839–856 (2003)
123. Galyamin, S.N., Tyukhtin, A.V.: Electromagnetic field of a moving charge in the presence of a left-handed medium. Phys. Rev. B **81**(23), 235134 (2010)
124. Gay-Balmaz, P., Martin, O.J.F.: Efficient isotropic magnetic resonators. Appl. Phys. Lett. **81**, 939–941 (2002)
125. Gedney, S.D.: An anisotropic PML absorbing medium for the FDTD simulation of fields in lossy and dispersive media. Electromagnetics **16**, 399–415 (1996)
126. Giles, M.B., Süli, E.: Adjoint methods for PDEs: a posteriori error analysis and postprocessing by duality. Acta Numer. **11**, 145–236 (2002)
127. Girault, V., Raviart, P.A.: Finite Element Methods for Navier-Stokes Equations – Theory and Algorithms. Springer, Berlin (1986)
128. Goodsell, G., Whiteman, J.R.: Superconvergence of recovered gradients of piecewise quadratic finite element approximations. Part II: L_∞-error estimates. Numer. Methods PDEs **7**, 85–99 (1991)
129. Gopalakrishnan, J., Pasciak, J.E., Demkowicz, L.F.: Analysis of a multigrid algorithm for time harmonic Maxwell equations. SIAM J. Numer. Anal. **42**, 90–108 (2004)
130. Greenleaf, A., Lassas, M., Uhlmann, G.: On non-uniqueness for Calderón's inverse problem. Math. Res. Lett. **10**, 685–693 (2003)
131. Greenleaf, A., Lassas, M., Uhlmann, G.: Anisotropic conductivities that cannot be detected by EIT. Physiol. Meas. **24**, 413–419 (2003)
132. Greenleaf, A., Kurylev, Y., Lassas, M., Uhlmann, G.: Cloaking devices, electromagnetics wormholes and transformation optics. SIAM Rev. **51**, 3–33 (2009)
133. Grote, M.J., Schneebeli, A., Schötzau, D.: Interior penalty discontinuous Galerkin method for Maxwell's equations: energy norm error estimates. J. Comput. Appl. Math. **204**, 375–386 (2007)
134. Grote, M.J., Schneebeli, A., Schötzau, D.: Interior penalty discontinuous Galerkin method for Maxwell's equations: optimal L^2-norm error estimates. IMA J. Numer. Anal. **28**, 440–468 (2008)
135. Guenneau, S., McPhedran, R.C., Enoch, S., Movchan, A.B., Farhat, M., Nicorovici, N.-A.P.: The colours of cloaks. J. Opt. **13**, 024014 (2011)
136. Hackbusch, W.: Multi-Grid Methods and Applications. Springer, New York (1985)
137. Hao, Y., Mittra, R.: FDTD Modeling of Metamaterials: Theory and Applications. Artech House Publishers, Boston (2008)
138. Harrington, R.F.: Field Computation by Moment Methods. Wiley-IEEE, Hoboken (1993)
139. Harutyunyan, D., Izsak, F., van der Vegt, J.J.W., Botchev, M.A.: Adaptive finite element techniques for the Maxwell equations using implicit a posteriori error estimates. Comput. Methods Appl. Mech. Eng. **197**, 1620–1638 (2008)
140. Hesthaven, J.S., Warburton, T.: High-order nodal methods on unstructured grids. I. Time-domain solution of Maxwell's equations. J. Comput. Phys. **181**, 186–221 (2002)
141. Hesthaven, J.S., Warburton, T.: Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications. Springer, New York (2008)

142. Hesthaven, J.S., Gottlieb, S., Gottlieb, D.: Spectral Methods for Time-Dependent Problems. Cambridge University Press, Cambridge (2007)
143. Hetmaniuk, U., Liu, H.Y., Uhlmann, G.: On three dimensional active acoustic cloaking devices and their simulation. Preprint, University of Washington (2009)
144. Hiptmair, R.: Multigrid method for Maxwells equations. SIAM J. Numer. Anal. **36**, 204–225 (1998)
145. Hiptmair, R.: Finite elements in computational electromagnetism. Acta Numer. **11**, 237–339 (2002)
146. Hiptmair, R., Xu, J.: Nodal auxiliary space preconditioning in H(curl) and H(div) spaces. SIAM J. Numer. Anal. **45**, 2483–2509 (2007)
147. Houston, P., Perugia, I., Schötzau, D.: Energy norm a posteriori error estimation for mixed discontinuous Galerkin approximations of the Maxwell operator. Comput. Methods Appl. Mech. Eng. **194**, 499–510 (2005)
148. Houston, P., Perugia, I., Schötzau, D.: An a posteriori error indicator for discontinuous Galerkin discretizations of H(curl)-elliptic partial differential equations. IMA J. Numer. Anal. **27**, 122–150 (2007)
149. Hu, Q., Zou, J.: A nonoverlapping domain decomposition method for Maxwell's equations in three dimensions. SIAM J. Numer. Anal. **41**, 1682–1708 (2003)
150. Huang, J., Zhang, S.: A divergence-free finite element method for a type of 3D Maxwell equations. Appl. Numer. Math. **62**, 802–813 (2012)
151. Huang, Y., Li, J.: Interior penalty discontinuous Galerkin method for Maxwell's equation in cold plasma. J. Sci. Comput. **41**, 321–340 (2009)
152. Huang, Y., Li, J.: Numerical analysis of a PML model for time-dependent Maxwell's equations. J. Comput. Appl. Math. **235**, 3932–3942 (2011)
153. Huang, Y., Li, J., Lin, Q.: Superconvergence analysis for time-dependent Maxwell's equations in metamaterials. Numer. Methods Partial Differ. Equ. **28**, 1794–1816 (2012)
154. Huang, Y., Li, J., Wu, C.: Averaging for superconvergence: verification and application of 2D edge elements to Maxwell's equations in metamaterials. Preprint, Oct. 2011
155. Huang, Y., Li, J., Yang, W.: Interior penalty DG methods for Maxwell's equations in dispersive media. J. Comput. Phys. **230**, 4559–4570 (2011)
156. Huang, Y., Li, J., Yang, W., Sun, S.: Superconvergence of mixed finite element approximations to 3-D Maxwell's equations in metamaterials. J. Comput. Phys. **230**, 8275–8289 (2011)
157. Huang, Y., Li, J., Yang, W.: Modeling backward wave propagation in metamaterials by a finite element time domain method. SIAM J. Sci. Comput. (in press)
158. Hughes, T.J.R.: Finite Element Method – Linear Static and Dynamic Finite Element Analysis. Prentice-Hall, Englewood Cliffs (1987)
159. Izsak, F., Harutyunyan, D., van der Vegt, J.J.W.: Implicit a posteriori error estimates for the Maxwell equations. Math. Comput. **77**, 1355–1386 (2008)
160. Jiao, D., Jin, J.-M.: Time-domain finite-element modeling of dispersive media. IEEE Microw. Wirel. Compon. Lett. **11**, 220–222 (2001)
161. Jikov, V.V., Kozlov, S.M., Oleinik, O.A.: Homogenization of Differential Operators and Integral Functionals. Springer, New York (1994)
162. Jin, J.: The Finite Element Method in Electromagnetics, 2nd edn. Wiley-IEEE, Hoboken (2002)
163. Johnson, C.: Numerical Solution of Partial Differential Equations by the Finite Element Method. Cambridge University Press, New York (1988)
164. Kafesaki, M., Koschny, Th., Penciu, R.S., Gundogdu, T.F., Economou, E.N., Soukoulis, C.M.: Left-handed metamaterials: detailed numerical studies of the transmission properties. J. Opt. A **7**, S12–S22 (2005)
165. Kohn, R.V., Shipman, S.P.: Magnetism and homogenization of microresonators. Multiscale Model. Simul. **7**, 62–92 (2008)
166. Kohn, R., Shen, H., Vogelius, M., Weinstein, M.: Cloaking via change of variables in electrical impendance tomography. Inverse Probl. **24**, 015016 (2008)

167. Kohn, R.V., Onofrei, D., Vogelius, M.S., Weinstein, M.I.: Cloaking via change of variables for the Helmholtz equation. Commun. Pure Appl. Math. **63**, 973–1016 (2010)
168. Kopriva, D.A., Woodruff, S.L., Hussaini, M.Y.: Computation of electromagnetic scattering with a non-conforming discontinuous spectral element method. Int. J. Numer. Mech. Eng. **53**, 105–122 (2002)
169. Kristensson, G.: Homogenization of the Maxwell equations in an anisotropic material. Technical Report LUTEDX/(TEAT-7104)/1–12/(2001), Department of Electroscience, Lund Institute of Technology, Sweden (2001)
170. Krizek, M., Neittaanmaki, P.: Bibliography on superconvergence. In: Krizek, M., Neittaanmaki, P., Stenberg, R. (eds.) Finite Element Methods: Superconvergence, Postprocessing and A Posteriori Estimates, pp. 315–348. Marcel Dekker, New York (1997)
171. Krowne, C.M., Zhang, Y. (eds.): Physics of Negative Refraction and Negative Index Materials: Optical and Electronic Aspects and Diversified Approaches. Springer, New York (2007)
172. Kunert, G., Nicaise, S.: Zienkiewicz-Zhu error estimators on anisotropic tetrahedral and triangular finite element meshes. ESAIM: Math. Model Numer. Anal. **37**, 1013–1043 (2003)
173. Kuzuoglu, M., Mittra, R.: Frequency dependence of the constitutive parameters of causal perfectly matched anisotropic absorbers. IEEE Microw. Guid. Wave Lett. **6**, 447–449 (1996)
174. Langtangen, H.P.: Computational Partial Differential Equations: Numerical Methods and Diffpack Programming, 2nd edn. Springer, Berlin (2003)
175. Laroche, M., Carminati, R., Greffet, J.-J.: Near-field thermophotovoltaic energy conversion. J. Appl. Phys. **100**, 063704 (2006)
176. Ledger, P.D., Morgan, K.: The application of the hp-finite element method to electromagnetic problems. Arch. Comput. Methods Eng. **12**, 235–302 (2005)
177. Lee, H.-J., Yook, J.-G.: Biosensing using split-ring resonators at microwave regime. Appl. Phys. Lett. **92**, 254103 (2008)
178. Lee, J.-F., Lee, R., Cangellaris, A.C.: Time domain finite element methods. IEEE Trans. Antennas Propag. **45**, 430–442 (1997)
179. Lee, J.-H., Xiao, T., Liu, Q.H.: A 3-D spectral-element method using mixed-order curl conforming vector basis functions for electromagnetic fields. IEEE Trans. Microw. Theory Tech. **54**, 437–444 (2006)
180. Leonhardt, U.: Optical conformal mapping. Science **312**, 1777–1780 (2006)
181. Leonhardt, U., Philbin, T.: Geometry and Light: The Science of Invisibility. Dover, New York (2010)
182. Li, J.: Posteriori error estimation for an interiori penalty discontinuous Galerkin method for Maxwell's equations in cold plasma. Adv. Appl. Math. Mech. **1**, 107–124 (2009)
183. Li, J.: Numerical convergence and physical fidelity analysis for Maxwell's equations in metamaterials. Comput. Methods Appl. Mech. Eng. **198**, 3161–3172 (2009)
184. Li, J.: Finite element study of the Lorentz model in metamaterials. Comput. Methods Appl. Mech. Eng. **200**, 626–637 (2011)
185. Li, J.: Development of discontinuous Galerkin methods for Maxwell's equations in metamaterials and perfectly matched layers. J. Comput. Appl. Math. **236**, 950–961 (2011)
186. Li, J.: Optimal L^2 error estimates for the interior penalty DG method for Maxwell's equations in cold plasma. Commun. Comput. Phys. **11**, 319–334 (2012)
187. Li, J., Chen, Y.: Computational Partial Differential Equations Using MATLAB. CRC, Boca Raton (2008)
188. Li, J., Huang, Y.: Mathematical simulation of cloaking metamaterial structures. Adv. Appl. Math. Mech. **4**, 93–101 (2012)
189. Li, J., Wood, A.: Finite element analysis for wave propagation in double negative metamaterials. J. Sci. Comput. **32**, 263–286 (2007)
190. Li, J., Zhang, Z.: Unified analysis of time domain mixed finite element methods for Maxwell's equations in dispersive media. J. Comput. Math. **28**, 693–710 (2010)
191. Li, J., Chen, Y., Elander, V.: Mathematical and numerical study of wave propagation in negative-index materials. Comput. Methods Appl. Mech. Eng. **197**, 3976–3987 (2008)

192. Li, J., Chen, Y., Liu, Y.: Mathematical simulation of metamaterial solar cells. Adv. Appl. Math. Mech. **3**, 702–715 (2011)
193. Li, J., Huang, Y., Lin, Y.: Developing finite element methods for Maxwell's equations in a Cole-Cole dispersive medium. SIAM J. Sci. Comput. **33**, 3153–3174 (2011)
194. Li, J., Huang, Y., Yang, W.: Developing a time-domain finite-element method for modeling of invisible cloaks. J. Comput. Phys. **231**, 2880–2891 (2012)
195. Li, J., Huang, Y., Yang, W.: Numerical study of the Plasma-Lorentz model in metamaterials. J. Sci. Comput. doi:10.1007/s10915-012-9608-5
196. Li, Z., Aydin, K., Ozbay, E.: Determination of the effective constitutive parameters of bianisotropic metamaterials from reflection and transmission coefficients. Phys. Rev. E **79**, 026610 (2009)
197. Liang, Z., Yao, P., Sun, X., Jiang, X.: The physical picture and the essential elements of the dynamical process for dispersive cloaking structures. Appl. Phys. Lett. **92**, 131118 (2008)
198. Lin, Q., Li, J.: Superconvergence analysis for Maxwell's equations in dispersive media. Math. Comput. **77**, 757–771 (2008)
199. Lin, Q., Lin, J.F.: High accuracy approximation of mixed finite element for 2-D Maxwell equations (in Chinese). Acta Math. Sci. Ser. A Chin. Ed. **23**, 499–503 (2003)
200. Lin, Q., Yan, N.: Superconvergence of mixed element methods for Maxwells equations (in Chinese). Gongcheng Shuxue Xuebao **13**, 1–10 (1996)
201. Lin, Q., Yan, N.: The Construction and Analysis of High Accurate Finite Element Methods (in Chinese). Hebei University Press, Hebei (1996)
202. Lin, Q., Yan, N.: Global superconvergence for Maxwells equations. Math. Comput. **69**, 159–176 (1999)
203. Lin, Q., Li, J., Zhou, A.: A rectangle test for the Stokes equations. In: Prof. of Sys. Sci. and Sys. Engrg., pp. 240–241. Culture Publish Co., Great Wall (H.K.) (1991)
204. Lin, Q., Yan, N., Zhou, A.: A rectangle test for interpolated finite elements. In: Prof. of Sys. Sci. and Sys. Engrg., pp. 217–229. Culture Publish Co., Great Wall (H.K.) (1991)
205. Liu, Z., Lee, H., Xiong, Y., Sun, C., Zhang, X.: Far-field optical hyperlens magnifying sub-diffraction-limited objects. Science **315**, 1686–1686 (2007)
206. Liu, R., Ji, C., Mock, J.J., Chin, J.Y., Cui, T.J., Smith, D.R.: Science **323**, 366–369 (2009)
207. Lu, T., Zhang, P., Cai, W.: Discontinuous Galerkin methods for dispersive and lossy Maxwell's equations and PML boundary conditions. J. Comput. Phys. **200**, 549–580 (2004)
208. Maradudin, A.A. (eds.): Stuctured Surfaces as Optical Metamaterials. Cambridge University Press, Cambridge (2011)
209. Markos, P., Soukoulis, C.M.: Wave Propagation: From Electrons to Photonic Crystals and Left-Handed Materials. Princeton University Press, Princeton (2008)
210. Marques, R., Martin, F., Sorolla, M.: Metamaterials with Negative Parameters: Theory, Design and Microwave Applications. Wiley-IEEE, New York (2008)
211. Milton, G.W.: The Theory of Composites. Cambridge University Press, Cambridge (2002)
212. Milton, G.W., Nicorovici, N.P.: On the cloaking effects associated with anomalous localized resonance. Proc. R. Soc. A **462**, 3027–3059 (2006)
213. Mittra, R., Pekel, U.: A new look at the perfectly matched layer (PML) concept for the reflectionless absorption of electromagnetic waves. IEEE Microw. Guid. Wave Lett. **53**, 84–86 (1995)
214. Milton, G.W., Briane, M., Willis, J.R.: On cloaking for elasticity and physical equations with a transformation invariant form. New J. Phys. **8**, 248 (2006)
215. Monk, P.: Superconvergence of finite element approximations to Maxwells equations. Numer. Methods Partial Differ. Equ. **10**, 793–812 (1994)
216. Monk, P.: A posteriori error indicators for Maxwell's equations. J. Comput. Appl. Math. **100**, 173–190 (1998)
217. Monk, P.: Finite Element Methods for Maxwell's Equations. Oxford Science Publications, New York (2003)
218. Monk, P., Parrott, A.K.: A dispersion analysis of finite element methods for Maxwell's equations. SIAM J. Sci. Comput. **15**, 916–937 (1994)

219. Montseny, E., Pernet, S., Ferriéres, X., Cohen, G.: Dissipative terms and local time-stepping improvements in a spatial high order Discontinuous Galerkin scheme for the time-domain Maxwell's equations. J. Comput. Phys. **227**, 6795–6820 (2008)
220. Munk, B.A.: Metamaterials: Critique and Alternatives. Wiley-Interscience, Hoboken (2009)
221. Narimanov, E.E., Shalaev, V.M.: Beyond diffraction. Nature **447**, 266–267 (2007)
222. Nédélec, J.-C.: Mixed finite elements in $\mathscr{R}^3$. Numer. Math. **35**, 315–341 (1980)
223. Nédélec, J.-C.: A new family of mixed finite elements in $\mathscr{R}^3$. Numer. Math. **50**, 57–81 (1986)
224. Nicaise, S.: On Zienkiewicz-Zhu error estimators for Maxwell's equations. C. R. Math. Acad. Sci. Paris **340**, 697–702 (2005)
225. Nicaise, S., Creusé, E.: A posteriori error estimation for the heterogeneous Maxwell equations on isotropic and anisotropic meshes. Calcolo **40**, 249–271 (2003)
226. Nicolaides, R.A., Wang, D.-Q.: Convergence analysis of a covolume scheme for Maxwell's equations in three dimensions. Math. Comput. **67**, 947–963 (1998)
227. Nochetto, R.H., Veeser, A.: Primer of adaptive finite element methods. In: Naldi, G., Russo, G. (eds.) Multiscale and Adaptivity: Modeling, Numerics and Applications: C.I.M.E. Summer School, Cetraro, Italy 2009, pp. 125–226. Springer, Berlin (2012)
228. Noginov, M.A., Podolskiy, V. (eds.): Tutorials in Metamaterials. Series in Nano-Optics and Nanophotonics. CRC, Boca Raton (2011)
229. Norris, A.N.: Acoustic cloaking theory. Proc. R. Soc. A. **464**, 2411–2434 (2008)
230. O'Hara, J.F., Singh, R., Brener, I., Smirnova, E., Han, J., Taylor, A.J., Zhang, W.: Thin-film sensing with planar terahertz metamaterials: sensitivity and limitations. Opt. Express **16**, 1786–1795 (2008)
231. Ouchetto, O., Zouhdi, S., Bossavit, A., Griso, G., Miara, B., Razek, A.: Homogenization of structured electromagnetic materials and metamaterials. J. Mater. Process. Technol. **181**, 225–229 (2007)
232. Padilla, W.J.: Group theoretical description of artificial electromagnetic metamaterials. Opt. Express **15**, 1639–1646 (2007)
233. Parnell, W.J.: Nonlinear pre-stress for cloaking from antiplane elastic waves. Proc. R. Soc. A **468**, 563–580 (2012)
234. Pendry, J.B.: Negative refraction makes a perfect lens. Phys. Rev. Lett. **85**, 3966–3969 (2000)
235. Pendry, J.B., Holden, A.J., Stewart, W.J., Youngs, I.: Extremely low frequency plasmons in metallic meso structures. Phys. Rev. Lett. **76**, 4773–4776 (1996)
236. Pendry, J.B., Holden, A.J., Robbins, D.J., Stewart, W.J.: Magnetism from conductors and enhanced nonlinear phenomena. IEEE Trans. Microw. Theory Tech. **47**, 2075–2084 (1999)
237. Pendry, J.B., Schurig, D., Smith, D.R.: Controlling electromagnetic fields. Science **312**, 1780–1782 (2006)
238. Pernet, S., Ferrieres, X.: HP A-priori error estimates for a non-dissipative spectral discontinuous Galerkin method to solve the Maxwell equations in the time domain. Math. Comput. **76**, 1801–1832 (2007)
239. Piperno, S., Remaki, M., Fezoui, L.: A non-diffusive finite volume scheme for the 3D Maxwell equations on unstructured meshes. SIAM J. Numer. Anal. **39**, 2089–2108 (2002)
240. Pozrikidis, C.: Introduction to Finite and Spectral Element Methods Using MATLAB. Chapman & Hall/CRC, Boca Raton (2005)
241. Prokopidis, K.P.: On the development of efficient FDTD-PML formulations for general dispersive media. Int. J. Numer. Model. **21**, 395–411 (2008)
242. Qiao, Z., Yao, C., Jia, S.: Superconvergence and extrapolation analysis of a nonconforming mixed finite element approximation for time-harmonic Maxwell's equations. J. Sci. Comput. **46**, 1–19 (2011)
243. Quarteroni, A., Valli, A.: Numerical Approximation of Partial Differential Equations. Springer, Berlin (1994)
244. Rahm, M., Schurig, D., Roberts, D.A., Cummer, S.A., Smith, D.R., Pendry, J.B.: Design of electromagnetic cloaks and concentrations using form-invariant coordinate transformations of Maxwell's equations. Photonics Nanostructures – Fundam. Appl. **6**, 87–95 (2008)

245. Ramakrishna, S.A., Grzegorczyk, T.M.: Physics and Applications of Negative Refractive Index Materials. CRC, Boca Raton (2008)
246. Rappaport, C.M.: Perfectly matched absorbing conditions based on anisotropic lossy mapping of space. IEEE Microw. Guid. Wave Lett. **53**, 90–92 (1995)
247. Riviere, B.: Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations: Theory and Implementation. SIAM, Philadelphia (2008)
248. Roden, J.A., Gedney, S.D.: Convolutional PML (CPML): an efficient FDTD implementation of the CFS-PML for arbitrary media. Microw. Opt. Technol. Lett. **27**, 334–339 (2000)
249. Sacks, Z.S., Kingsland, D.M., Lee, R., Lee, J.-F.: A perfectly matched anisotropic absorber for use as an absorbing boundary condition. IEEE Trans. Antennas Propag. **43**, 1460–1463 (1995)
250. Sanchez-Palencia, E.: Non-Homogeneous Media and Vibration Theory. Springer, Berlin (1980)
251. Scheid, C., Lanteri, S.: Convergence of a discontinuous Galerkin scheme for the mixed time domain Maxwell's equations in dispersive media, IMA J Numer Anal (2012). doi: 10.1093/imanum/drs008
252. Schmidt, A., Siebert. K.G.: Design of Adaptive Finite Element Software: The Finite Element Toolbox ALBERTA. Springer, Berlin (2005)
253. Schöberl, J.: A posteriori error estimates for Maxwell equations. Math. Comput. **77**, 633–649 (2008)
254. Schurig, D., Mock, J.J., Justice, B.J., Cummer, S.A., Pendry, J.B., Starr, A.F.S., Smith, D.R.: Metamaterial electromagnetic cloak at microwave frequencies. Science **314**, 977–980 (2006)
255. Schwab, C.: p- and hp- Finite Element Methods, Theory and Applications to Solid and Fluid Mechanics. Oxford University Press, New York (1998)
256. Shalaev, V.M., Sarychev, A.K.: Electrodynamics of Metamaterials. World Scientific, Hackensack (2007)
257. Shamonina, E., Solymar, L.: Properties of magnetically coupled metamaterial elements. J. Magn. Magn. Mater. **300**, 38–43 (2006)
258. Shaw, S.: Finite element approximation of Maxwell's equations with Debye memory. Adv. Numer. Anal. **2010**, Article ID 923832 (2010). doi:10.1155/2010/923832
259. Shelby, R.A., Smith, D.R., Nemat-Nasser, S.C., Schultz, S.: Microwave transmission through a two-dimensional, isotropic, left-handed metamaterial. Appl. Phys. Lett. **78**, 489–491 (2001)
260. Shelby, R.A., Smith, D.R., Schultz, S.: Experimental verification of a negative index of refraction. Science **292**, 489–491 (2001)
261. Shen, J., Tang, T., Wang, L.-L.: Spectral Methods: Algorithms, Analysis and Applications. Springer, New York (2011)
262. Shi, Y., Li, Y., Liang, C.H.: Perfectly matched layer absorbing boundary condition for truncating the boundary of the left-handed medium. Microw. Opt. Technol. Lett. **48**, 57–62 (2006)
263. Shvets, G., Tsukerman, I. (eds.): Plasmonics and Plasmonic Metamaterials: Analysis and Applications. World Scientific, Hackensack (2011)
264. Sihvola, A.H.: Electromagnetic Mixing Formulas and Applications. The Institute of Electrical Engineer, London (1999)
265. Silveirinha, M., Belov, P., Simovski, C.: Sub-wavelength imaging at infrared frequencies using an array of metallic nanorods. Phys. Rev. B **75**, 035108 (2007)
266. Silveirinha, M., Belov, P., Simovski, C.: Ultimate limit of resolution of subwavelength imaging devices formed by metallic rods. Opt. Lett. **33**, 1726–1728 (2008)
267. Silvester, P.P., Ferrari, R.L.: Finite Elements for Electrical Engineers, 3rd edn. Cambridge University Press, London (1996)
268. Sjöberg, D., Engström, C., Kristensson, G., Wall, D.J.N., Wellander, N.: A Floquet-Bloch decomposition of Maxwell's equations applied to homogenization. Multiscale Model. Simul. **4**, 149–171 (2005)
269. Smith, D.R., Kroll, N.: Negative refractive index in left-handed materials. Phys. Rev. Lett. **85**, 2933–2936 (2000)
270. Smith, D., Pendry, J.: Homogenization of metamaterials by field averaging. J. Opt. Soc. Am. B **23**, 391–403 (2006)

271. Smith, D.R., Padilla, W.J., Vier, D.C., Nemat-Nasser, S.C., Schultz, S.: Composite medium with simultaneously negative permeability and permittivity. Phys. Rev. Lett. **84**, 4184–4187 (2000)
272. Smolyaninov, I.I., Hung, Y.-J., Davis, C.C.: Magnifying superlens in the visible frequency range. Science **315**, 1699–1701 (2007)
273. Solin, P., Dubcova, L., Cerveny, J., Dolezel, I.: Adaptive hp-FEM with arbitrary-level hanging nodes for Maxwell's equations. Adv. Appl. Math. Mech. **2**, 518–532 (2010)
274. Solymar, L., Shamonina, E.: Waves in Metamaterials. Oxford University Press, Oxford (2009)
275. Syms, R.R.A., Shamonina, E., Kalinin, V., Solymar, L.: A theory of metamaterials based on periodically loaded transmission lines: interaction between magnetoinductive and electromagnetic waves. J. Appl. Phys. **97**, 064909 (2005)
276. Taflove, A., Hagness, S.C.: Computational Electrodynamics: The Finite-Difference Time-Domain Method, 2nd edn. Artech House Publishers, Boston (2000)
277. Teixeira, F.L.: Time-domain finite-difference and finite-element methods for Maxwell equations in complex media. IEEE Trans. Antennas Propag. **56**, 2150–2166 (2008)
278. Teixeira, F.L., Chew, W.C.: PML-FDTD in cylindrical and spherical coordinates. IEEE Microw. Guid. Wave Lett. **7**, 285–287 (1997)
279. Tobon, L., Chen, J., Liu, Q.H.: Spurious solutions in mixed finite element method for Maxwell's equations: dispersion analysis and new basis functions. J. Comput. Phys. **230**, 7300–7310 (2011)
280. Toselli, A., Widlund, O.: Domain Decomposition Methods: Theory and Algorithms. Springer Series in Computational Mathematics, vol. 34. Springer, New York (2004)
281. Toselli, A., Widlund, O., Wohlmuth, B.: A FETI preconditioner for two dimensional edge element approximations of Maxwell's equations on nonmatching grids. SIAM J. Sci. Comput. **23**, 92–108 (2001)
282. Trefethen, L.N.: Spectral Methods in MATLAB. SIAM, Philadephia (2001)
283. Tsuji, P., Engquist, B., Ying, L.: A sweeping preconditioner for time-harmonic Maxwell's equations with finite elements. J. Comput. Phys. **231**, 3770–3783 (2012)
284. Turkel, E., Yefet, A.: Absorbing PML boundary layers for wave-like equations. Appl. Numer. Math. **27**, 533–557 (1998)
285. Valentine, J., Zhang, S., Zentgraf, Th., Ulin-Avila, E., Genov, D.A., Bartal, G., Zhang, X.: Three-dimensional optical metamaterial with a negative refractive index. Nature **455**, 376–380 (2008)
286. Verfürth, R.: A posteriori error estimation and adaptive mesh-refinement techniques. J. Comput. Appl. Math. **50**, 67–83 (1994)
287. Verfürth, R.: A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques. Wiley, Teubner (1996)
288. Veselago, V.G.: Electrodynamics of substances with simultaneously negative values of sigma and mu. Sov. Phys. Usp. **10**, 509–514 (1968)
289. Wahlbin, L.B.: Superconvergence in Galerkin Finite Element Methods. Springer, Berlin (1995)
290. Wang, B., Xie, Z., Zhang, Z.: Error analysis of a discontinuous Galerkin method for Maxwell equations in dispersive media. J. Comput. Phys. **229**, 8552–8563 (2010)
291. Wellander, N.: Homogenization of the Maxwell equations. Case I. Linear theory. Appl. Math. **46**, 29–51 (2001)
292. Wheeler, M.F., Whiteman, J.R.: Superconvergence of recovered gradients of discrete time/piecewise linear Galerkin approximations for linear and nonlinear parabolic problems. Numer. Methods PDEs **10**, 271–294 (1994)
293. Weinan, E.: Principles of Multiscale Modeling. Cambridge University Press, Cambridge (2011)
294. Whitney, H.: Geometric Integration Theory. Princeton University Press, Princeton (1957)
295. Wu, C., Avitzour, Y., Shvets, G.: Ultra-thin, wide-angle perfect absorber for infrared frequencies. In: Noginov, M.A., Zheludev, N.I., Boardman, A.D., Engheta, N. (eds.) Metamaterials: Fundamentals and Applications, Proceedings of SPIE, vol. 7029, 70290W (2008)

296. Xu, J., Zhang, Z.: Analysis of recovery type a posteriori error estimators for mildly structured grids. Math. Comput. **73**, 1139–1152 (2003)
297. Yan, N.: Superconvergence Analysis and A Posteriori Error Estimation in Finite Element Methods. Science Press, Beijing (2008)
298. Yan, N., Zhou, A.: Gradient recovery type a posteriori error estimates for finite element approximations on irregular meshes. Comput. Methods Appl. Mech. Eng. **190**, 4289–4299 (2001)
299. Yee, K.S.: Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media. IEEE Trans. Antennas Propag. **14**, 302–307 (1966)
300. Yserentant, H.: Old and new convergence proofs for multigrid methods. Acta Numer. **2**, 285–326 (1993)
301. Zhang, S., Fan, W., Panoiu, N.C., Malloy, K.J., Osgood, R.M., Brueck, S.R.: Experimental demonstration of near-infrared negative-index metamaterials. Phys. Rev. Lett. **95**, 137404 (2005)
302. Zhang, Y., Cao, L.-Q., Wong, Y.-S.: Multiscale computations for 3D time-dependent Maxwell's equations in composite materials. SIAM J. Sci. Comput. **32**, 2560–2583 (2010)
303. Zhao, Y., Hao, Y.: Full-wave parallel dispersive finite-difference time-domain modeling of three-dimensional electromagnetic cloaking structures. J. Comput. Phys. **228**, 7300–7312 (2009)
304. Zhao, Y., Argyropoulos, C., Hao, Y.: Full-wave finite-difference time-domain simulation of electromagnetic cloaking structures. Opt. Express **16**, 6717–6730 (2008)
305. Zheng, W.Y., Chen, Z., Wang, L.: An adaptive finite element method for the $H - \psi$ formulation of time-dependent eddy current problems. Numer. Math. **103**, 667–689 (2006)
306. Zhong, L., Chen, L., Shu, S., Wittum, G., Xu, J.: Convergence and optimality of adaptive edge finite element methods for time-harmonic Maxwell equations. Math. Comput. **81**, 623–642 (2012)
307. Zhong, L., Shu, S., Wang, J., Xu, J.: Two-grid methods for time-harmonic Maxwell equations. Linear Algebra Appl. 2012, Early View. doi:10.1002/nla.1827
308. Zhou, A., Li, J.: The full approximation accuracy for the stream function-vorticity-pressure method. Numer. Math. **68**, 427–435 (1994)
309. Zienkiewicz, O.C., Zhu, J.Z.: A simple error estimator and adaptive procedure for practical engineering analysis. Internat. J. Numer. Methods Eng. **24**, 337–357 (1987)
310. Ziolkowski, R.W.: Maxwellian material based absorbing boundary conditions. Comput. Methods Appl. Mech. Eng. **169**, 237–262 (1999)
311. Ziolkowski, R.W.: Pulsed and CW Gaussian beam interactions with double negative metamaterial slabs. Opt. Express **11**, 662–681 (2003)
312. Ziolkowski, R.W., Erentok, A.: Metamaterial-based efficient electrically small antennas. IEEE Trans. Antennas Propag. **AP-54**(7), 2113–2130 (2006)
313. Ziolkowski, R.W., Heyman, E.: Wave propagation in media having negative permittivity and permeability. Phys. Rev. E **64**, 056625 (2001)
314. Zouhdi, S., Sihvola, A., Vinogradov, A.P. (eds.): Metamaterials and Plasmonics: Fundamentals, Modelling, Applications. Springer, Berlin (2009)

Index

J. Li and Y. Huang, *Time-Domain Finite Element Methods for Maxwell's Equations in Metamaterials*, Springer Series in Computational Mathematics 43,
DOI 10.1007/978-3-642-33789-5, © Springer-Verlag Berlin Heidelberg 2013