

NUMERICAL MATHEMATICS  
AND SCIENTIFIC COMPUTATION

# Finite Element Methods for Maxwell's Equations

PETER MONK



OXFORD SCIENCE PUBLICATIONS

# NUMERICAL MATHEMATICS AND SCIENTIFIC COMPUTATION

*Series Editors*

G. H. GOLUB Ch. SCHWAB  
W. A. LIGHT E. SÜLI

# NUMERICAL MATHEMATICS AND SCIENTIFIC COMPUTATION

\*P. Dierckx: *Curve and surface fittings with splines*

\*H. Wilkinson: *The algebraic eigenvalue problem*

\*I. Duff, A. Erisman, and J. Reid: *Direct methods for sparse matrices*

\*M. J. Baines: *Moving finite elements*

\*J. D. Pryce: *Numerical solution of Sturm–Liouville problems*

K. Burrage: *Parallel and sequential methods for ordinary differential equations*

Y. Censor and S. A. Zenios: *Parallel optimization: theory, algorithms, and applications*

M. Ainsworth, J. Levesley, M. Marletta, and W. Light: *Wavelets, multilevel methods, and elliptic PDEs*

W. Freeden, T. Gervens, and M. Schreiner: *Constructive approximation on the sphere: theory and applications to geomathematics*

Ch. Schwab: *p- and hp- finite element methods: theory and applications to solid and fluid mechanics*

J. W. Jerome: *Modelling and computation for applications in mathematics, science, and engineering*

Alfio Quarteroni and Alberto Valli: *Domain decomposition methods for partial differential equations*

G. E. Karniadakis and S. J. Sherwin: *Spectral/hp element methods for CFD*

I. Babuška and T. Strouboulis: *The finite element method and its reliability*

B. Mohammadi and O. Pironneau: *Applied shape optimization for fluids*

S. Succi: *The lattice Boltzmann equation for fluid dynamics and beyond*

P. Monk: *Finite element methods for Maxwell's equations*

A. Bellen and M. Zennaro: *Numerical methods for delay differential equations*

Monographs marked with an asterisk (\*) appeared in the series 'Monographs in Numerical Analysis' which has been folded into, and is continued by, the current series

# Finite Element Methods for Maxwell's Equations

Peter Monk

*Department of Mathematical Sciences*

*University of Delaware*

*Newark, USA*

CLARENDON PRESS • OXFORD  
2003

**OXFORD**  
UNIVERSITY PRESS

Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.  
It furthers the University's objective of excellence in research, scholarship,  
and education by publishing worldwide in  
Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi  
Kuala Lumpur Madrid Melbourne Mexico City Nairobi  
New Delhi Shanghai Taipei Toronto  
With offices in

Argentina Austria Brazil Chile Czech Republic France Greece  
Guatemala Hungary Italy Japan South Korea Poland Portugal

Singapore Switzerland Thailand Turkey Ukraine Vietnam  
Oxford is a registered trade mark of Oxford University Press  
in the UK and in certain other countries

Published in the United States  
by Oxford University Press Inc., New York  
© Oxford University Press 2003

The moral rights of the author have been asserted

Database right Oxford University Press (maker)  
First published 2003  
Reprinted 2004, 2006

All rights reserved. No part of this publication may be reproduced,  
stored in a retrieval system, or transmitted, in any form or by any means,  
without the prior permission in writing of Oxford University Press,  
or as expressly permitted by law, or under terms agreed with the appropriate  
reprographics rights organization. Enquiries concerning reproduction  
outside the scope of the above should be sent to the Rights Department,  
Oxford University Press, at the address above.

You must not circulate this book in any other binding or cover  
and you must impose this same condition on any acquirer.

British Library Cataloguing in Publication Data  
Data available

Library of Congress Cataloging in Publication Data  
ISBN 0 19 850888 3  
10 9 8 7 6 5 4 3

# PREFACE

In writing a book on the mathematical foundations of the finite element method for approximating Maxwell's equations I am well aware that I am on very dangerous ground. In his recent textbook *Functional Analysis*, Lax [202] says that “Two souls dwell in the bosom of scattering theory. One is mathematical and handles the unitary equivalence of operators with continuous spectra. The other is physics ...”. This quotation seems to me to describe scattering theory remarkably well, except that from the point of view of this book we need to substitute “electrical engineering” for physics. There is currently an enormous effort in the electrical engineering community to simulate electromagnetic phenomena using a variety of numerical methods including finite elements, which are the subject of this book. On the mathematical side there has recently been increased interest in the understanding of the mathematical properties of Maxwell's equations relevant to numerical analysis. The purpose of this book is to describe some of the basic mathematical theory of Maxwell's equations as it pertains to finite element methods, and hence to provide some mathematical underpinnings for the finite element method in this context. Along the way I shall try to point out some of the more obvious problems still remaining. Inevitably, such a book can be criticized on the grounds of being insufficiently mathematical or insufficiently practical (a more likely criticism), depending on the background of the reader — which brings us back to Lax's quotation!

The book is intended to be self-contained from the point of view of finite element theory. Therefore, there is a detailed discussion of convergence theory for mixed finite element methods, basic definitions of finite elements, and error estimates. However, it is much less detailed from the point of view of practical implementation — for this aspect of the finite element method there are already excellent sources in the electrical engineering literature including [177, 272]. Inevitably, it is necessary to assume some mathematics background for the book. Two subjects form the basis of the theory here: functional analysis and Sobolev space theory. For these topics, the excellent book of McLean [215] covers more than is necessary for this book. I have not assumed that the reader is familiar with Sobolev spaces of vector functions. Thus, in Chapter 3, I have summarized some more or less classical material on these spaces. The main source for this chapter is the book of Girault and Raviart [143]. This is a lovely book and well worth reading.

After the preparatory work in Chapters 2 (functional analysis and abstract error estimates) and 3 (Sobolev spaces and vector function spaces) we move on, in Chapter 4, to discuss a simple model problem for Maxwell's equations. This is a cavity or interior problem, which is posed on a bounded domain, but with

boundary conditions motivated by scattering applications (as first described in Chapter 1). This chapter uses the spaces from Chapter 3 to write down and analyze a standard variational formulation for the cavity problem. The analysis motivates the function spaces involved and the analytical techniques used to investigate such a problem.

At this stage we face a decision: what class of domains to allow for the scatterer. On the one hand, the theory of partial differential equations is much simplified if the domain has a smooth boundary. But this vastly complicates the discussion of finite element methods and the effects of the approximation of smooth boundaries is not well understood for Maxwell's equations. Therefore, I have decided to focus my discussion on Lipschitz polyhedra. These allow the use of standard tetrahedral meshes. In addition, some of the subtle problems related to approximating Maxwell's equations (such as the non-convergence of standard finite element methods in some cases [105]) appear in this situation. Finally, some of the most interesting recent advances in finite element theory and function space theory for Maxwell's equations has taken place in the context of Lipschitz polyhedral domains (see, e.g. [63, 106, 12]).

Using the discussion of Chapter 4 as motivation, we see that some special finite elements — the edge elements of Nédélec [233] — are particularly well suited to discretizing the Maxwell system. Therefore, in Chapters 5 and 6 we present a detailed description of these spaces, together with an associated scalar space for the electrostatic potential and other spaces needed to complete the theory. These chapters are a central part of the book and, besides presenting the original Nédélec finite element spaces, also emphasize some more recent viewpoints, including in particular the discrete de Rham diagram which summarizes the relationships between the relevant function spaces, their finite element discretizations and interpolation operators.

Having obtained a suitable variational formulation of the cavity problem and suitable finite element spaces, we then move to the finite element discretization of the cavity problem in Chapter 7. I present in detail two proofs of convergence for this method. To date, the first proof can only be applied in a special case, but has the advantages of simplicity and of providing a very clean result. In addition, this theory will be used later when we investigate the frequency dependence of the error in finite element methods in Chapter 13 and when discussing an overlapping Schwarz method for solving the associated matrix problem. The second proof uses the theory of collectively compact operators to prove convergence in a rather general case allowing spatially dependent electromagnetic parameters. Another proof, due to Hiptmair [164], is not included but a similar technique is used later in Chapter 10. A fourth proof, due to Boffi and Gastaldi [50], is also not included since it rests on the theory of eigenvalue problems, which are not an emphasis of this book (although we do provide some theory in Chapters 4 and 7). The three chapters, 4, 5 and 7, form the core of the book and could be useful in a graduate course on finite element methods. Together with some material from Chapter 13 and some from the engineering texts mentioned above, an entire course could be

constructed — and indeed this book is partially a result of such a course taught at the University of Delaware. These chapters contain the principal technical results used in all analyses of edge elements to date.

A central task of computational electromagnetism is the approximation of scattering problems. In these problems a known incident field (e.g. from a radar transmitter) interacts with an object (e.g. an aircraft) to create a scattered field. The approximation of this scattered field (or the total field) is the goal of the finite element method. In this book we shall only consider the case of a bounded scatterer (like an aircraft). This reflects my interests, but of course there are many very important applications of scattering from unbounded media. Examples include the classical problem of computing scattering from an infinite periodic structure (or diffraction grating) [25] or a periodic structure with defects [10]. Although we shall not be handling these problems here, the techniques presented also appear in the analysis of more complex problems. For example the theory of Chapter 10 has been used in the analysis of scattering from objects coated by thin layers [11]. Our presentation of scattering problems starts with classical scattering by a sphere in Chapter 9, where we derive the famous integral representation of the solution to Maxwell's equations called the Stratton–Chu formula. In addition, we derive classical series representations of the solution of Maxwell's equations. These are used in Chapter 10 to derive a semi-discrete method for the scattering problem utilizing the electromagnetic equivalent of the Dirichlet to Neumann map. A fully discrete domain-decomposed version of this algorithm is proposed and analyzed in Chapter 11. The methods in Chapters 10 and 11 have the disadvantage of needing a truncated domain with a spherical truncation boundary. Obviously, using this method, high aspect ratio scatterers would require a domain with a large volume and, hence, large computational cost. Therefore, in Chapter 12 we turn to a coupled integral equation and finite element method due to Hazard and Lenoir [159] and Cutzach and Hazard [111]. In this method the Stratton–Chu formula is used to represent the solution outside the scatterer and simultaneously the finite element method is also used on a truncated domain extending outside the scatterer. There is thus a region where both methods represent the solution. It has to be admitted that this overlapping scheme is not the standard one in widespread use. I prefer this method because it avoids computing singular integrals and provides the basis for an alternating Schwarz iterative scheme for solving the problem. Readers interested in the more standard approach should consult the book of Jin [177] and the paper of Hiptmair [163].

There are of course many more problems associated with the finite element discretization of Maxwell's equations than those discussed in Chapters 7–12. In particular, the matrix problem resulting from the discretization of the Maxwell system is indefinite (regardless of the frequency of the radiation). Thus, the solution of this linear system (which is large and sparse) presents a serious challenge. Indeed, an efficient solution of this linear system is perhaps the main challenge currently facing finite element analysis of scattering problems. We discuss this

problem in Chapter 13. This chapter also contains shorter discussions of a number of other practical aspects of the solution of Maxwell's equations. For example, we discuss the sensitivity of the error in the calculation to the frequency of the radiation and explain the need for a “sufficiently fine” grid compared to the wavelength of the radiation. We also consider *a posteriori* error estimation and the extraction of the far field pattern of the scattered wave from a knowledge of the near field. In addition, we examine the domain truncation problem further and, in particular, touch on the perfectly matched layer and infinite elements. These topics are much less well understood from the theoretical point of view than the error analysis presented earlier in the book.

The final chapter (Chapter 14) of the book hardly fits with the title, but since inverse problems are my main reason for studying scattering theory I cannot resist a brief introduction to inverse scattering. Besides its intrinsic interest, the chapter provides an example of the application of some of the analytical results derived earlier in the book.

There are a number of books that overlap to a greater or lesser extent with this work. The electrical engineering books of Jin [177] and Sylvester and Ferrari [272] provide much more detail on coding finite element methods and, of course, more details of engineering applications. Thus, they complement my book rather well, with the book of Jin being most relevant because it focuses on edge elements. From the point of view of scattering theory in a variational setting, the book of Cessenat [73] is very useful but does not deal with numerical methods or (in the main) Lipschitz domains. Similarly, the book of Colton and Kress [94], although a vital source for much of the basic material in this book, uses a function space setting different from the one used here. In addition, finite element methods are not tackled. Perhaps closest to this book is the book of the founding father of this area, Professor Nédélec [236]. However, the emphasis of Nédélec's book is different in that he does not focus on finite element methods. Finally, although not a book, the massive survey article of Hiptmair [164] deserves mention. This article covers much of the material in Chapters 4 – 7 but at a more sophisticated level using discrete differential forms. In the same way as the book of Jin complements my book from the point of view of implementation, so does Hiptmair's article complement my presentation of finite elements and cavity problems.

Some comment needs to be made about the bibliography and references. I have roughly 300 references and have tried very hard to reference basic papers in the field. One area where the references are somewhat scarce is to the practical engineering literature. This does not represent a lack of enthusiasm for that literature. In fact, the widespread and successful engineering use of finite element methods and the need to buttress this success with a theoretical understanding are the motivations for this book. Since most of the theoretical work on finite elements has taken place in the mathematics literature, such papers appear in a disproportionate way in the bibliography.

Inevitably, there is an enormous amount of interesting material left out of this book. In essence, the contents are a reflection of my own research interests. In

my defense, I can only quote Wittgenstein: “Whereof one cannot speak, thereof one must be silent” [297].

Of course I have tried to rid the book of as many typos as possible. But I am mindful that some bugs will have escaped detection. I plan to post any typos reported to me on the web page

<http://www.math.udel.edu/~monk/FEBook/index.html>.

In addition I will record there any interesting suggestions regarding arguments in the book (but I reserve the right to define what is “interesting”!).

Thanks are due to many people. My parents and the Falkland Island government gave me an excellent school education. My PhD adviser Rick Falk introduced me to finite elements, gave me tremendous encouragement as a graduate student, and even suggested the University of Delaware for postgraduate employment. In my professional life I have benefited tremendously from my collaboration and friendship with David Colton, who encouraged me to write this book. Outside the department, my family, and particularly my wife Ellen, have supported me and provided a wonderful antidote to depression and self-absorption. Particular thanks are also due to Pam Irwin, who cheerfully typed much of the book from my execrable notes, and to David Colton and Fioralba Cakoni who helped with the manuscript. Last, but by no means least, I would like to thank Dr Arje Nachman and the Air Force Office of Scientific Research for grant support which has made my research possible.

*Newark P.M.*

*August 2002*

*This page intentionally left blank*

# CONTENTS

|  |    |
|--|----|
| <b>1 Mathematical models of electromagnetism</b>               | 1  |
| 1.1 Introduction   | 1  |
| 1.2 Maxwell's equations  | 2  |
| 1.2.1 Constitutive equations for linear media                  | 5  |
| 1.2.2 Interface and boundary conditions                        | 7  |
| 1.3 Scattering problems and the radiation condition            | 9  |
| 1.4 Boundary value problems                                    | 12 |
| 1.4.1 Time-harmonic problem in a cavity                        | 12 |
| 1.4.2 Cavity resonator   | 13 |
| 1.4.3 Scattering from a bounded object                         | 13 |
| 1.4.4 Scattering from a buried object                          | 14 |
| <b>2 Functional analysis and abstract error estimates</b>      | 15 |
| 2.1 Introduction   | 15 |
| 2.2 Basic functional analysis and the Fredholm alternative     | 15 |
| 2.2.1 Hilbert space  | 15 |
| 2.2.2 Linear operators and duality                             | 18 |
| 2.2.3 Variational problems                                     | 19 |
| 2.2.4 Compactness and the Fredholm alternative                 | 22 |
| 2.2.5 Hilbert-Schmidt theory of eigenvalues                    | 24 |
| 2.3 Abstract finite element convergence theory                 | 25 |
| 2.3.1 Cea's lemma  | 25 |
| 2.3.2 Discrete mixed problems                                  | 26 |
| 2.3.3 Convergence of collectively compact operators            | 32 |
| 2.3.4 Eigenvalue estimates                                     | 35 |
| <b>3 Sobolev spaces, vector function spaces and regularity</b> | 36 |
| 3.1 Introduction   | 36 |
| 3.2 Standard Sobolev spaces                                    | 36 |
| 3.2.1 Trace spaces   | 42 |
| 3.3 Regularity results for elliptic equations                  | 45 |
| 3.4 Differential operators on a surface                        | 48 |
| 3.5 Vector functions with well-defined curl or divergence      | 49 |
| 3.5.1 Integral identities                                      | 50 |
| 3.5.2 Properties of $H(\text{div}; \Omega)$                    | 52 |
| 3.5.3 Properties of $H(\text{curl}; \Omega)$                   | 55 |
| 3.6 Scalar and vector potentials                               | 61 |
| 3.7 The Helmholtz decomposition                                | 65 |
| 3.8 A function space for the impedance problem                 | 69 |
| 3.9 Curl or divergence conserving transformations              | 77 |

|  |     |
|--|-----|
| <b>4 Variational theory for the cavity problem</b>         | 81  |
| 4.1 Introduction   | 81  |
| 4.2 Assumptions on the coefficients and data               | 83  |
| 4.3 The space $X$ and the nullspace of the curl            | 84  |
| 4.4 Helmholtz decomposition                                | 86  |
| 4.4.1 Compactness properties of $X_0$                      | 87  |
| 4.5 The variational problem as an operator equation        | 89  |
| 4.6 Uniqueness of the solution                             | 92  |
| 4.7 Cavity eigenvalues and resonances                      | 95  |
| <b>5 Finite elements on tetrahedra</b>                     | 99  |
| 5.1 Introduction   | 99  |
| 5.2 Introduction to finite elements                        | 101 |
| 5.2.1 Sets of polynomials                                  | 108 |
| 5.3 Meshes and affine maps                                 | 112 |
| 5.4 Divergence conforming elements                         | 118 |
| 5.5 The curl conforming edge elements of Nédélec           | 126 |
| 5.5.1 Linear edge element                                  | 139 |
| 5.5.2 Quadratic edge elements                              | 140 |
| 5.6 $H^1(\Omega)$ conforming finite elements               | 143 |
| 5.6.1 The Clément interpolant                              | 147 |
| 5.7 An $L^2(\Omega)$ conforming space                      | 149 |
| 5.8 Boundary spaces  | 150 |
| <b>6 Finite elements on hexahedra</b>                      | 155 |
| 6.1 Introduction   | 155 |
| 6.2 Divergence conforming elements on hexahedra            | 155 |
| 6.3 Curl conforming hexahedral elements                    | 158 |
| 6.4 $H^1(\Omega)$ conforming elements on hexahedra         | 162 |
| 6.5 An $L^2(\Omega)$ conforming space and a boundary space | 164 |
| <b>7 Finite element methods for the cavity problem</b>     | 166 |
| 7.1 Introduction   | 166 |
| 7.2 Error analysis via duality                             | 168 |
| 7.2.1 The discrete Helmholtz decomposition                 | 170 |
| 7.2.2 Preliminary error analysis                           | 171 |
| 7.2.3 Duality estimate                                     | 174 |
| 7.3 Error analysis via collective compactness              | 176 |
| 7.3.1 Pointwise convergence                                | 178 |
| 7.3.2 Collective compactness                               | 180 |
| 7.3.3 Numerical results for the cavity problem             | 188 |
| 7.4 The ellipticized Maxwell system                        | 189 |
| 7.4.1 Discrete ellipticized variational problem            | 191 |
| 7.5 The discrete eigenvalue problem                        | 195 |
| <b>8 Topics concerning finite elements</b>                 | 199 |
| 8.1 Introduction   | 199 |

|   |     |
|---|-----|
| 8.2 The second family of elements on tetrahedra       | 202 |
| 8.2.1 Divergence conforming element                   | 202 |
| 8.2.2 Curl conforming element                         | 205 |
| 8.2.3 Scalar functions and the de Rham diagram        | 209 |
| 8.3 Curved domains                                    | 209 |
| 8.3.1 Locally mapped tetrahedral meshes               | 210 |
| 8.3.2 Large-element fitting of domains                | 214 |
| 8.4 $hp$ finite elements                              | 217 |
| 8.4.1 $H^1(\Omega)$ conforming $hp$ element           | 218 |
| 8.4.2 $hp$ curl conforming elements                   | 219 |
| 8.4.3 $hp$ divergence conforming space                | 221 |
| 8.4.4 de Rham diagram for $hp$ elements               | 222 |
| <b>9 Classical scattering theory</b>                  | 225 |
| 9.1 Introduction                                      | 225 |
| 9.2 Basic integral identities                         | 225 |
| 9.3 Scattering by a sphere                            | 234 |
| 9.3.1 Spherical harmonics                             | 236 |
| 9.3.2 Spherical Bessel functions                      | 238 |
| 9.3.3 Series solution of the exterior Maxwell problem | 241 |
| 9.4 Electromagnetic Calderon operators                | 248 |
| 9.4.1 The electric-to-magnetic Calderon operator      | 249 |
| 9.4.2 The magnetic-to-electric Calderon operator      | 252 |
| 9.5 Scattering of a plane wave by a sphere            | 254 |
| 9.5.1 Uniqueness and Rellich's lemma                  | 254 |
| 9.5.2 Series solution                                 | 256 |
| <b>10 The scattering problem using Calderon maps</b>  | 261 |
| 10.1 Introduction                                     | 261 |
| 10.2 Reduction to a bounded domain                    | 262 |
| 10.3 Analysis of the reduced problem                  | 264 |
| 10.3.1 Extended Helmholtz decomposition               | 267 |
| 10.3.2 An operator equation on $X_0^\sim$             | 269 |
| 10.4 The discrete problem                             | 274 |
| <b>11 Scattering by a bounded inhomogeneity</b>       | 280 |
| 11.1 Introduction                                     | 280 |
| 11.2 Derivation of the domain-decomposed problem      | 281 |
| 11.3 The finite-dimensional problem                   | 289 |
| 11.4 Analysis of the interior finite element problem  | 290 |
| 11.5 Error estimates for the fully discrete problem   | 298 |
| <b>12 Scattering by a buried object</b>               | 302 |
| 12.1 Introduction                                     | 302 |
| 12.2 Homogeneous isotropic background                 | 303 |
| 12.2.1 Analysis of the scheme                         | 308 |
| 12.2.2 The fully discrete problem                     | 311 |

|   |     |
|---|-----|
| 12.2.3 Computational considerations                           | 314 |
| 12.3 Perfectly conducting half space                          | 315 |
| 12.4 Layered medium   | 318 |
| 12.4.1 Incident plane waves                                   | 318 |
| 12.4.2 The dyadic Green's function                            | 321 |
| 12.4.3 Reduction to a bounded domain                          | 328 |
| <b>13 Algorithmic development</b>                             | 332 |
| 13.1 Introduction   | 332 |
| 13.2 Solution of the linear system                            | 333 |
| 13.3 Phase error in finite element methods                    | 344 |
| 13.3.1 Wavenumber dependent error estimates                   | 345 |
| 13.3.2 Phase error in three dimensional edge elements         | 351 |
| 13.4 <i>A posteriori</i> error estimation                     | 355 |
| 13.4.1 A residual-based error estimator                       | 356 |
| 13.4.2 Numerical experiments                                  | 362 |
| 13.5 Absorbing boundary conditions                            | 364 |
| 13.5.1 Silver-Müller absorbing boundary condition             | 365 |
| 13.5.2 Infinite element method                                | 370 |
| 13.5.3 The perfectly matched layer                            | 375 |
| 13.6 Far field recovery                                       | 386 |
| <b>14 Inverse problems</b>                                    | 394 |
| 14.1 Introduction   | 394 |
| 14.2 The linear sampling method                               | 397 |
| 14.2.1 Implementing the LSM                                   | 399 |
| 14.2.2 Numerical results with the LSM                         | 405 |
| 14.3 Mathematical aspects of inverse scattering               | 409 |
| 14.3.1 Uniqueness for the inverse problem                     | 411 |
| 14.3.2 Herglotz wave functions                                | 414 |
| 14.3.3 The far field operators $\mathbf{F}$ and $\mathcal{B}$ | 417 |
| 14.3.4 Mathematical justification of the LSM                  | 422 |
| <b>Appendices</b>   |     |
| <b>A Coordinate systems</b>                                   | 425 |
| A.1 Cartesian coordinates                                     | 425 |
| A.2 Spherical coordinates                                     | 425 |
| <b>B Vector and differential identities</b>                   | 427 |
| B.1 Vector identities   | 427 |
| B.2 Differential identities                                   | 427 |
| B.3 Differential identities on a surface                      | 427 |
| <b>References</b>   | 428 |
| <b>Index</b>  | 446 |

# 1 MATHEMATICAL MODELS OF ELECTROMAGNETISM

## 1.1 Introduction

In 1873 Maxwell founded the modern theory of electromagnetism with the publication of his *Treatise on Electricity and Magnetism*, in which he formulated the equations that now bear his name. These equations consist of two pairs of coupled partial differential equations relating six fields, two of which model sources of electromagnetism. It turns out that these equations are not sufficient to uniquely determine the electromagnetic field and that additional constitutive equations are needed to model the way in which the fields interact with matter. There is considerable flexibility in the constitutive equations. Because of this, we need to carefully state the problems to be analyzed in this book, and we start this chapter by summarizing the classical Maxwell equations governing an electromagnetic field in a linear medium. We then reduce this system to its time-harmonic form by assuming propagation at a single frequency. The time-harmonic Maxwell system will be the focus of this book. Besides Maxwell's equations, it is also necessary to describe appropriate physical boundary conditions. These include radiation conditions that select the outgoing field relevant to scattering problems.

Once the basic boundary value problem is formulated, it is often expedient to reduce the full Maxwell system to a simpler system relevant to the physical problem at hand. For example, it is often reasonable to assume that the electromagnetic field is time invariant or static. This reduces Maxwell's equations to a potential problem. Simpler models can also be derived at long and short wavelengths. We do not consider any of these reduced models here. We shall be concerned with approximating the time-harmonic Maxwell system for linear media in the “resonance region”. By this we mean that the wavelength of the radiation is commensurate with the dimensions of features of the scatterer.

We end this chapter with a summary of the relevant boundary value problems from the point of view of this book. Our presentation, at this stage, is purely formal (we simply assume the existence of appropriate solutions) and follows the format of standard texts on electromagnetism, such as [274]. Later chapters will give a careful variational formulation of the equations in this chapter, followed by finite element methods.

First a word about notation: vectors are distinguished from scalars by the use of bold typeface (but this convention does not, in general, carry over to operators). Unless otherwise stated, vectors will all be three dimensional and either real (in  $\mathbf{R}^3$ ) or complex (in  $\mathbf{C}^3$ ). For example,  $\mathbf{x} \in \mathbf{R}^3$  denotes position

in three-space and has components  $x_1$ ,  $x_2$  and  $x_3$  ( $x = (x_1, x_2, x_3)^\top$  where  $\top$  denotes transpose). For two vectors  $a \in \mathbb{C}^N$  and  $b \in \mathbb{C}^N$  we define the dot product on  $\mathbb{C}^N$  by

$$a \cdot b = \sum_{j=1}^N a_j b_j.$$

The reason for not including complex conjugation in the dot product is that we will need to write down expressions like  $v \cdot E$ , where  $v$  is a real vector and  $E$  is complex. In this case we do not want to conjugate  $E$ . Later, when we start to write down variational formulations, it will be important to recall that the dot product does not have complex conjugation built in. If  $a \in \mathbb{C}^N$  we define the Euclidean norm of  $a$  by  $|a| = \sqrt{a \cdot \bar{a}}$ , where  $\bar{a} = (\bar{a}_1, \dots, \bar{a}_N)^\top$  and  $\bar{a}_j$  is the complex conjugate of  $a_j$ .

As usual in mathematics texts,  $i = \sqrt{-1}$ , and  $j$  is just an integer variable. In our error estimates we shall use a generic constant  $C$  everywhere different. Apart from this, I have tried to avoid using the same symbol for two quantities (at least on the same page!).

## 1.2 Maxwell's equations

The classical macroscopic electromagnetic field is described by four vector functions of position  $x \in \mathbb{R}^3$  and time  $t \in \mathbb{R}$  denoted by  $\boldsymbol{\epsilon}$ ,  $D$ ,  $H$  and  $B$ . The fundamental field vectors  $\boldsymbol{\epsilon}$  and  $H$  are called the electric and magnetic field intensities, respectively (we shall refer to them as the electric field and the magnetic field, respectively). The vector functions  $D$  and  $B$ , which will later be eliminated from the description of the electromagnetic field via suitable constitutive relations, are called the electric displacement and magnetic induction, respectively.

An electromagnetic field is created by a distribution of sources consisting of static electric charges and the directed flow of electric charge, which is called current. The distribution of charges is given by a scalar charge density function  $\rho$ , while currents are described by the vector current density function  $J$ . Maxwell's equations then state that the field variables and sources are related by the following equations which apply throughout the region of space in  $\mathbb{R}^3$  occupied by the electromagnetic field:(1.1a)

$$\frac{\partial \mathcal{B}}{\partial t} + \nabla \times \boldsymbol{\epsilon} = 0, \tag{1.1b}$$

$$\nabla \cdot D = \rho, \tag{1.1c}$$

$$\frac{\partial D}{\partial t} \nabla \times \mathcal{H} = -J, \tag{1.1d}$$

$$\nabla \cdot B = 0,$$

Equation (1.1a) is called Faraday's law and gives the effect of a changing magnetic field on the electric field. The divergence condition (1.1b) is Gauss's law and gives the effect of the charge density on the electric displacement. The next equation,

(1.1c), is Ampère's circuital law as modified by Maxwell. Finally, eqn (1.1d) expresses the fact that the magnetic induction  $\mathbf{B}$  is solenoidal. A table of SI units relevant to electromagnetism is given in Table 1.1.

The divergence conditions (1.1b) and (1.1d) are consequences of the fundamental field equations, (1.1a) and (1.1c), provided charge is conserved. Formally, this is shown by taking the divergence of (1.1a) and (1.1c) and recalling that  $\nabla \cdot (\nabla \times \mathbf{A}) = 0$  for any vector function  $\mathbf{A}$ . Hence

$$\nabla \cdot \frac{\partial \mathbf{B}}{\partial t} = 0 \text{ and } \nabla \cdot \frac{\partial \mathbf{D}}{\partial t} = -\nabla \cdot \mathcal{J}.$$

But if charge is conserved,  $\rho$  and  $\mathcal{J}$  are connected by the relation(1.2)

$$\nabla \cdot \mathcal{J} + \frac{\partial \rho}{\partial t} = 0,$$

and hence

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

Thus if (1.1b) and (1.1d) hold at one time, they hold for all time. However, the fact that (1.1b) and (1.1d) are consequences of (1.1a) and (1.1c) for the continuous electromagnetic field does not mean that these divergence conditions can be entirely ignored when designing a numerical scheme to discretize (1.1). A successful scheme must produce a numerical approximation that in some sense satisfies discrete analogs of (1.1b) and (1.1d).

Either by using the Fourier transform in time, or because we wish to analyze electromagnetic propagation at a single frequency (e.g. if the source currents and charges vary sinusoidally in time), the time-dependent problem (1.1) can be reduced to the time-harmonic Maxwell system. If the radiation has a temporal frequency  $\omega > 0$ , then the electromagnetic field is said to be time-harmonic, provided(1.3a)

$$\mathbf{E}(x, t) = \Re \left( \exp(-i\omega t) \hat{\mathbf{E}}(x) \right), \quad (1.3b)$$

$$\mathbf{D}(x, t) = \Re \left( \exp(-i\omega t) \hat{\mathbf{D}}(x) \right), \quad (1.3c)$$

$$\mathbf{H}(x, t) = \Re \left( \exp(-i\omega t) \hat{\mathbf{H}}(x) \right),$$

Table 1.1 *A table giving the SI units appropriate for electromagnetic quantities.*

| Quantity Units   | Quantity Units                                  |
|--|---|
| Electric field intensity $\mathbf{E}$ Vm <sup>-1</sup> | Magnetic field intensity H Am <sup>-1</sup>     |
| Electric displacement D Cm <sup>-2</sup>               | Magnetic induction B T                          |
| Electric current density J Am <sup>-2</sup>            | Electric charge density $\rho$ Cm <sup>-3</sup> |

(1.3d)

$$\mathcal{B}(x, t) = \Re(\exp(-i\omega t)\hat{B}(x)),$$

where  $i = \sqrt{-1}$ , and  $\Re(\cdot)$  denotes the real part of the expression in parentheses. Note that  $\hat{E}$  (and similarly other hat variables) are now complex-valued vector functions of position but not time. Some authors instead choose a time dependence of  $\exp(i\omega t)$ . Of course, the choice is arbitrary and, provided it is used consistently, produces no difficulties. Our choice is fairly standard in the mathematics literature.

For consistency we also need the current density and charge density to be time-harmonic, so we assume

$$\mathcal{J}(x, t) = \Re(\exp(-i\omega t)\hat{J}(x)),$$

$$\rho(x, t) = \Re(\exp(-i\omega t)\hat{\rho}(x)).$$

Substituting these relations into (1.1) leads to the time-harmonic Maxwell equations:(1.4a)

$$-i\omega\hat{B} + \nabla \times \hat{E} = 0,$$

$$\nabla \cdot \hat{D} = \hat{\rho}, \quad (1.4b)$$

$$-i\omega\hat{D} + \nabla \times \hat{H} = \hat{J}, \quad (1.4c)$$

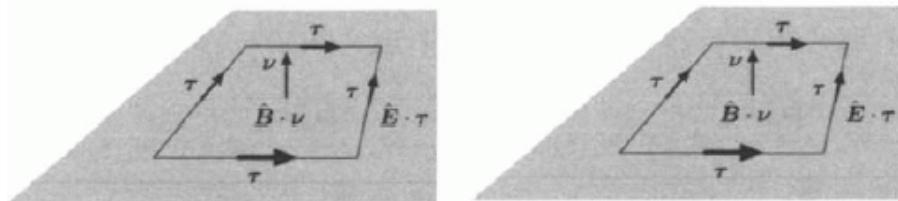
$$\nabla \cdot \hat{B} = 0, \quad (1.4d)$$

where the time-harmonic charge density  $\hat{\rho}$  is given via charge conservation (1.2) or by taking the divergence of (1.4c) and using (1.4b) as  $i\omega\hat{\rho} = \nabla \cdot \hat{J}$  and hence can be eliminated from the equations.

Equations (1.4) give the time-harmonic Maxwell equations in differential form. Frequently, particularly in the physics literature, they are stated in integral form. As an example, consider (1.4a) and let  $S$  be a smooth surface in  $\mathbb{R}^3$  with boundary  $\partial S$  and unit normal  $\nu$ . Then, using Stokes theorem, we find that(1.5)

$$\int_S \hat{B} \cdot \nu dA = \int_S (\nabla \times \hat{E}) \cdot \nu dA = \int_{\partial S} \hat{E} \cdot \tau ds,$$

Fig. 1.1. For a surface  $S$  with normal  $\nu$  the integrated flux of  $B$  normal to  $S$  is given by the integral of the tangential component of  $\hat{E}$  around the edges shown. Here we show schematics for a triangle and rectangle, two important surfaces from the point of view of numerical methods.



where  $\tau$  is the unit tangent to  $\partial S$  oriented by the right-hand rule relative to  $v$ . In the integral formulation we see that  $\hat{E}$  is naturally associated to line integrals, whereas  $B$  is naturally associated to surface integrals. For example, in Fig. 1.1 we show this when  $S$  is a triangle or a rectangle, two important cases that will appear later in the book.

Motivated by this integral formulation, finite difference schemes (in particular the famous FDTD scheme of Yee [301, 225]) usually associate the electric field  $\hat{E}$  with edges in a rectilinear mesh and the magnetic induction  $B$  with faces. This is also the arrangement of discrete unknowns in a generalization of the rectangular finite difference scheme to tetrahedral grids called the co-volume scheme [214, 240, 241]. As we shall see in Chapter 5, we can also design finite elements that have a similar arrangement of unknowns. Finally, we note that (1.5) is also a starting point for the description of Maxwell's equations in terms of differential forms [164].

### 1.2.1 Constitutive equations for linear media

Equations (1.4) must be augmented by two constitutive laws that relate  $\hat{E}$  and  $\hat{H}$  to  $D$  and  $B$ , respectively. These laws depend on the properties of the matter in the domain occupied by the electromagnetic field. We can distinguish three cases:

- (1) *Vacuum or free space* In free space the fields are related by the equations(1.6)

$$\hat{D} = \epsilon_0 \hat{E} \text{ and } \hat{B} = \mu_0 \hat{H},$$

where the constants  $\epsilon_0$  and  $\mu_0$  are called, respectively, the *electric permittivity* and *magnetic permeability*. The values of  $\epsilon_0$  and  $\mu_0$  depend on the system of units used. In the standard SI or MKS units

$$\begin{aligned} \mu_0 &= 4\pi \times 10^{-7} \text{ Hm}^{-1}, \\ \epsilon_0 &\approx 8.854 \times 10^{-12} \text{ Fm}^{-1}. \end{aligned}$$

Furthermore the speed of light in a vacuum, denoted by  $c$ , is given by  $c = \sqrt{\epsilon_0 \mu_0}^{-1} (c \approx 2.998 \times 10^8 \text{ ms}^{-1})$  [274] ..

- (2) *Inhomogeneous, isotropic materials* The most commonly occurring case in practice is that various different materials (e.g. copper, air, etc.) occupy the domain of the electromagnetic field. The medium is then called inhomogeneous. If the material properties do not depend on the direction of the field and the material is linear, we have(1.7)

$$\hat{D} = \epsilon \hat{E} \text{ and } \hat{B} = \mu \hat{H},$$

where  $\epsilon$  and  $\mu$  are positive, bounded, scalar functions of position (we shall give a more careful description of these functions in Section 4.2).

- (3) *Inhomogeneous, anisotropic materials* In some materials the electric or magnetic properties of the constituent materials depends on the direction of the

field (e.g. in the macroscopic description of a finely layered medium). In such cases  $\epsilon$  and  $\mu$  in (1.7) are  $3 \times 3$  positive-definite matrix functions of position. Usually, the finite element method is equally applicable to isotropic or anisotropic materials in that programs can be written from the onset for the anisotropic case. The theoretical justification of the convergence of the method is more difficult in these cases. Of course, in the presence of extreme anisotropy, special techniques may be necessary.

Although the methods in this book can be applied to anisotropic media, we will not analyze methods with matrix-valued coefficients. This is mainly due to the difficulty of verifying uniqueness of the solution of Maxwell's equations in this case. Although uniqueness is known (see [287]), the proof is too complex for this book.

One further constitutive relation needs to be discussed. In a conducting material, the electromagnetic field itself gives rise to currents. If the field strengths are not large, we can assume that Ohms law holds so that:(1.8)

$$\hat{J} = \sigma \hat{E} + \hat{J}_a,$$

where  $\sigma$  is called the *conductivity* and is a non-negative function of position. The vector function  $\hat{J}_a$  describes the applied current density. Regions where  $\sigma$  is positive are called *conductors*. Where  $\sigma = 0$  and  $\epsilon \neq \epsilon_0$ , the material is termed a *dielectric*, and  $\epsilon$  is referred to as the *dielectric constant*. In a vacuum (or air at low field strengths)  $\sigma = 0$ ,  $\epsilon = \epsilon_0$  and  $\mu = \mu_0$ . More generally, in anisotropic media, the conductivity  $\sigma$  can be a symmetric, positive semi-definite matrix function of position. However, we shall not consider this case here.

Using the linear, inhomogeneous constitutive equations in (1.7) and the constitutive relation for the currents in (1.8), we arrive at the following time-harmonic Maxwell system:(1.9a)

$$-i\omega\mu\hat{H} + \nabla \times \hat{E} = 0,$$

$$\nabla \cdot (\epsilon \hat{E}) = \frac{1}{i\omega} \nabla \cdot (\sigma \hat{E} + \hat{J}_a), \quad (1.9b)$$

$$-i\omega\mu\hat{E} + \sigma\hat{E} - \nabla \times \hat{H} = -\hat{J}_a, \quad (1.9c)$$

$$\nabla \cdot (\mu\hat{H}) = 0, \quad (1.9d)$$

where we recall that  $\hat{J}_a$  denotes a given applied current density.

There is one last reduction to perform on the equations. It is convenient to work with relative parameter values. Following Colton and Kress [93], we define

$$E = \epsilon_0^{1/2} \hat{E} \text{ and } H = \mu_0^{1/2} \hat{H}.$$

Using these definitions in (1.9), and defining the relative permittivity and permeability by

$$\epsilon_r = \frac{1}{\epsilon_0} \left( \epsilon + \frac{i\sigma}{\omega} \right) \text{ and } \mu_r = \frac{\mu}{\mu_0},$$

we obtain the final version of the first-order Maxwell system, where we note that  $\epsilon_r = \mu_r = 1$  in vacuum:(1.10a)

$$\begin{aligned} -ik\mu_r H + \nabla \times E &= 0, \\ -ik\epsilon_r E - \nabla \times H &= -\frac{1}{ik} F, \end{aligned} \quad (1.10b)$$

where  $F = ik\mu_0^{1/2} \hat{J}_a$  and the wavenumber  $k = \omega \sqrt{\epsilon_0 \mu_0}$ . We also obtain the divergence conditions (which follow from the differential equations when  $\kappa > 0$ )(1.11a)

$$\begin{aligned} \nabla \cdot (\epsilon_r E) &= -\frac{1}{k^2} \nabla \cdot F, \\ \nabla \cdot (\mu_r H) &= 0. \end{aligned} \quad (1.11b)$$

Although it is possible to derive numerical schemes for the first-order system (1.10)–(1.11), it is more usual to eliminate the magnetic field  $H$  by solving (1.10a) for  $H$  and substituting into (1.10b) to obtain the second-order Maxwell system(1.12)

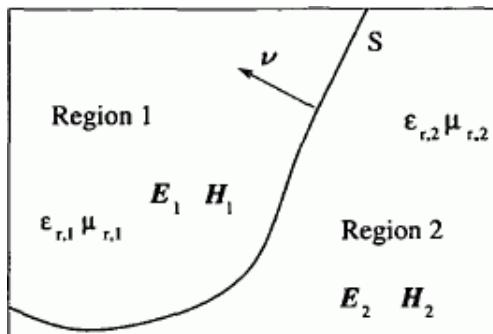
$$\nabla \times (\mu_r^{-1} \nabla \times E) - k^2 \epsilon_r E = F$$

together with (1.11a). Of course, the choice of eliminating  $H$ , rather than  $E$ , is arbitrary. We shall generally use (1.12) in this book, rather than the first-order system, since there are fewer dependent variables.

## 1.2.2 Interface and boundary conditions

Equations (1.10) or (1.12) are not a complete classical description of the electromagnetic field since the equations do not hold at boundaries between different materials where either  $\mu_r$  or  $\epsilon_r$  are discontinuous (e.g. at a copper–air interface). Let us consider the case of two media with differing electric and magnetic properties separated by a surface  $S$  with unit normal  $\nu$  pointing from region 2 to region 1 (see Fig. 1.2.). As we shall see later in Lemma 5.3, for  $\nabla \times E$  in (1.12) to be well defined in a least-squares sense we must have the tangential component

Fig. 1.2. Geometry of the surface and subdomains in our discussion of interface boundary conditions.



of the electric field to be continuous across  $S$  and so  $\mathbf{v} \times \mathbf{E}$  is continuous across  $S$ . Thus if  $E_1$  denotes the limiting value of the electric field as  $S$  is approached from region 1 and  $E_2$  denotes the limit of the field from the other region, we must have(1.13)

$$\mathbf{v} \times (E_1 - E_2) = 0 \text{ on } S.$$

On the other hand, we shall see (Lemma 5.3 again) that for  $\mu_r H$  in (1.11b) to have a well defined divergence in the least-squares sense, the normal components of  $\mu_r H$  must be continuous across  $S$  so that(1.14)

$$\mathbf{v} \cdot (\mu_{r,1} H_1 - \mu_{r,2} H_2) = 0 \text{ on } S,$$

where, again, the subscripts denote limiting values of the coefficients and field variables on either side of the surface  $S$ . The continuity conditions (1.13) and (1.14) hold for any electromagnetic field. However, we cannot assume that the analogue of (1.13) holds for the magnetic field. In general

$$\mathbf{v} \times (H_1 - H_2) = 0 \text{ on } J_S,$$

where this relation defines the tangential vector field  $J_s$  termed the surface current density on  $S$ . In most instances the magnetic field has continuous tangential components (i.e.  $J_s = 0$ ). This is true unless the surface  $S$  models a thin conductive layer giving rise to the conductive boundary condition (see [15]) or singularities in  $F$  give rise to surface currents on  $S$ . Thus we will also usually assume that(1.15)

$$\mathbf{v} \times (H_1 - H_2) = 0 \text{ on } S.$$

The presence of singularities in the charge density  $\rho$  may cause jumps in the normal component of  $\epsilon_r E$ . We write(1.16)

$$\mathbf{v} \cdot (\epsilon_{r,1} E_1 - \epsilon_{r,2} E_2) = \rho s \text{ on } S,$$

where  $\rho s$  is termed the surface charge density. From (1.14) and (1.16), even in the case of a negligible surface charge and current density, we see that the electric and magnetic field vectors are not continuous if  $\epsilon_r$  or  $\mu_r$  are discontinuous across  $S$ . Any numerical scheme for approximating Maxwell's equations in the presence of material discontinuities must take into account that tangential components of the field are continuous, but that normal components jump across a material boundary. As we shall see, the variational or weak formulation of Maxwell's equations used in this book automatically takes care of these jump conditions.

A particularly important case occurs when the material on one side of the interface discussed above is a perfect conductor. From Ohm's law (1.8), we see heuristically that if the conductivity  $\sigma \rightarrow \infty$  and if the current density  $J$  is to remain bounded then  $E \rightarrow 0$ . This suggests that in a perfect conductor the electric field vanishes. If the side of the surface  $S$  labeled 2 in Fig. 1.2 is a

perfect conductor then  $\mathbf{E}_2 = 0$  in (1.13) and we arrive at the perfect conducting boundary condition for  $\mathbf{E}_i$ ,(1.17)

$$\mathbf{v} \times \mathbf{E}_1 = 0 \text{ on } S,$$

where we can drop the index 1 since only the field outside the perfect conductor needs to be modeled.

If the material on one side of the boundary is not a perfect conductor, but allows the field to penetrate only a small distance, a more appropriate boundary condition is the impedance or imperfectly conducting boundary condition. Suppose again that the good conductor is in region 1 and that the normal  $\mathbf{v}$  points from region 2 into region 1. Then this boundary condition is(1.18)

$$\mathbf{v} \times \mathbf{H}_1 - \lambda(\mathbf{v} \times \mathbf{E}_1) \times \mathbf{v} = 0,$$

where the impedance  $\lambda$  is a positive function of position on the surface of the material.

### 1.3 Scattering problems and the radiation condition

So far we have not been specific about the region in space occupied by the electromagnetic field. The first case we shall discuss is scattering from a bounded, inhomogeneous object (e.g. radar scattering from an aircraft). We assume that the object consists of a bounded perfect conductor occupying a domain  $D$ , perhaps surrounded by an inhomogeneous medium where  $\epsilon_r \neq 1$  or  $\mu_r \neq 1$ . The electromagnetic field occupies the domain  $\mathbb{R}^3 \setminus D^-$ . We assume that sufficiently far from  $D$  the object is surrounded in all directions by air (or vacuum), so there is a radius  $a$  such that  $\epsilon_r(x) = \mu_r(x) = 1$  when  $|x| > a$  (see Fig. 1.3 ).

On the boundary of  $D$ , denoted by  $\Gamma$ , we impose the perfect conducting boundary condition. It turns out to be necessary to impose another boundary condition “at infinity” in order to obtain a well-posed problem. To do this we need to distinguish a given incident field (perhaps due to a radar or other electromagnetic source) and the resulting scattered field. The incident field is denoted by  $\mathbf{E}^i$  and is assumed to satisfy the Maxwell system in the absence of the scatterer (in the background medium) so that, in this case,(1.19)

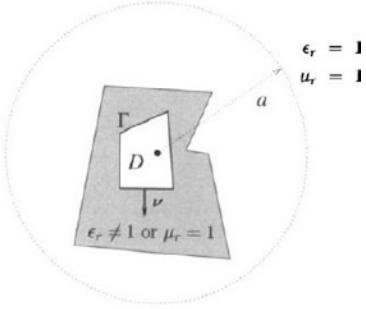
$$\nabla \times \nabla \times \mathbf{E}^i - k^2 \mathbf{E}^i = \mathbf{F} \text{ in } \mathbb{R}^3,$$

where  $\mathbf{F}$  is a given function describing the current source. A typical example might be the *plane wave* given by(1.20)

$$\mathbf{E}^i = p \exp(i k \mathbf{x} \cdot \mathbf{d}),$$

where  $\mathbf{d} \in \mathbb{R}^3$  is a unit vector giving the *direction of propagation* of the wave, and the vector  $p \neq 0$  is called the *polarization* and must be orthogonal to the

Fig. 1.3. Geometry of the scatterer and boundaries for the scattering problem. A bounded scatterer consisting of a perfectly conducting part and a penetrable part where the electromagnetic properties differ from the background is surrounded by air or vacuum.



direction of propagation so  $p \cdot d = 0$ . In this case  $F = 0$ . The total field  $E$  consists of the incident field  $E^i$  and the scattered field  $E^s$ , so(1.21)

$$E = E^i + E^s.$$

The scattered field is out-going (i.e. originates at the scatterer and propagates outwards) and this is imposed by requiring the scattered field to satisfy the Silver–Müller radiation condition [228]:(1.22)

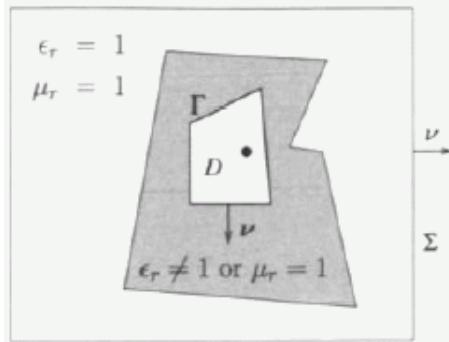
$$\lim_{\rho \rightarrow \infty} \rho \left( (\nabla \times E^s) \times \hat{x} - ik E^s \right) = 0,$$

where  $\rho = |x|$  and the limit is uniform in  $\phi = x/|x|$ .

The wavelength of the incident field in (1.20) is  $2\pi/\kappa$  since  $|d| = 1$ . If this wavelength is much smaller than a typical length  $b$  of relevant features of the scatterer (i.e. if  $\kappa b$  is large) or if this wavelength is much larger than  $b$  (i.e. if  $\kappa b$  is close to zero), it is possible to apply asymptotic methods to simplify the scattering problem. For example, when  $\kappa b$  is small, a popular approximation is the eddy-current model [272, 165]. For large  $\kappa b$  one can use the geometric theory of diffraction [181]. In this book we shall be concerned with computations in the “resonance region”, where  $\kappa b = O(1)$ , so asymptotic methods are not applicable.

An obvious difficulty with approximating the scattering problem by a finite element method is that the problem is posed on an infinite domain. One simple way to avoid this difficulty is to approximate the scattering problem by imposing the radiation condition (1.22) on a surface  $\Sigma$  far from the scatterer (where  $\mu_r = 1$  and  $\epsilon_r = 1$ ). Thus, in this approximation, the domain occupied by the computational electromagnetic field, denoted by  $\Omega$ , is the region between  $\Gamma$  and  $\Sigma$ , which

Fig. 1.4. Geometry of the scatterer and boundaries for the interior problem with an absorbing boundary condition on the auxiliary boundary  $\Sigma$ .



are assumed to be disjoint surfaces (see Fig. 1.4), and Maxwell's equations are satisfied in  $\Omega$ . On  $\Gamma$  we have the perfect conducting boundary condition, but on  $\Sigma$  we impose a boundary condition inspired by the Silver–Müller condition:(1.23)

$$(\nabla \times E) \times \nu - ikE_T = (\nabla \times E^i) \times \nu - ikE_T^i \text{ on } \Sigma,$$

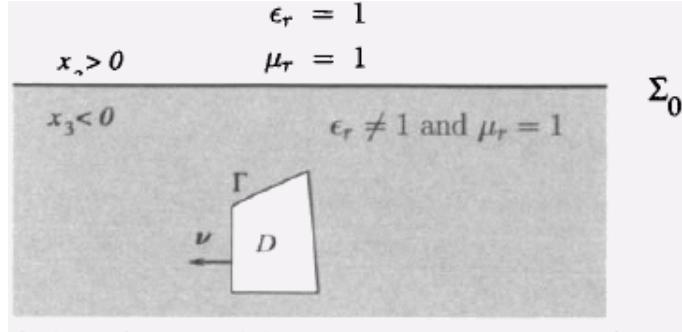
where  $\nu$  is the unit outward normal to  $\Sigma$  and  $E_T = (\nu \times E|_{\Sigma}) \times \nu$  (similarly for  $E_T^i$ ). This is just an impedance boundary condition of the form (1.18) with a special choice of the impedance. Equation (1.23) is an example of an absorbing boundary condition used to simulate the infinite domain outside  $\Omega$ . We shall discuss some other possible choices in Section 13.5. Obviously, the solution of the true scattering problem and the problem on a bounded domain are not equal, but the difference can be made small by taking  $\Sigma$  far enough from the scatterer.

One other interesting problem arises if the electromagnetic field is entirely contained in a perfect conducting cavity. In this case  $\Omega$  is a bounded domain with boundary  $\Gamma$  and Maxwell's equations are satisfied in  $\Omega$ . If there is no conductor present (i.e.  $\sigma = 0$ ), there are values of the wave number  $\kappa$  for which the Maxwell system no longer has a unique solution. These values of  $\kappa$  are resonant wave numbers for cavity modes and mathematically are eigenvalues of the Maxwell system.

The second scattering problem we wish to consider is a simple case when the scatterer is unbounded. More complex “gratings” and “rough surfaces” will not be considered in this book (see, e.g. [128]). The problem we shall consider is a simple model for scattering from buried objects (arising, e.g., in simulating ground penetrating radar). The background medium now consists of two regions. The region  $x_3 > 0$  is assumed occupied by air and the region  $x_3 < 0$  is earth. The relative permeability of air and earth is assumed to be unity, and the relative permittivity of earth is assumed constant (with a possibly non-zero imaginary part since the earth is usually a conductor).

At the air–earth interface, we impose the jump conditions discussed in the previous section. We suppose the scatterer (consisting of perfect conductors,

Fig. 1.5. Geometry of the problem of scattering from perfectly conducting obstacles in a layered medium. This is a model for scattering from buried objects. The scatterer  $D$  lies in the lower half space.



although it is also possible to allow inhomogeneous scatterers) occupies a region  $D$  entirely contained in the earth layer (see Fig. 1.5). The sources of the field are assumed to be in the air layer (so  $F = 0$  for  $x_3 < 0$ ). The resulting incident field is assumed to satisfy the Maxwell system in the background layered medium.

Since the air–earth interface is of infinite extent we cannot directly use the simple Silver–Müller condition and instead use an integral radiation condition [243]. Let  $\partial B_R^+$  denote the hemisphere of radius  $R$  on which  $x_3 > 0$  (and similarly  $\partial B_R^-$  for  $x_3 < 0$ ). We require(1.24a)

$$\lim_{x \rightarrow \infty} \int_{\partial B_R^+} |(\nabla \times E^s) \times v - ikE^s|^2 dA = 0, \quad (1.24b)$$

$$\lim_{x \rightarrow \infty} \int_{\partial B_R^-} |(\nabla \times E^s) \times v - ikE^s|^2 dA = 0.$$

## 1.4 Boundary value problems

We shall now summarize the principal boundary value problems for Maxwell's equations in this book. We shall make more precise the assumptions on the domain  $D$  and coefficients  $\epsilon_r$  and  $\mu_r$  in later chapters.

### 1.4.1 Time-harmonic problem in a cavity

Suppose  $\Omega$  is a bounded domain with two disjoint connected boundaries  $\Gamma$  and  $\Sigma$ . We seek to compute the time-harmonic electric field  $E$  corresponding to a given current density  $F$  by solving the time-harmonic electric field equation (1.12) subject to the perfect conducting boundary condition (1.17) and the impedance boundary condition (1.23) as follows:(1.25a)

$$\nabla \times (\mu_r^{-1} \nabla \times E) - k^2 \epsilon_r E = F \text{ in } \Omega, \quad (1.25b)$$

$$v \times E = 0 \text{ in } \Gamma,$$

$$\mu_r^{-1}(\nabla \times E) \times v - i\kappa\lambda E_T = g \text{ on } \Sigma, \quad (1.25c)$$

where  $g$  is a given tangential vector field on  $\Sigma$  (see (1.23) for an example of  $g$  computed from an incident field). We shall allow  $\Sigma$  to be empty in which case these equations model propagation in a cavity with a perfectly conducting wall. For an absorbing boundary condition approximation of a scattering problem,  $\mu_r = 1$  and  $\lambda = 1$  on  $\Sigma$ , and  $\epsilon_r = \mu_r = 1$  in a neighborhood of  $\Sigma$ .

### 1.4.2 Cavity resonator

Given a bounded domain  $\Omega$  with boundary  $\Gamma$ , we seek scalars  $\kappa$  and non-trivial (i.e. not identically zero) electric fields  $E$  which satisfies eqn (1.12) with  $F = 0$ , so that

$$\begin{aligned} \nabla \times (\mu_r \nabla \times E) - \kappa^2 \epsilon_r E &= 0 \text{ in } \Omega, \\ v \times E &= 0 \text{ on } \Gamma. \end{aligned}$$

In addition, since there is no applied current, we require that  $E$  satisfy the divergence condition (1.11a) with  $\varrho = 0$ , so that

$$\nabla \cdot (\epsilon_r E) = 0 \text{ in } \Omega.$$

The effect of this latter condition is to guarantee that there are at most finitely many linearly independent solutions to this problem when  $\kappa = 0$ .

### 1.4.3 Scattering from a bounded object

In this problem the domain of the electromagnetic field is the unbounded region  $\mathbb{R}^3 \setminus D^-$ , where  $D$  is a bounded domain with connected complement. Given a known incoming electric field  $E^i$  satisfying (1.19), we seek to compute the total field  $E$  and scattered field  $E^s$  such that the time-harmonic electric field equation (1.12) holds together with (1.21), so that(1.26)

$$\begin{aligned} \nabla \times (\mu_r^{-1} \nabla \times E) - \kappa^2 \epsilon_r E &= F \text{ in } \mathbb{R}^3 \setminus \bar{D}, \\ E &= E^i + E^s \text{ in } \mathbb{R}^3 \setminus D. \end{aligned} \quad (1.27)$$

We assume that the scatterer is bounded so that  $D$  is bounded and  $\epsilon_r = \mu_r = 1$  outside a sufficiently large ball. On the boundary  $\Gamma$  of the unbounded component of  $\mathbb{R}^3 \setminus D^-$ , we impose the perfect conducting boundary condition,(1.28)

$$E \times v = 0 \text{ on } \Gamma.$$

In addition  $E^s$  must satisfy the Silver–Müller radiation condition (1.22)(1.29)

$$\lim_{\rho \rightarrow \infty} \rho \left( (\nabla \times E^s) \times \hat{x} - i\kappa E^s \right) = 0 \text{ as } r \rightarrow \infty,$$

where  $\rho = |x|$ , uniformly in  $\mathcal{O} = x/|x|$ .

### 1.4.4 Scattering from a buried object

Let

$$\mathbb{R}_+^3 = \left\{ x \in \mathbb{R}^3 \mid x_3 > 0 \right\} \text{ and } \mathbb{R}_-^3 = \left\{ x \in \mathbb{R}^3 \mid x_3 < 0 \right\}.$$

We suppose that the scatterers are contained in a bounded region in the lower half space. The interface between layers is denoted by  $\Sigma_0$  and is the plane  $x_3 = 0$  (see Fig. 1.5). For simplicity, we only consider a perfectly conducting scatterer occupying a bounded domain  $D$  entirely contained in  $\mathbb{R}_-^3$  ( $\text{so } \bar{D} \subset \mathbb{R}_-^3$ ), and we assume that the complement of  $D$  is connected.

The electric field satisfies Maxwell's equations in  $\mathbb{R}_+^3$  (with  $\epsilon_r = \mu_r = 1$ , since the domain is supposed to contain air) and the general Maxwell equation in  $\mathbb{R}_-^3 \setminus \bar{D}$  with  $\mu_r = 1$  and constant  $\epsilon_r = \epsilon_i$ . The integral radiation condition is imposed at infinity. Thus, for given  $F$  having support in  $\mathbb{R}_+^3$  ( $F = 0$  is a possible choice), the total field  $E$  satisfies(1.30)

$$\begin{aligned} \nabla \times \nabla \times E - \kappa^2 E &= F \text{ in } \mathbb{R}_+^3, \\ \nabla \times \nabla \times E - \kappa^2 \epsilon_r E &= 0 \text{ in } \mathbb{R}_-^3 \setminus \bar{D}. \end{aligned} \quad (1.31)$$

Imposing the jump conditions (1.13) and (1.15) we have(1.32)

$$[v \times E] = 0 \text{ and } [v \times (\nabla \times E)] = 0 \text{ on } \Sigma_0,$$

where  $[ \cdot ]$  denotes the jump in its argument across  $\Sigma_0$ . As usual, on the boundary of  $D$ ,

$$E \times v = 0 \text{ on } \Gamma.$$

We suppose that the scattered field is due to a given incident field  $E^i$  which satisfies the background Maxwell system:

$$\begin{aligned} \nabla \times \nabla \times E^i - \kappa^2 E^i &= F \text{ in } \mathbb{R}_+^3, \\ \nabla \times \nabla \times E^i - \kappa^2 \epsilon_r E^i &= 0 \text{ in } \mathbb{R}_-^3, \end{aligned}$$

and the jump conditions (1.32) on  $\Sigma_0$ . Here  $F$  is the function of compact support in  $\mathbb{R}_+^3$  representing the source of the incident field appearing in (1.30). Then we have

$$E = E^i + E^s \text{ in } \mathbb{R}^3 \setminus \bar{D}.$$

and the following integral radiation conditions on the scattered field  $E^s$

$$\lim_{R \rightarrow \infty} \int_{\partial B_R^+} |(\nabla \times E^s) \times v - ik E^s|^2 dA = 0,$$

$$\lim_{R \rightarrow \infty} \int_{\partial B_R^-} |(\nabla \times E^s) \times v - ik E^s|^2 dA = 0,$$

# 2 FUNCTIONAL ANALYSIS AND ABSTRACT ERROR ESTIMATES

## 2.1 Introduction

In our analysis of weak formulations of Maxwell's equations, we shall appeal to certain basic theorems from functional analysis. The reader is presumed to be familiar with these rudimentary concepts. As a result, the first part of this chapter is simply a convenient summary of notation, definitions and theorems with references to the literature. As a background source, a good book is that of McLean [215].

In the second part of the chapter, we turn to some abstract finite element error estimates that will be used later in our proofs of finite element convergence rates. These results, although standard, are verified in detail due to their basic role in the analysis of finite element methods. The notation here is fairly standard, although we use calligraphic symbols like  $\mathcal{X}$  to denote general Hilbert spaces. This is to distinguish them from particular spaces appearing in later chapters.

## 2.2 Basic functional analysis and the Fredholm alternative

The material for this section is mainly taken from the books of Kress [193] and McLean [215], which also contain proofs of the relevant results. In many cases we have quoted theorems for Hilbert spaces even though the theorems hold for more general spaces. Hilbert spaces will be enough for our needs.

### 2.2.1 Hilbert space

If  $\mathcal{X}$  is a vector space over the complex numbers, then a scalar product on  $\mathcal{X}$  is a map  $(\cdot, \cdot)_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{C}$  such that

- (1) if  $u \in \mathcal{X}$  then  $(u, u)_{\mathcal{X}} = 0$  if and only if  $u = 0$ ;
- (2) for all  $u, v \in \mathcal{X}$  we have  $\overline{(u, v)_{\mathcal{X}}} = (v, u)_{\mathcal{X}}$ ;
- (3) for all  $u, v, w \in \mathcal{X}$  and  $\alpha, \beta \in \mathbb{C}$  we have

$$(\alpha u + \beta v, w)_{\mathcal{X}} = \alpha(u, w)_{\mathcal{X}} + \beta(v, w)_{\mathcal{X}}.$$

The norm associated with  $(\cdot, \cdot)_{\mathcal{X}}$  is

$$\|\varphi\|_{\mathcal{X}} = \sqrt{(\varphi, \varphi)_{\mathcal{X}}} \text{ for all } \varphi \in \mathcal{X}.$$

This norm satisfies the usual triangle inequality

$$\|\varphi + \xi\|_{\mathcal{X}} \leq \|\varphi\|_{\mathcal{X}} + \|\xi\|_{\mathcal{X}}, \text{ for all } \xi, \varphi \in \mathcal{X}.$$

**Definition 2.1** Let  $X$  be a vector space with scalar product  $(\cdot, \cdot)_X$ . If  $X$  is complete with respect to the norm  $\|\cdot\|_X$  it is called a *Hilbert space*.

A basic example of such a Hilbert space is  $L^2(\Omega)$ , the space of square-integrable functions on an open domain  $\Omega \subset \mathbb{R}^3$ , which has the scalar product

$$(\varphi, \xi) = \int_{\Omega} \varphi \bar{\xi} dV,$$

where  $\bar{\xi}$  is the complex conjugate of  $\xi$ . Here we have used the notation  $(\cdot, \cdot)$  instead of the more correct  $(\cdot, \cdot)_{L^2(\Omega)}$ . In this case the  $L^2$  scalar product is so important and used so frequently that it is worth immediately breaking our own rules of notation!

Two elementary estimates are used over and over again in our error analysis. The first is the Cauchy–Schwarz inequality.

**Lemma 2.2** (Cauchy–Schwarz inequality) *For all  $u, v \in X$*  (2.1)

$$|(u, v)_X| \leq \|u\|_X \|v\|_X.$$

This is easily proved (when  $u \neq 0$ ) by expanding the inequality

$$\left( t \overline{(u, v)_X} u - v, t \overline{(u, v)_X} u - v \right)_X \geq 0.$$

with  $t = 1 / \|u\|_X^2$ .

The second basic estimate follows from the Cauchy–Schwarz inequality using the observation that for any  $\delta > 0$ , and real numbers  $\alpha$  and  $\beta$  we have  $(\delta^{1/2}\alpha - \delta^{1/2}\beta)^2 \geq 0$ . Expanding this inequality proves the basic arithmetic–geometric mean inequality

$$|\alpha\beta| \leq \frac{\delta}{2} \alpha^2 + \frac{1}{2\delta} \beta^2.$$

Using this result and the Cauchy–Schwarz inequality proves the following lemma.

**Lemma 2.3** (Arithmetic–geometric mean inequality) *Let  $u, v \in X$  and  $\delta > 0$  then* (2.2)

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

A sequence  $\{u_n\}_{n=1}^{\infty} \subset X$  is said to be *convergent* to a function  $v \in X$  if

$$\lim_{n \rightarrow \infty} \|u - u_n\|_X = 0.$$

Sometimes we will emphasize this convergence in norm by speaking of *strong convergence*. This is to distinguish from weak convergence. In particular, a sequence  $\{u_n\}_{n=1}^{\infty} \subset X$  is said to *converge weakly* to a function  $v \in X$  if for each  $\varphi \in X$  we have  $(v_n, \varphi)_X \rightarrow (v, \varphi)_X$  as  $n \rightarrow \infty$ .

Unfortunately, bounded sets in a Hilbert space do not necessarily contain a convergent subsequence. However, we shall make use of the following weaker result.

**Lemma 2.4** Let  $\{u_n\}_{n=1}^{\infty} \subset X$  be bounded. Then this sequence has a weakly convergent subsequence.

We often work with subspaces of suitable vector spaces. Indeed the finite element method is just a method to construct useful subspaces of various function spaces. An important class of subsets of  $X$  is defined next:

**Definition 2.5** A subset  $U$  of a Hilbert space  $X$  is *closed* if it contains all limits of convergent sequences in  $U$ .

We shall frequently encounter situations in which we know a subspace of a Hilbert space which is not closed. We can then create a closed subspace from this subspace as follows (the definition mentions subsets—we shall only use it in the more restrictive case of subspaces).

**Definition 2.6** Given a subset  $U \subset X$ , the *closure* of  $U$  in  $X$  (denoted by  $\text{closure}(U)$ ) is the set of all limits of convergent subsequences of  $U$  using the  $X$  norm. We say a subset  $U \subset X$  is *dense* in  $X$  if  $\text{closure}(U) = X$ .

A particularly simple case will occur frequently throughout the book. If  $\Omega$  is an open subset of  $R^3$ , we denote by  $\bar{\Omega}$  the closure of this subset.

A convenient property of Hilbert spaces is that there exists a best approximation to a given function  $f \in X$  from a closed subspace (Theorem 1.26 of [193]).

**Theorem 2.7** Let  $U \subset X$  be a closed subspace of the Hilbert space  $X$  and let  $f \in X$ . Then there exists a unique  $g \in U$  such that

$$\|f - g\|_X = \inf_{u \in U} \|f - u\|_X.$$

This theorem has some important consequences, in particular the decomposition of a Hilbert space into orthogonal subspaces. We shall use this decomposition to write vector functions as a sum of a gradient and a curl (the Helmholtz decomposition). In general, let  $U$  be a subspace of a Hilbert space  $X$ . We have the following definition and theorem.

**Definition 2.8** The *orthogonal complement* of  $U$ , denoted by  $U^\perp$ , is the closed subspace such that

$$U^\perp = \{v \in X | (v, u)_X = 0 \text{ for all } u \in U\}.$$

**Theorem 2.9** Let  $U$  be a closed subspace of a Hilbert space  $X$ . Then, if  $f \in X$ , there exist unique functions  $u \in U$  and  $v \in U^\perp$  such that

$$f = u + v$$

and we write  $X = U \oplus U^\perp$ .

## 2.2.2 Linear operators and duality

We now need to discuss operators mapping one Hilbert space to another such space. Consider an operator  $A : X \rightarrow Y$ , where  $X$  and  $Y$  are Hilbert spaces. The operators in this book are usually bounded and linear, by which we mean the following:

**Definition 2.10** An operator  $A : X \rightarrow Y$  is said to be *linear* if

$$A(\alpha u + \beta v) = \alpha Au + \beta Av \text{ for all } \alpha, \beta \in \mathbb{C}, u, v \in \mathcal{X}.$$

and is *bounded* if there exists a constant  $C$  such that

$$\|A\varphi\|_Y \leq C\|\varphi\|_{\mathcal{X}} \text{ for all } \varphi \in \mathcal{X},$$

where  $C$  is independent of  $\varphi$ . In addition,  $A$  is said to be *continuous* if, for every  $\varphi \in X$  and sequence  $\{\varphi_n\}_{n=1}^{\infty}$  converging to  $\varphi$  in  $X$ , we have  $A\varphi_n \rightarrow A\varphi$  in  $Y$  as  $n \rightarrow \infty$ .

A useful theorem is then the following (see, e.g. Theorem 2.5 of [193]).

**Theorem 2.11** A linear operator is continuous if and only if it is bounded.

We can also define a norm for an operator to be the optimal boundedness constant or more precisely as follows:

**Definition 2.12** The *natural norm* of a bounded linear operator  $A : X \rightarrow Y$  is given by

$$\|A\|_{\mathcal{X} \rightarrow \mathcal{Y}} = \sup_{\varphi \neq 0, \varphi \in \mathcal{X}} \frac{\|A\varphi\|_Y}{\|\varphi\|_{\mathcal{X}}}.$$

The identity operator  $I : X \rightarrow X$  is the operator such that  $Ix = x$  for every  $x \in X$ . Obviously  $\|I\|_{X \rightarrow X} = 1$ .

Standard spaces related to the operator  $A$  are as follows:

**Definition 2.13** The *range* of the operator  $A : X \rightarrow Y$  is denoted by  $A(X)$  and given by

$$A(\mathcal{X}) = \{y \in \mathcal{Y} \mid y = Ax \text{ for some } x \in \mathcal{X}\}.$$

We denote by  $N(A)$  the *null-space* of  $A$ , so that

$$N(A) = \{x \in \mathcal{X} \mid Ax = 0\}.$$

Let  $A : X \rightarrow Y$  be a bounded linear operator. We now wish to define some useful operators related to  $A$ . There exists a unique linear operator  $A^* : Y \rightarrow X$  called the *adjoint operator* such that(2.3)

$$(Ax, y)_Y = (x, A^*y)_X \text{ for all } x \in \mathcal{X} \text{ and } y \in \mathcal{Y}.$$

Sometimes it is desirable to use the *dual operator* to  $A$  instead of the adjoint. To define the dual operator we need first to define the dual space of a Hilbert space  $X$  as follows:

**Definition 2.14** For a given Hilbert space  $X$ , the *dual space*  $X'$  is the space of bounded linear functionals on  $X$ . If  $f \in X'$  then the norm of  $f$  is

$$\|f\|_{X'} = \sup_{x \in X, x \neq 0} \frac{|f(x)|}{\|x\|_X}.$$

We define the *dual pairing*  $\langle\langle \cdot, \cdot \rangle\rangle_x$  by

$$\langle\langle g, u \rangle\rangle_x = g(u) \text{ for all } u \in X \text{ and } g \in X'.$$

Having defined the dual space and dual pairing, we can now define the *dual operator* denoted by  $A^\top : Y' \rightarrow X'$ , where  $X'$  and  $Y'$  are the dual spaces of  $X$  and  $Y$ , respectively. If  $\langle\langle \cdot, \cdot \rangle\rangle_y$  denotes the corresponding pairing for  $Y$  then  $A^\top$  is defined by

$$\langle\langle Ax, y \rangle\rangle_y = \langle\langle x, A^\top y \rangle\rangle_x \text{ for all } x \in X \text{ and } y \in Y'.$$

Note that the dual pairing does not imply conjugation of the second argument of  $\langle\langle \cdot, \cdot \rangle\rangle_x$  or  $\langle\langle \cdot, \cdot \rangle\rangle_y$ . We need one more concept. If  $V \subseteq Y'$  then the *annihilator* of  $V$  denoted by  ${}^aV$  is defined by

$${}^aV = \{u \in Y \mid \langle\langle g, u \rangle\rangle_y = 0 \text{ for all } g \in V\}$$

The following result is well-known (see, e.g, Theorem 4.6 of [94] and Theorem 2.10 of [215] and a density argument):

**Theorem 2.15** Let  $A : X \rightarrow Y$  be bounded and linear. Then

$$\begin{aligned} A(X)^\perp &= N(A^*), \quad N(A^*)^\perp = \text{closure}(A(X)) \text{ and} \\ \text{closure}(A(X)) &= {}^a(N(A^\top)). \end{aligned}$$

**Remark 2.16** The chief use of this result will be to prove “density” results. We prove that either  $N(A^*) = \{0\}$  or  $N(A^\top) = \{0\}$  and then can conclude that  $A(X)$  is dense in  $Y$ .

### 2.2.3 Variational problems

We shall be interested in approximating variational problems posed in Hilbert spaces. Various theories exist that provide conditions on the underlying variational problem to guarantee the existence and uniqueness of a solution. We start by recalling the simplest of these: the Riesz representation theorem in the form given by Theorem 2.30 of [215]. This theorem justifies the claim of the existence of  $A^*$  in (2.3).

**Theorem 2.17** Let  $X$  be a Hilbert space. For each  $g \in X'$  there exists a unique  $u \in X$  such that

$$(u, v)_X = g(v) \text{ for all } v \in X.$$

Furthermore,  $\|u\|_X = \|g\|_{X'}$ .

Unfortunately, the Riesz representation theorem will not be sufficient for our purposes. We need a famous generalization called the Lax–Milgram lemma. Before stating this result, we need the following definitions.

**Definition 2.18** Let  $X$  and  $Y$  be Hilbert spaces. A mapping  $a(\cdot, \cdot) : X \times Y \rightarrow \mathbb{C}$  is called a *sesquilinear form* if

$$\begin{aligned} a(a_1 u + a_2 v, \varphi) &= a_1 a(u, \varphi) + a_2 a(v, \varphi) \\ &\quad \text{for all } a_1, a_2 \in \mathbb{C}, u, v \in X \text{ and } \varphi \in Y. \\ a(u, \beta_1 \varphi + \beta_2 \chi) &= \beta_1 a(u, \varphi) + \beta_2 a(u, \chi) \\ &\quad \text{for all } \beta_1, \beta_2 \in \mathbb{C}, u \in X \text{ and } \varphi, \chi \in Y. \end{aligned}$$

As before, over-bar (e.g.  $\bar{\beta}$ ) denotes complex conjugation.

An obvious example of a sesquilinear form is the  $L^2(\Omega)$  scalar product

$$(u, \varphi) = \int_{\Omega} \bar{u} \varphi dV.$$

**Definition 2.19** A sesquilinear form  $a(\cdot, \cdot)$  defined on  $X \times Y$ , where  $X$  and  $Y$  are Hilbert spaces, is said to be *bounded* if there is a constant  $C$  independent of  $u \in X$  and  $\varphi \in Y$  such that

$$|a(u, \varphi)| \leq C \|u\|_X \|\varphi\|_Y \text{ for all } u \in X \text{ and } \varphi \in Y.$$

**Definition 2.20** The sesquilinear form  $a : X \times X \rightarrow \mathbb{C}$ , where  $X$  is a Hilbert space, is said to be *coercive* if there is a constant  $\alpha > 0$  independent of  $u \in X$  such that

$$|a(u, u)| \geq \alpha \|u\|_X^2 \text{ for all } u \in X.$$

Note that many books use the term “strictly coercive” for the form of coercivity defined here.

Given a Hilbert space  $X$  and a bounded coercive sesquilinear form  $a(\cdot, \cdot)$  on  $X \times X$ , we now consider the variational problem of finding  $u \in X$  such that(2.4)

$$a(u, \varphi) = f(\varphi) \text{ for all } \varphi \in X,$$

where  $f \in X'$  is a given linear functional. The following lemma summarizes the existence and uniqueness theory for this problem.

**Lemma 2.21** (Lax–Milgram) Suppose  $a : X \times X \rightarrow \mathbb{C}$  is a bounded and coercive sesquilinear form. Then for each  $f \in X'$  there exists a unique solution  $u \in X$  to(2.4)and

$$\|u\|_X \leq \frac{C}{\alpha} \|f\|_{X'}$$

where  $C$  and  $\alpha$  are the constants in the boundedness and coercivity definitions above.

In one case later in this chapter the Lax–Milgram lemma will not be sufficient and we need a further generalization (see Theorem 1.4.3 of [244] and also [24]). This generalization uses a sesquilinear form defined on the product of two different spaces.

**Theorem 2.22** (Generalized Lax–Milgram lemma) *Let  $X$  and  $Y$  be Hilbert spaces and let  $a(\cdot, \cdot)$  denote a bounded sesquilinear form on  $X \times Y$  which has the following properties:*

(1) *There is a constant  $\alpha$  such that*

$$\inf_{u \in X, \|u\|_X = 1} \sup_{v \in Y, \|v\|_Y \leq 1} |a(u, v)| \geq \alpha > 0.$$

(2) *For every  $v \in Y$ ,  $v \neq 0$*

$$\sup_{u \in X} |a(u, v)| > 0.$$

*Suppose  $g \in Y'$  then there exists a unique  $u \in X$  such that*

$$a(u, \varphi) = g(\varphi) \text{ for all } \varphi \in Y.$$

*Moreover,*

$$\|u\|_X \leq \frac{C}{\alpha} \|g\|_{Y'}.$$

Condition (i) in this theorem is one form of the Babuška–Brezzi or inf-sup condition and generalizes the coercivity property. The Babuška–Brezzi condition is often stated a little differently in the context of the variational theory of mixed problems. In this theory we have two Hilbert spaces  $X$  and  $S$  and sesquilinear forms,

$$a: X \times X \rightarrow \mathbb{C} \text{ and } b: X \times S \rightarrow \mathbb{C}.$$

These are assumed to be bounded, so there is a constant  $C > 0$  such that

$$\begin{aligned} |a(u, \varphi)| &\leq C \|u\|_X \|\varphi\|_X \text{ for all } u, \varphi \in X, \\ |a(u, \xi)| &\leq C \|u\|_X \|\xi\|_X \text{ for all } u \in X, \xi \in S. \end{aligned}$$

In order to develop an existence theory for the upcoming mixed variational problem, we need to assume that  $a(\cdot, \cdot)$  is coercive, but not on all of  $X$ . To this end, let(2.5)

$$\mathcal{Z} = \{u \in X \mid b(u, \xi) = 0 \text{ for all } \xi \in S\}.$$

**Definition 2.23** The sesquilinear form  $a(\cdot, \cdot)$  is said to be *Z-coercive* if there exists a constant  $\alpha > 0$ , such that(2.6)

$$|a(u, u)| \geq \alpha \|u\|_X^2 \text{ for all } u \in \mathcal{Z}$$

where  $\alpha$  is independent of  $u$ .

In addition, we need to assume an appropriate condition on  $b(\cdot, \cdot)$ . It follows from the inf-sup condition in Theorem 2.22 that the appropriate condition, usually referred to as the *Babuška–Brezzi condition*, is the following.

**Definition 2.24** The sesquilinear form  $b(\cdot, \cdot)$  is said to satisfy the *Babuška–Brezzi condition* if there exists a constant  $\beta > 0$  such that, for all  $p \in S$ ,

$$\sup_{w \in \mathcal{X}} \frac{|b(w, p)|}{\|w\|_{\mathcal{X}}} \geq \beta \|p\|_S,$$

where  $\beta$  is independent of  $p$ .

We can now state the following theorem, which is a consequence of the generalized Lax–Milgram lemma and can be found in a special case in Section 10.2 of [60] and in more generality in Theorem 1.1, p. 42, of [61] (where we also use Lemma 4.2 of [57]).

**Theorem 2.25** Let  $X$  and  $S$  be Hilbert spaces and let  $a : X \times X \rightarrow C$  and  $b : X \times S \rightarrow C$  be bounded sesquilinear forms that satisfy the Z-coercivity and Babuška–Brezzi conditions given in (2.6) and (2.7), respectively. Suppose  $f \in X'$  and  $g \in S'$  and consider the problem of finding  $u \in X$  and  $p \in S$  such that (2.8a)

$$\begin{aligned} a(u, \varphi) + b(\varphi, p) &= f(\varphi) \text{ for all } \varphi \in \mathcal{X}, \\ (2.8b) \end{aligned}$$

$$b(u, \xi) = g(\xi) \text{ for all } \xi \in S.$$

Then there exists a unique solution  $(u, p)$  to (2.8) and

$$\|u\|_{\mathcal{X}} + \|p\|_S \leq C(\|f\|_{\mathcal{X}'} + \|g\|_{S'}) .$$

**Remark 2.26** The system (2.8) is often referred to as a “mixed” variational problem because it first arose in studies of mixed variational problems in elasticity theory. There are many excellent books devoted to the study of mixed methods where the reader will find proofs and examples. Our presentation follows Brenner and Scott [60] for the most part, with some material taken from Brezzi and Fortin [61].

Lemma 4.2 of [57] shows that, if  $b(\cdot, \cdot)$  satisfies the Babuška–Brezzi condition, there is at least one function  $u_0 \in X$  such that  $b(u_0, \xi) = g(\xi)$  for all  $\xi \in S$  and such that  $\|u_0\|_X \leq C\|g\|_{S'}$ , where  $C$  is independent of  $g$ .

## 2.2.4 Compactness and the Fredholm alternative

Unfortunately, the theories outlined in the previous section, which are mainly aimed at strictly coercive elliptic problems, do not settle the question of existence for solutions of Maxwell's equations. For this we need to know how certain perturbations of basic elliptic problems behave. In particular, if  $X$  is a Hilbert space and  $F \in X$ , we wish to solve the operator problem of finding  $u \in X$ , such that

$$(I + A)u = F,$$

where  $A : X \rightarrow X$  is bounded and linear. We need conditions under which  $(I + A)^{-1} : X \rightarrow X$  exists and is bounded. Under very restrictive assumptions

on the operator  $\mathcal{A}$ , a simple extension of the binomial theorem can be used to guarantee this. The proof is via a Neumann series and the result is summarized in the following theorem (Theorem 2.8 of [193]):

**Theorem 2.27** *Let  $X$  be a Hilbert space and  $\mathcal{A} : X \rightarrow X$  be a bounded linear operator with  $\|\mathcal{A}\|_{X \rightarrow X} < 1$ . Then  $I + \mathcal{A}$  has a bounded inverse given by the Neumann series*

$$(I + \mathcal{A})^{-1} = \sum_{n=0}^{\infty} (-1)^n \mathcal{A}^n$$

and

$$\|(I + \mathcal{A})^{-1}\|_{X \rightarrow X} \leq \frac{1}{1 - \|\mathcal{A}\|_{X \rightarrow X}}.$$

Particularly for low-frequency problems, we can sometimes prove the existence and uniqueness of solutions to scattering problems by the previous theorem. However, for higher wavenumbers this is not sufficient. We remedy this by restricting  $\mathcal{A}$  to be in a special class of operators. To describe this class requires some more definitions.

**Definition 2.28** A subset  $U$  of a Hilbert space  $X$  is said to be *compact* if every sequence of elements from  $U$  contains a subsequence converging to an element of  $U$ .

In fact, we have defined here the notion of sequential compactness. For a Hilbert space this notion is equivalent to more general definitions (see Theorem 1.15 of [193]).

**Definition 2.29** A subset  $U$  of a Hilbert space  $X$  is *relatively compact* if its closure is compact.

Now we can define a class of operators that plays a central role in scattering theory.

**Definition 2.30** A linear operator  $\mathcal{A} : X \rightarrow Y$  from a Hilbert space  $X$  to a Hilbert space  $Y$  is said to be *compact* if it maps bounded sets in  $X$  to relatively compact sets in  $Y$ .

Thus, to prove compactness of an operator  $\mathcal{A} : X \rightarrow Y$ , we need to show that for every bounded sequence  $\{\varphi_n\}_{n=0}^{\infty}$  in  $X$ , the sequence  $\{\mathcal{A}\varphi_n\}_{n=0}^{\infty}$  in  $Y$  contains a convergent subsequence. An alternative approach is justified by the next theorem, if we can decompose the operator into the product of a compact and bounded operator (Theorem 2.15 of [193]).

**Theorem 2.31** *Let  $X, Y, Y$  be Hilbert spaces and let  $\mathcal{A} : X \rightarrow Y$  and  $\mathcal{B} : Y \rightarrow Y$  be bounded linear operators. Then the product  $\mathcal{B}\mathcal{A} : X \rightarrow Y$  is compact if one of the operators  $\mathcal{A}$  or  $\mathcal{B}$  is compact.*

Another useful result is the following (Theorem 2.19 of [193]).

**Lemma 2.32** *Let  $X$  be a Hilbert space. Then the identity map  $I : X \rightarrow X$  is compact if and only if  $X$  is finite dimensional.*

The method we shall adopt for proving that certain variational formulations of Maxwell's equations have a solution is to appeal to Fredholm theory. This theory can be stated in much more generality than we shall give here (see [193]). The next theorem is a combination of Theorems 2.22 and 2.27 from [215].

**Theorem 2.33** *Let  $B : X \rightarrow X$  be a bounded linear operator where  $X$  is a Hilbert space. Suppose  $B = I + A$ , where  $A$  is a compact operator and  $I$  is the identity operator. Then either*

- (1) *The homogeneous equation  $Bu = 0$  has only the trivial solution  $u = 0$  in  $X$ . In this case, for every  $f \in X$ , the inhomogeneous equation  $Bu = f$  has a unique solution depending continuously on  $f$ ; or*
- (2) *The homogeneous equation  $Bu = 0$  has exactly  $p$  linearly independent solutions for some finite integer  $p > 0$ .*

## 2.2.5 Hilbert–Schmidt theory of eigenvalues

If case (2) of Theorem 2.33 holds, we have that  $Bu = 0$  has at least one nontrivial solution and thus there exists a  $u \in X$ ,  $u \neq 0$ , such that

$$Au = -u.$$

The function  $u$  is said to be an eigenfunction of  $A$  corresponding to the eigenvalue (-1). More generally, we have the following definition.

**Definition 2.34** A function  $u \in X$  and a scalar  $\gamma \in C$  are respectively an *eigenfunction* and corresponding *eigenvalue* of an operator  $A : X \rightarrow X$  if

$$Au = \gamma u \text{ and } u \neq 0.$$

For a general compact operator  $A$ , we cannot conclude that there exist eigenvalues and eigenvectors without further conditions on  $A$ . An important case in electromagnetism occurs when  $A$  is self-adjoint, by which we mean:

**Definition 2.35** An operator  $A : X \rightarrow X$  is *self-adjoint* if

$$(Au, v)_X = (u, Av)_X \text{ for all } u, v \in X.$$

For self-adjoint and compact operators we have the classical Hilbert–Schmidt theory. The following version of this theory is Theorem 2.36 from [215]

**Theorem 2.36** *If  $A : X \rightarrow X$  is a compact, self-adjoint, linear operator on a Hilbert space  $X$ , then there exists a possibly finite sequence of eigenfunctions  $u_1, u_2, \dots$  and real eigenvalues  $\gamma_1, \gamma_2, \dots$  such that*

- (1)  $Au_j = \gamma_j u_j$  and  $u_j \neq 0$ ,  $j = 1, 2, \dots$ ;
- (2)  $u_j$  is orthogonal to  $u_n$  if  $j \neq n$ ;
- (3)  $|\gamma_1| \geq |\gamma_2| \geq \dots > 0$ ;
- (4) if the sequence of eigenvalues is infinite  $\lim_{j \rightarrow \infty} \gamma_j = 0$ ;
- (5)  $Au = \sum_{j \geq 1} \gamma_j(u, u_j)u_j$  with convergence in  $X$  when the sum has infinitely many terms;
- (6) let  $W = \text{span}\{u_1, u_2, \dots\}$ , then  $X = \text{closure}(W) \oplus N(A)$ , where, as usual, the null-space of  $A$  is  $N(A) = \{u \in X \mid Au = 0\}$ .

## 2.3 Abstract finite element convergence theory

Now we turn our attention to some standard convergence theories for abstract finite element variational problems. Due to the central importance of these results to our later error estimates, we provide detailed proofs of these theorems. Each of the existence results quoted in the previous section (Lax–Milgram, mixed problem, Fredholm alternative) has a corresponding finite element convergence theory.

We suppose that we have a sequence of finite-dimensional subspaces denoted by  $X_b$ ,  $b > 0$ , of a Hilbert space  $X$ . These will actually be spaces of finite element functions, and  $b$  will denote the maximum diameter of the elements in the underlying mesh. However, at this stage it suffices to assume that the spaces satisfy  $X_b \subset X$ ,  $b > 0$ , and that they are finite dimensional. Since  $X_b \subset X$  for each  $b$ , we say that the approximation is *conforming*. We can think that  $X_b$  becomes larger (having an increasing dimension) as  $b \rightarrow 0$ , but we make no use of that fact here.

In our error estimates, frequent use will be made of a generic constant  $C$  everywhere different. This avoids the use of more and more subscripts to keep track of constants in the estimates, and is entirely standard in publications on error estimates. Rarely we shall be forced to keep track of constants and then denote them by  $C_1, C_2, \dots$ .

### 2.3.1 Cea's lemma

The simplest convergence result is the finite-dimensional, or discrete, analogue of the Lax–Milgram lemma [80] termed Cea's lemma.

**Lemma 2.37** (Cea) *Suppose  $X_b \subset X$ ,  $b > 0$ , is a family of finite-dimensional subspaces of a Hilbert space  $X$ . Suppose  $a : X \times X \rightarrow C$  is a bounded, coercive sesquilinear form and  $f \in X'$ . Then the problem of finding  $u_b \in X_b$  such that (2.9)*

$$a(u_b, \varphi_b) = f(\varphi_b) \text{ for all } \varphi_b \in X_b$$

*has a unique solution. If  $u \in X$  is the exact solution solving (2.4) then there is a constant  $C$  independent of  $u$ ,  $u_b$  and  $b$  such that (2.10)*

$$\|u - u_b\|_X \leq C \inf_{\mathcal{X}_b \in \mathcal{X}_b} \|u - \mathcal{X}_b\|_X.$$

**Remark 2.38** Estimate (2.10) is said to be a quasi-optimal error estimate since, up to the constant  $C$ , the actual error  $\|u - u_b\|_X$  is bounded by the best approximation error  $\inf_{\mathcal{X}_b \in \mathcal{X}_b} \|u - \mathcal{X}_b\|_X$ . An optimal estimate would have  $C = 1$ . Note that a best-approximation function exists by Theorem 2.7, but it is not generally the solution  $u_b$  of the variational problem (2.9).

**Proof of Lemma 2.37** Since  $X_b \subset X$  we can see that  $a : X_b \times X_b \rightarrow C$  inherits the boundedness and coercivity properties from the sesquilinear form on  $X \times X$  with the same constants. Hence, by the Lax–Milgram Lemma 2.21 applied to

(2.9), we know that a unique solution  $u_b \in X_b$  exists. Now taking  $\varphi = \varphi_b$  in (2.4) and subtracting (2.9) from (2.4) gives the Galerkin orthogonality relation,

$$\alpha(u - u_h, \varphi_h) = 0 \text{ for all } \varphi_h \in \mathcal{X}_h.$$

But for any  $X_b \in X_b$  using this result,

$$\begin{aligned} \alpha(u - u_h, u - u_h) &= \alpha(u - u_h, u - \mathcal{X}_h) + \alpha(u - u_h, \mathcal{X}_h - u_h) \\ &= \alpha(u - u_h, u - \mathcal{X}_h). \end{aligned}$$

Using this equality and the coercivity and boundedness properties of  $\alpha(\cdot, \cdot)$ ,

$$\begin{aligned} \alpha\|u - u_h\|_{\mathcal{X}}^2 &\leq |\alpha(u - u_h, u - u_h)| \\ &= |\alpha(u - u_h, u - u_h)| \leq C\|u - u_h\|_{\mathcal{X}}\|u - u_h\|_{\mathcal{X}}. \end{aligned}$$

Hence

$$\|u - u_h\|_{\mathcal{X}} \leq \left(\frac{C}{\alpha}\right)\|u - \mathcal{X}_h\|_{\mathcal{X}} \text{ for all } \mathcal{X}_h \in \mathcal{X}_h.$$

□

### 2.3.2 Discrete mixed problems

We now need to perform an error analysis for discretizations of the mixed system (2.8). Here we no longer have a coercive bilinear form. Suppose we have finite-dimensional subspaces  $X_b \subset X$  and  $S_b \subset S$  (again indexed by  $b > 0$ ). The discrete mixed problem is to find  $u_b \in X_b$  and  $p_b \in S_b$  such that (2.11a)

$$\begin{aligned} \alpha(u_h, \varphi_h) + b(\varphi_h, p_h) &= f(\varphi_h) \text{ for all } \varphi_h \in \mathcal{X}_h, \\ (2.11b) \end{aligned}$$

$$b(u_h, \xi_h) = g(\xi_h) \text{ for all } \xi_h \in S_h.$$

As in the continuous mixed problem (2.8),  $f \in X'$  and  $g \in S'$  are given functionals. It turns out that we need to assume that  $\alpha(\cdot, \cdot)$  is coercive on a subset of  $X_b \times X_b$  and that  $b(\cdot, \cdot)$  satisfies the Babuška-Brezzi condition at the discrete level. In particular, let us define the discrete analogue of the space  $Z$  defined in (2.5) as follows: (2.12)

$$\mathcal{Z}_h = \{u_h \in \mathcal{X}_h \mid b(u_h, \xi_h) = 0 \text{ for all } \xi_h \in S_h\}.$$

We assume that  $\alpha(\cdot, \cdot)$  is uniformly  $Z$ -coercive so there is a constant  $\alpha > 0$  such that (2.13)

$$|\alpha(u_h, u_h)| \geq \alpha\|u_h\|_{\mathcal{X}}^2 \text{ for all } u_h \in \mathcal{Z}_h,$$

where  $\alpha$  is independent of  $b$  and  $u_b \in Z_b$ .

Similarly, we also need a *discrete Babuška–Brezzi condition*, so we assume there is a constant  $\beta > 0$  independent of  $h$  and  $p_b$  such that(2.14)

$$\sup_{\varphi_h \in \mathcal{X}_h} \frac{|b(\varphi_h, p_h)|}{\|\varphi_h\|_{\mathcal{X}}} \geq \beta \|p_h\|_s .$$

In what follows it will also be useful to define

$$\mathcal{Z}_h(g) = \{u_h \in \mathcal{X}_h \mid b(u_h, \xi_h) = g(\xi_h) \text{ for all } \xi_h \in S_h\} .$$

We can now state and prove a basic existence and uniqueness result for (2.11) [139, 60, 61]:

**Theorem 2.39** Assume that  $a : X \times X \rightarrow C$  and  $b : X \times S \rightarrow C$  are bounded sesquilinear forms satisfying, respectively, the discrete coercivity condition(2.13)and the discrete Babuška–Brezzi condition (2.14). Then, provided  $Y_b(g)$  is not empty, there exists a unique solution to (2.11).

**Remark 2.40** For this theorem it suffices that(2.13)and(2.14)hold with constants  $\alpha = \alpha(h) > 0$  and  $\beta = \beta(h) > 0$  for each  $h$  (i.e. they may depend on  $h$ ). But, to obtain quasi-optimal error estimates later in this section, we need  $\alpha$  and  $\beta$  to be positive independent of  $h$ , as was assumed above.

The proof of this theorem shows that, if  $Y_b(g) \neq \emptyset$  and if  $a(\cdot, \cdot)$  is  $Y_b$ -coercive, a solution  $u_b$  to the discrete problem exists even though  $p_b$  may not be uniquely determined.

**Proof of Theorem 2.39** This is a direct consequence of Theorem 2.25. We provide a proof here to allow us to explain where the  $Y_b$ -coercivity property (2.13) and the Babuška–Brezzi condition (2.14) are used in analyzing (2.11). Since  $Y_b(g)$  is not empty, there is a function  $u_h^{(0)} \in \mathcal{Z}_h(g)$ . Then we may write  $u_h = u_h^{(1)} + u_h^{(0)}$ , with  $u_h^{(1)} \in \mathcal{Z}_h$ . So, using (2.11a),

$$a(u_h^{(0)} + u_h^{(1)}, \varphi_h) + b(\varphi_h, p_h) = f(\varphi_h) \text{ for all } \varphi_h \in \mathcal{X}_h .$$

Choosing  $\varphi_h \in Y_b$ , we see that  $b(\varphi_h, P_b) = 0$ , so that  $u_h^{(0)} \in \mathcal{Z}_h$  satisfies

$$a(u_h^{(1)}, \varphi_h) = f(\varphi_h) - a(u_h^{(0)}, \varphi_h) \text{ for all } \varphi_h \in \mathcal{Z}_h .$$

But  $a(\cdot, \cdot)$  is  $Y_b$ -coercive and, of course, bounded, so by the Lax–Milgram lemma there is a unique solution  $u_h^{(1)} \in \mathcal{Z}_h$  to this equation and the existence of  $u_b$  is verified. Once we have found  $u_b$ , we can find  $p_b \in S_b$  by solving

$$b(\varphi_h, p_b) = -a(u_b, \varphi_h) + f(\varphi_h) \text{ for all } \varphi_h \in \mathcal{X}_h .$$

This is a generalized variational problem. First note that if  $\varphi_b \in Y_b$  then  $b(\varphi_b, p_b) = 0$  and

$$-\alpha(u_h, \varphi_h) + f(\varphi_h) = -\alpha(u_h, \varphi_h) - b(\varphi_h, p_h) + f(\varphi_h) = 0,$$

so the equation is trivial for  $\varphi_h \in Y_h$ . But by the Projection Theorem 2.9 we have  $X_h = Z_h \oplus Z_h^\perp$ , so we need only analyze the problem of finding  $p_h \in S_h$  such that (2.15)

$$b(\varphi_h, p_h) = -\alpha(u_h, \varphi_h) + f(\varphi_h) \text{ for all } \varphi_h \in Z_h^\perp.$$

Using the Babuška–Brezzi condition (we can replace  $X_h$  by  $Z_h^\perp$  since  $b(\varphi_h, q_h) = 0$  for  $\varphi_h \in Y_h$ ),

$$\sup_{\varphi_h \in Z_h^\perp} \frac{|b(\varphi_h, q_h)|}{\|\varphi_h\|_X} \geq \alpha \|q_h\|_S$$

and, since  $\varphi_h \in Z_h^\perp$ , we must have

$$\sup_{q_h \in S_h} |b(\varphi_h, q_h)| > 0,$$

so by the generalized Lax–Milgram Lemma 2.22 a unique solution exists to (2.15).

It remains to show that  $(u_h, p_h)$  is unique. By linearity, we need only consider equations (2.11) with  $f = g = 0$ . Since  $g = 0$ , we see that  $u_h \in Y_h$  and, taking  $\varphi_h = u_h$  and  $\xi_h = p_h$  in (2.11) and subtracting the two equations, we see that  $a(u_h, u_h) = 0$  so that the  $Y_h$ -coercivity property implies  $u_h = 0$ . Hence we have  $b(\varphi_h, p_h) = 0$  for all  $\varphi_h \in X_h$ , and the Babuška–Brezzi condition implies that  $p_h = 0$ . We have thus verified the uniqueness of the solution.  $\square$

It might be helpful to see in a more concrete way how the discrete system (2.11) behaves at the matrix level. Let  $\{\varphi_{h,j}\}_{j=1}^{m_1}$  be a basis for  $Y_h$ . Then, using a basis in  $Z_h^\perp$ , we can extend the basis for  $Y_h$  to a basis  $\{\varphi_{h,j}\}_{j=1}^m$  of  $X_h$ , where  $m$  is the dimension of  $X_h$ . Let  $\{\xi_{h,j}\}_{j=1}^n$  be a basis for  $S_h$ . Then

$$u_h = \sum_{j=1}^m u_j \varphi_{h,j}, \quad p_h = \sum_{j=1}^n p_j \xi_{h,j}.$$

If  $\vec{u} = (u_1, u_2, \dots, u_m)^T$  and  $\vec{p} = (p_1, p_2, \dots, p_n)^T$  then  $\vec{u}$  and  $\vec{p}$  satisfy the following matrix problem

$$A \vec{u} + B \vec{p} = \vec{F},$$

$$B^* \vec{u} = \vec{G},$$

where  $B^*$  is the conjugate-transpose of  $B$ . The entries of the  $m \times m$  matrix  $A$  are given by

$$A_{l,j} = \alpha(\varphi_{h,j}, \varphi_{h,l}), \quad 1 \leq l, j \leq m$$

and of the  $m \times n$  matrix  $B$  by

$$B_{l,j} = b(\varphi_{h,l}, \varphi_{h,j}), \quad 1 \leq l \leq m, \quad 1 \leq j \leq n.$$

The vectors  $\vec{F}$  and  $\vec{G}$  have entries  $F_l = f(\varphi_{l,h})$ ,  $1 \leq l \leq m$ , and  $G_j = g(\xi_{j,h})$ ,  $1 \leq j \leq n$ . The construction of the basis implies that we may partition

$$\vec{u} = \begin{pmatrix} \vec{u}_1 \\ \vec{u}_2 \end{pmatrix},$$

where  $\vec{u}_1 \in \mathbb{C}^{m_1}$  and  $\vec{u}_2 \in \mathbb{C}^{m-m_1}$  are the coefficients of the components of  $\vec{u}$  in  $Z_h$  and  $Z_h^\perp$  respectively.

But since  $b(\varphi_h, q_h) = 0$  for all  $\varphi_h \in Z_h$  and  $q_h \in S_h$  we see that

$$B = \begin{pmatrix} 0 \\ B_1 \end{pmatrix} \text{ and } A = \begin{pmatrix} A_{11} | A_{12} \\ A_{21} | A_{22} \end{pmatrix},$$

where  $B_1$  is an  $(m-m_1) \times n$  matrix. The discrete system is then (2.16)

$$\begin{pmatrix} A_{11} | A_{12} \\ A_{21} | A_{22} \end{pmatrix} \begin{pmatrix} \vec{u}_1 \\ \vec{u}_2 \end{pmatrix} + \begin{pmatrix} 0 \\ B_1 \end{pmatrix} \vec{p} = \begin{pmatrix} \vec{F}_1 \\ \vec{F}_2 \end{pmatrix}, \quad (2.17)$$

$$(0 | B_1^*) \begin{pmatrix} \vec{u}_1 \\ \vec{u}_2 \end{pmatrix} = \vec{G}.$$

By the assumption that  $Z_h(\varrho)$  is not empty, there is a vector  $\vec{u}_2^{(0)}$  such that  $B_1^* \vec{u}_2^{(0)} = \vec{G}$ . Then from the top row of (2.16)

$$A_{11} \vec{u}_1 = \vec{F}_1 - A_{12} \vec{u}_2^{(0)},$$

and the  $Z_h$ -coercivity property guarantees that the matrix  $A_{11}$  is non-singular. Thus the above equation can be solved for  $\vec{u}_1$ . Now we have satisfied the top row of (2.16) and also (2.17). The second row of (2.16) reads

$$B_1 \vec{p} = \vec{F}_2 - A_{21} \vec{u}_1 - A_{22} \vec{u}_2^{(0)}.$$

The Babuška–Brezzi condition guarantees  $B_1$  is of full rank so that  $p$  is uniquely determined.

Now we wish to prove error estimates. We start with the following lemma due to Falk and Osborn [139].

**Lemma 2.41** Suppose the bounded sesquilinear form  $b : X \times Y \rightarrow \mathbb{C}$  satisfies the discrete Babuška–Brezzi condition (2.14). Then for any function  $v \in X$  there is a unique function  $u_h \in Z_h^\perp$  such that

$$b(v - u_h, \varphi_h) = 0 \text{ for all } \varphi_h \in S_h.$$

Furthermore,

$$\|u_h\|_X \leq \frac{C}{\alpha} \|v\|_X.$$

**Remark 2.42** By Remark 2.26 there is a function  $v \in Z^\perp$  such that  $b(v, \varphi) = g(\varphi)$  for all  $\varphi \in S$ . Thus, by Lemma 2.41, there is a  $v_b \in Z_h^\perp$  such that  $b(v_b, \varphi_b) = b(v, \varphi_b) = g(\varphi_b)$  for all  $\varphi_b \in S_b$ . Hence  $Z_b(g) \neq \emptyset$  whenever the discrete and continuous Babuška–Brezzi conditions are satisfied.

The operator mapping  $v$  to  $v_b$  is sometimes termed the Fortin interpolation operator.

**Proof of Lemma 2.41** This is essentially a repeat of the last part of the proof of existence for Theorem 2.39 using the generalized Lax–Milgram lemma. We do not give it here.  $\square$

Next we show that, even without the discrete Babuška–Brezzi condition,  $u_b$  is a good approximation to  $u$  (see Theorem 10.3.7 of [60]). The problem with the upcoming estimate is that it requires a knowledge of the approximation properties of  $Z_b(g)$ , which is unlikely to be readily available.

**Theorem 2.43** Suppose that the sesquilinear form  $a : X \times X \rightarrow \mathbb{C}$  is bounded and  $Z_b$ -coercive so that (2.13) holds. In addition, suppose  $b : X \times S \rightarrow \mathbb{C}$  is bounded. Let  $(u, p) \in X \times S$  satisfy (2.8) and let  $(u_b, p_b) \in X_b \times S_b$  satisfy (2.11). Then there is a constant  $C$  independent of  $b$ ,  $(u_b, p_b)$  and  $(u, p)$  such that

$$\|u - u_b\|_X \leq C \left\{ \inf_{v_b \in Z_b(g)} \|u - v_b\|_X + \inf_{q_b \in S_b} \|p - q_b\|_S \right\}.$$

**Proof** This is essentially an application of the Strang lemma [80] but we proceed from first principles. Note that  $Z_b(g) \neq \emptyset$  ( $p_b \in Z_b(g)$ ). If  $v_b \in Z_b(g)$  then  $v_b - u_b \in Z_b$ . So, for any  $v_b \in Z_b(g)$ , using the triangle inequality and the  $Z_b$ -coercivity of  $a(\cdot, \cdot)$  we have (2.18)

$$\begin{aligned} \|u - u_b\|_X &\leq \|u - u_b\|_X + \|u_b - u\|_X \\ &\leq \|u_b - u\|_X + \frac{1}{\alpha} \frac{|a(u_b - u, u_b - u)|}{\|u_b - u\|_X} \\ &\leq \|u_b - u\|_X + \sup_{w_b \in Z_b} \frac{|a(u_b - u, w_b)|}{\|w_b\|_X} \\ &\leq \|u_b - u\|_X + \frac{1}{\alpha} \left\{ \sup_{w_b \in Z_b} \frac{|a(u_b - u, w_b)|}{\|w_b\|_X} + \sup_{w_b \in Z_b} \frac{|a(u_b - u, w_b)|}{\|w_b\|_X} \right\} \\ &\leq \left(1 + \frac{C}{\alpha}\right) \|u_b - u\|_X + \frac{1}{\alpha} \sup_{w_b \in Z_b} \frac{|a(u_b - u, w_b)|}{\|w_b\|_X}, \end{aligned}$$

where we have also used the continuity of  $a(\cdot, \cdot)$ . However, using eqn (2.8a) for  $u$ , and the fact that  $w_b \in Z_b$ ,

$$\begin{aligned} a(u - u_b, w_b) &= a(u, w_b) - a(u_b, w_b) \\ &= a(u, w_b) - f(w_b) \\ &= -b(w_b, p) = -b(w_b, p - q_b) \text{ for all } q_b \in S_b. \end{aligned}$$

Hence, using the boundedness of  $b(\cdot, \cdot)$ , we have  $|a(u - u_b, w_b)| \leq C \|w_b\| X \|p - q_b\|_S$ . Use of this inequality in (2.18) completes the estimate.  $\square$

The next theorem shows that  $p - p_b$  can be estimated provided that both the  $Z_b$ -coercivity condition and discrete Babuška–Brezzi condition hold (see Theorem 10.5.12 of [60])

**Lemma 2.44** Suppose the conditions of Theorem 2.43 hold, and in addition that the discrete Babuška–Brezzi condition given in (2.14) holds. Then

$$\|p - p_h\|_S \leq \frac{C}{\beta} \|u - u_h\|_{\mathcal{X}} + \left(1 + \frac{C}{\beta}\right) \inf_{q_h \in S_h} \|p - q_h\|_S.$$

**Proof** Let  $\varphi = \varphi_b$  in (2.8a). Subtracting (2.11a) from (2.8a) we obtain

$$b(\varphi_h, p - p_h) = -a(u - u_h, \varphi_h) \text{ for all } \varphi_h \in \mathcal{X}_h.$$

Using this equality, the discrete Babuška–Brezzi condition and eqns (2.8) and (2.11) we find that for any  $qb \in S_b$ ,

$$\begin{aligned} \beta \|q - p_h\|_S &\leq \sup_{\varphi_h \in \mathcal{X}_h} \frac{|b(\varphi_h, q_h - p_h)|}{\|\varphi_h\|_{\mathcal{X}}} \\ &= \sup_{\varphi_h \in \mathcal{X}_h} \frac{|b(\varphi_h, p - p_h) + b(\varphi_h, q_h - p)|}{\|\varphi_h\|_{\mathcal{X}}} \\ &= \sup_{\varphi_h \in \mathcal{X}_h} \frac{|-a(u - u_h, \varphi_h) + b(\varphi_h, q_h - p)|}{\|\varphi_h\|_{\mathcal{X}}} \\ &\leq C(\|u - u_h\|_{\mathcal{X}} \|q_h - p\|_S), \end{aligned}$$

where we have used the boundedness of  $a(\cdot, \cdot)$  and  $b(\cdot, \cdot)$  in deriving the last line. Writing

$$\|p - p_h\|_S \leq \|p - q_h\|_S + \|q_h - p_h\|_S$$

and using the above inequality provides the proof.  $\square$

We can now state and prove our final error estimate for mixed methods (see Corollary 10.5.18 of [60]). This result should be compared to that in Lemma 2.43. The important difference is that  $Z_b(q)$  in the estimate of that lemma has been replaced by  $X_b$  so we need only know about the approximation properties of the full finite element space to understand convergence.

**Theorem 2.45** Suppose that the discrete and continuous coercivity conditions given by (2.6) and (2.13) and the discrete and continuous Babuška–Brezzi conditions given by (2.7) and (2.14) are satisfied. Then there is a unique solution  $(u, p) \in X \times S$  satisfying (2.8) and a unique solution  $(u_b, p_b) \in X_b \times S_b$  satisfying (2.11). There is also a constant  $C$  independent of  $h$ ,  $(u, p)$  and  $(u_b, p_b)$  such that

$$\|u - u_h\|_{\mathcal{X}} + \|p - q_h\|_S \leq C \left\{ \inf_{\mathcal{X}_h \in \mathcal{X}_h} \|u - \mathcal{X}_h\|_{\mathcal{X}} + \inf_{q_h \in S_h} \|p - q_h\|_S \right\}.$$

**Remark 2.46** Often the discrete Babuška–Brezzi condition is difficult to prove, and, for example, [61, 60] devote considerable space to methods for doing this. For Maxwell's equations the continuous and discrete Babuška–Brezzi condition are easily seen to be satisfied, but the  $Z$  or  $Z_b$ -coercivity condition are a good deal more difficult to verify.

**Proof of Theorem 2.45** Existence and uniqueness of the solutions of problems (2.8) and (2.11) follows from Theorems 2.25 and 2.39, together with Remark 2.42. Combining Lemmas 2.43 and 2.44, we have

$$\|u - u_h\|_{\mathcal{X}} + \|p - p_h\|_S \leq C \left\{ \inf_{\mathcal{X}_h \in \mathcal{Z}_h(\mathcal{G})} \|u - \mathcal{X}_h\|_{\mathcal{X}} + \inf_{q_h \in S_h} \|p - q_h\|_S \right\}.$$

It remains to show that in the first term on the right-hand side we can replace  $Z_b(\mathcal{G})$  by  $X_b$ . For any  $v_b \in X_b$  let  $w_b \in z_b^\perp$  satisfy  $b(w_b, qb) = b(u - v_b, qb)$  for all  $qb \in S_b$ . The existence and uniqueness of  $w_b$  is proved in Lemma 2.41. Now, since  $b(u, qb) = g(qb)$ , we have  $b(w_b + v_b, qb) = g(qb)$  for all  $qb \in S_b$  so  $w_b + v_b \in Z_b(\mathcal{G})$ . Using the estimate for  $\|w_b\|_X$  from Lemma 2.41,

$$\|u - (v_h + w_h)\|_{\mathcal{X}} \leq \|u - v_h\|_{\mathcal{X}} + \|w_h\|_{\mathcal{X}} \leq \left(1 + \frac{C}{\alpha}\right) \|u - v_h\|_{\mathcal{X}}.$$

Hence

$$\inf_{\mathcal{X}_h \in \mathcal{Z}_h(\mathcal{G})} \|u - \mathcal{X}_h\|_{\mathcal{X}} \leq \left(1 + \frac{C}{\alpha}\right) \inf_{v_h \in \mathcal{X}_h} \|u - v_h\|_{\mathcal{X}},$$

and the theorem is proved.  $\square$

### 2.3.3 Convergence of collectively compact operators

This material is taken from Kress [193] (see [16] for the original work in this area). First we define collective compactness. As usual,  $X$  denotes a general Hilbert space.

**Definition 2.47** A set  $K = \{K_n : X \rightarrow X, n = 0, 1, 2, \dots\}$  of bounded linear operators is called *collectively compact* if, for each bounded set  $U \subset X$ , the image set

$$K(U) = \{K_n u \mid \text{for all } u \in U, \text{ and } K_n \in K\}$$

is relatively compact (i.e. its closure is compact).

Note that this definition implies that the operators  $\{K_n\}_{n=0}^\infty$  are uniformly bounded. To see this we apply the definition choosing  $U = \{u \in X \mid \|u\|_X = 1\}$ . The image set  $K(U)$  is bounded (since it is relatively compact) and this implies a uniform bound on  $\|K_n\|_{X \rightarrow X}$ ,  $n = 0, 1, 2, \dots$ .

In our applications  $n$  will index a sequence of successively finer finite element meshes (not necessarily nested). We want to estimate the error in the solution as  $n$  tends to  $\infty$ . Part of this process is to verify that the finite element solution operators satisfy the following definition, which corresponds to the standard notion of convergence for a finite element method.

**Definition 2.48** The operators  $\{K_n\}_{n=0}^{\infty}$  are said to *converge pointwise* to an operator  $K : X \rightarrow X$  if, for each  $f \in X$ ,  $K_n f \rightarrow Kf$  in  $X$  as  $n \rightarrow \infty$ .

Recalling now that

$$\|(K_n - K)K\|_{X \rightarrow X} = \sup_{f \in X, \|f\|_X = 1} \|(K_n - K)Kf\|_X,$$

with a similar definition for  $\|(K_n - K)K_n\|_{X \rightarrow X}$ , we have our first result on collectively compact operators (Theorem 10.6 of [193]).

**Lemma 2.49** Suppose  $\{K_n : X \rightarrow X\}_{n=0}^{\infty}$  is a collectively compact set of bounded linear operators and that the operators are pointwise convergent to a compact operator  $K : X \rightarrow X$ . Then

$$\|(K_n - K)K\|_{X \rightarrow X} \rightarrow 0 \text{ and } \|(K_n - K)K_n\|_{X \rightarrow X} \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Before proving this lemma, we recall the following result that follows from the uniform boundedness principle (Corollary 10.4 of [193], but stated for Hilbert spaces).

**Lemma 2.50** Let  $X$  and  $Y$  be Hilbert spaces and let  $A_n : X \rightarrow Y$ ,  $n = 1, 2, \dots$  be a family of bounded linear and pointwise convergent operators with limit operator  $A : X \rightarrow Y$ . Then convergence is uniform on compact subsets  $U$  of  $X$  or, equivalently,

$$\sup_{\varphi \in U} \|A_n \varphi - A \varphi\|_Y \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Using this lemma, the proof of Lemma 2.49 follows from the compactness of  $A$ .

**Proof of Lemma 2.49** We sketch the proof and direct the reader to Kress [193] for details. Let

$$\mathcal{Q} = \{K_n f \mid \|f\|_X = 1, n = 0, 1, \dots\}.$$

Since  $\{K_n\}_{n=0}^{\infty}$  is collectively compact,  $\mathcal{Q}$  is relatively compact. Since the family of operators  $\{K_n\}_{n=0}^{\infty}$  is uniformly bounded and  $\mathcal{Q}$  is relatively compact, by Lemma 2.50 the convergence of  $K_n X \rightarrow K_X$  as  $n \rightarrow \infty$  is uniform for  $X \in \mathcal{Q}$ . Thus, given  $\varepsilon > 0$ , there exists  $N > 0$  such that  $\|K_n X - K_X\|_X < \varepsilon$  for all  $n > N$  and all  $X \in \mathcal{Q}$ . But  $X \in \mathcal{Q}$  is arbitrary, so we see that for every  $\xi \in X$  such that  $\|\xi\|_X = 1$  we have  $\|(K_n - K)K_m \xi\|_X < \varepsilon$  for all  $m$ , for all  $\xi \in X$  and for all  $n > N$ . Taking  $m = n$  we get the required result. Using  $K$  in place of  $K_n$  in the definition of  $\mathcal{Q}$  (recalling that  $K$  is assumed to be compact), and proceeding as above, proves the other case.  $\square$

Now suppose  $K$  is such that for every  $F \in X$  there is a unique solution  $u \in X$  of the equation (2.19)

$$(I + K)u = F,$$

and  $\|(I + K)^{-1}\|_{X \rightarrow X} < \infty$ . We wish to show that the discrete problem of finding  $u_n \in X$  such that

(2.20)

$$(I + K_n)u_n = \mathcal{F}_n,$$

for  $F_n \in X$  has a unique solution, and derive an error estimate. This is done in the following theorem.

**Theorem 2.51** *Let  $K: X \rightarrow X$  be a compact operator such that  $(I + K)$  is invertible with bounded inverse and let  $u$  satisfy (2.19). Suppose  $\{K_n: X \rightarrow X\}_{n=0}^{\infty}$  is a collectively compact set of bounded linear operators that are pointwise convergent to  $K$ . Then for all  $n$  large enough  $(I + K_n)^{-1}$  exists and its uniformly bounded independent of  $n$  as a map from  $X$  to  $X$  so that (2.20) has a unique solution  $u_n \in X$ . In addition the following error estimate holds:*

$$\|u - u_n\|_X \leq C(\|\mathcal{F} - \mathcal{F}_n\|_X + \|(K - K_n)u\|_X).$$

**Proof** This theorem is from [119] and the proof is a slight modification of the proof of Theorem 10.8 of [193]. We start by showing that  $(I + K)^{-1}$  exists. Since  $(I + K)^{-1}$  is bounded, we may define  $B_n = I - (I + K)^{-1}K_n$ . Then

$$B_n(I + K_n) = I + (I + K)^{-1}(K - K_n)K_n.$$

Letting  $S_n = (I + K)^{-1}(K - K_n)K_n$ , we know that

$$\|S_n\|_{X \rightarrow X} \leq C\|(K - K_n)K_n\|_{X \rightarrow X} < 1$$

for  $n$  sufficiently large (by the previous lemma). Hence, using the Neumann series (see Theorem 2.27),  $(I + S_n)^{-1}$  exists and is bounded as follows

$$\|(I + S_n)^{-1}\|_{X \rightarrow X} \leq \frac{1}{1 - \|S_n\|_{X \rightarrow X}}.$$

Since  $I + S_n$  is invertible and  $B_n(I + K_n) = I + S_n$ , we see that  $(I + K_n)$  must be injective. Using the fact that each operator  $K_n$  is compact (the entire set is collectively compact) together with the Fredholm alternative (Theorem 2.33) shows that  $(I + K_n)^{-1}$  exists. Therefore,  $(I + K_n)^{-1} = (I + S_n)^{-1}B_n$ , so that

$$\|(I + K_n)^{-1}\|_{X \rightarrow X} \leq \frac{1 + \|(I + K)^{-1}K_n\|_{X \rightarrow X}}{1 - \|(I + K)^{-1}(K_n - K)K_n\|_{X \rightarrow X}},$$

and uniform boundedness of the inverse is verified. Using the equations for  $u_n$  and  $u$ , we can write

$$\begin{aligned} (I + K_n)(u - u_n) &= (I + K_n)u - \mathcal{F}_n = (I + K)u + (K_n - K)u - \mathcal{F}_n \\ &= (\mathcal{F}_n - \mathcal{F}) + (K_n - K)u. \end{aligned}$$

Hence, by the uniform boundedness of  $(I + K_n)^{-1}$ , we have  $\|u - u_n\|_X \leq C\|(\mathcal{F} - \mathcal{F}_n) + (K_n - K)u\|_X$  and use of the triangle inequality proves the desired result.

□

Note that we have not proved the stronger result of norm convergence of  $(I + K_n)^{-1}$  to  $(I + K)^{-1}$  as  $n \rightarrow \infty$ . Under further restrictions on the operators  $K_n$  and their adjoints, it is possible to conclude this stronger result (see [285]).

### 2.3.4 Eigenvalue estimates

In this section, we shall summarize some known results for the approximation of eigenvalue problems. Since our main finite element convergence result is proved using a pointwise convergent set of collectively compact discrete operators, we shall use the appropriate theory of Osborn [246]. However, because eigenvalue problems are not a focus of the book, we shall not provide proofs.

Suppose  $X$  is a Hilbert space and  $K : X \rightarrow X$  is a self-adjoint and compact operator. Then we know from the Hilbert–Schmidt theory (Section 2.2.5) that the problem of finding  $\mu \in \mathbb{R}$  and  $u \in X$ ,  $u \neq 0$ , such that(2.21)

$$Ku = \mu u$$

has a solution (in fact, in the case of Maxwell's equations, infinitely many solutions).

Now suppose that  $\Lambda$  is a countable set having only zero as the limit point. Suppose also that  $K = \{K_b : X \rightarrow X\}_{b \in \Lambda}$  is a set of collectively compact, self-adjoint operators, and that the operators converge pointwise to the operator  $K$  above (which is compact and self-adjoint).

The discrete eigenvalue problem is then to find  $\mu_b$  and  $u_b \in X$ ,  $u_b \neq 0$ , such that(2.22)

$$K_b u_b = \mu_b u_b .$$

We now wish to know under what conditions the eigenvalues and eigenvectors for problem (2.22) converge to the true eigenvalues and eigenvectors from (2.21).

Let us suppose that  $\mu$  is an eigenvalue of  $K$  of multiplicity  $m$ . Osborn [246] proves the following theorem (i.e. essentially Theorem 3 of Osborn' paper, but we have collected other results in the paper and assume that the operators are self-adjoint):

**Theorem 2.52** Suppose  $\varepsilon > 0$  is such that the disk of radius  $\varepsilon$  about  $\mu$  contains no other eigenvalues of  $K$ . Then for  $b$  small enough the disk of radius  $\varepsilon$  centered at  $\mu$  contains precisely  $m$  eigenvalues of the discrete problem denoted by  $\mu_{b,j}$ ,  $j = 1, \dots, m$ . The dimension of the eigenspace corresponding to  $\mu$ , denoted by  $E(\mu)$ , is equal to that of  $\bigoplus_{j=1}^m E(\mu_{b,j})$ . Finally, for  $1 \leq j \leq m$  there is a constant  $C$  such that(2.23)

$$|\mu - \mu_{b,j}| \leq C \left\{ \sum_{l,j=1}^m |((K - K_b) u_j, u_l)_X| + \| (K - K_b) \|_{E(\mu)}^2 \right\} .$$

Here  $\{u_j\}_{j=1}^m$  is an  $X$  orthonormal basis for  $E(\mu)$  and  $(K - K_b)|_{E(\mu)}$  is the restriction of  $(K - K_b)$  to  $E(\mu)$ .

**Remark 2.53** Osborn also provides an estimate for the distance of  $E(\mu)$  to  $\bigoplus_{j=1}^m E(\mu_{b,j})$ .

# 3 SOBOLEV SPACES, VECTOR FUNCTION SPACES AND REGULARITY

## 3.1 Introduction

The variational theory of Maxwell's equations is built on Sobolev spaces of scalar and vector functions. In this chapter we shall summarize some basic results concerning such function spaces. We start with Sobolev spaces of scalar functions. The reader is assumed to be familiar with the basic concepts for these spaces, so the first part of the chapter only serves to define some notation and collect some standard results in a convenient place.

In the latter part of the chapter we discuss some Sobolev spaces of vector-valued functions appropriate for analyzing Maxwell's equations. These spaces are a little less standard, so we shall give more details. In particular, we note that these spaces have rather delicate and surprising properties concerning the density of smooth functions. We shall also discuss various decompositions of vector fields (in particular, the Helmholtz decomposition of a vector function into a curl-free and a divergence-free part) and prove a critical regularity result for solutions of Maxwell's equations. Related to this, we discuss scalar and vector potentials. The basic reference for this material is the excellent book of Girault and Raviart [143].

## 3.2 Standard Sobolev spaces

We start by defining some standard spaces of functions (see, e.g. [215]). For any open set  $\Omega \subset \mathbb{R}^N$ ,  $N = 1, 2, 3$  we define

$C^k(\Omega)$ : the set of  $k$  times continuously differentiable functions on  $\Omega$ ;

$C_0^k(\Omega)$ : the set of functions  $\varphi \in C^k(\Omega)$  having compact support in  $\Omega$

$C_c^k(\Omega)$ : the set of functions in  $C^k(\Omega)$  which have bounded and uniformly continuous derivatives up to order  $k$  on  $\Omega$  (i.e. the restrictions of functions in  $C_0^k(\mathbb{R}^N)$  to  $\Omega$ ); and  $L^p(\Omega)$ ,  $1 \leq p < \infty$ : the set of functions  $\varphi$  on  $\Omega$  for which  $|\varphi|^p$  is integrable. More exactly, functions  $\varphi$  such that

$$\int_{\Omega} |\varphi|^p dV < \infty .$$

The most important case here is  $p = 2$ , which is the set of all square-integrable functions on  $\Omega$ .

We use the standard multi-index notation for derivatives. If

$$\alpha = (\alpha_1, \alpha_2, \dots, \alpha_N)^T \in \mathbb{Z}_+^N,$$

where  $\mathbb{Z}_+$  is the set of non-negative integers, we set  $|\alpha|_1 = \sum_{i=1}^N |\alpha_i|$  and for  $\varphi \in C^{|\alpha|_1}(\Omega)$  we define

$$\frac{\partial^\alpha \varphi}{\partial x^\alpha} = \frac{\partial^{|\alpha|_1} \varphi}{\partial x_1^{\alpha_1}, \dots, \partial x_N^{\alpha_N}}.$$

The space of distributions, denoted by  $C_0^\infty(\Omega)'$ , is the dual space of  $C_0^\infty(\Omega)$  in the sense that a linear functional  $T: C_0^\infty(\Omega) \rightarrow \mathbb{C}$  is contained in  $C_0^\infty(\Omega)'$ , provided that for every compact set  $K \subset \Omega$  there exist constants  $C$  and  $k$  such that

$$|T(\varphi)| \leq C \sum_{|\alpha|_1 \leq k} \sup_K |D^\alpha \varphi|$$

for all  $\varphi \in C_0^\infty(\Omega)$  (see [298, 215] for a more detailed discussion of distributions).

The standard definition of the distributional derivative of a function  $\varphi \in C_0^\infty(\Omega)'$  again uses the multi-index notation. The distributional derivative  $\partial^\alpha \varphi \in C_0^\infty(\Omega)'$  of a function  $\varphi \in C_0^\infty(\Omega)'$  is the unique distribution that satisfies (3.1)

$$\left( \frac{\partial^\alpha \varphi}{\partial x^\alpha}, \psi \right) = (-1)^{|\alpha|_1} \left( \varphi, \frac{\partial^\alpha \psi}{\partial x^\alpha} \right) \text{ for all } \psi \in C_0^\infty(\Omega).$$

For functions  $\varphi \in C^m(\Omega)$ , the distributional and standard (or strong) derivatives of  $\varphi$  agree provided  $|\alpha|_1 \leq m$ . Of course, for functions in  $L^2(\Omega)$ , the derivative must, in general, be understood in the distributional sense.

An open, connected set in  $\mathbb{R}^N$ ,  $N = 1, 2, 3$ , will be referred to as a domain. The fundamental Sobolev spaces are denoted  $W^{s,p}(\Omega)$ , where  $s \in \mathbb{Z}_+$ ,  $1 \leq p < \infty$  and  $\Omega$  is a domain in  $\mathbb{R}^N$ . These spaces are defined by

$$W^{s,p}(\Omega) = \left\{ \varphi \in L^p(\Omega) \mid \partial^\alpha \varphi \in L^p(\Omega) \text{ for all } |\alpha|_1 \leq s \right\}.$$

Associated with this space is the norm (3.2)

$$\|\varphi\|_{W^{s,p}(\Omega)} = \left( \sum_{|\alpha|_1 \leq s} \int_\Omega |\partial^\alpha \varphi(x)|^p dV(x) \right)^{1/p}.$$

The corresponding semi-norm, used later in our interpolation analysis of finite element methods, is (3.3)

$$\|\varphi\|_{W^{s,p}(\Omega)} = \left( \sum_{|\alpha|_1 = s} \int_\Omega |\partial^\alpha \varphi(x)|^p dV(x) \right)^{1/p}.$$

A particularly important case occurs when  $p = 2$ , and the majority of our use of these spaces will be in this case.

An alternative definition of Sobolev spaces for  $p = 2$  is to define the spaces  $H^s(\Omega)$ ,  $s \in \mathbb{Z}_+$ , using Fourier transforms. Rather than spending time to introduce this concept, we note that for  $\Omega = \mathbb{R}^N$ ,  $N = 2, 3$ , it can be shown that  $H(\mathbb{R}^N) = W^{s,2}(\mathbb{R}^N)$  (see Theorem 3.16 of [215]). Then for a bounded domain we define

$$H^s(\Omega) = \left\{ u \in C_0^\infty(\Omega)' \mid u = U|_\Omega \text{ for some } U \in W^{s,2}(\mathbb{R}^N) \right\}.$$

The norm on this space is defined using an auxiliary space [215],

$$H_{\mathbb{R}^N \setminus \overline{\Omega}}^s = \left\{ u \in W^{s,2}(\mathbb{R}^N) \mid \text{support}(u) \subset \mathbb{R}^N \setminus \overline{\Omega} \right\}.$$

Since this is a closed subspace of  $W^{s,2}(\mathbb{R}^N)$ , the Projection Theorem 2.9 guarantees the existence of a projection  $P: W^{s,2}(\mathbb{R}^N) \rightarrow H_{\mathbb{R}^N \setminus \overline{\Omega}}^s$  and we can define the  $H(\Omega)$  inner product by

$$(u, v)_{H^s(\Omega)} = ((I - P)U, (I - P)V)_{H^s(\mathbb{R}^N)}, \quad \text{where } u = U|_\Omega \text{ and } v = V|_\Omega.$$

As we shall see (Theorem 3.2), if the domain  $\Omega$  has a sufficiently well-behaved boundary we have  $H(\Omega) = W^{s,2}(\Omega)$ , with  $\|u\|_{H(\Omega)} \equiv \|u\|_{W^{s,2}(\Omega)}$ , so the two spaces and their properties can be used interchangeably.

Functions in the Sobolev spaces discussed so far do not satisfy any particular boundary condition. To define spaces of functions that satisfy a Dirichlet boundary condition (i.e. vanish on the boundary) we proceed as follows. We use the closure of  $C_0^\infty(\Omega)$  in the appropriate norm to define(3.4)

$$W_0^{s,p}(\Omega) = \text{closure of } C_0^\infty(\Omega) \text{ in the } W^{s,p}(\Omega) \text{ norm.}$$

Again, the special case  $p = 2$  deserves its own notation, so we set(3.5)

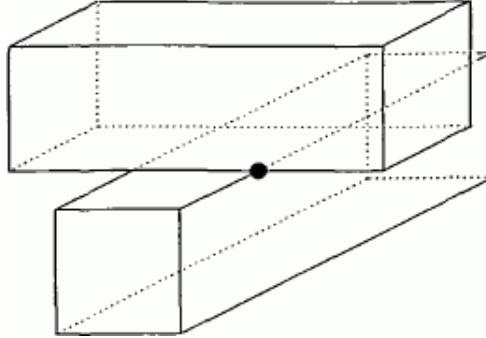
$$H_0^s(\Omega) = W_0^{s,2}(\Omega).$$

As we shall see, functions  $u$  in  $H_0^1(\Omega)$  satisfy the boundary condition  $u = 0$  on the boundary of  $\Omega$  (denoted  $\partial\Omega$ ) in an appropriate sense.

Most of our work concerns Maxwell's equations on bounded domains in  $\mathbb{R}^3$  (unbounded domains will be reduced to bounded domains by a truncation procedure). On bounded domains, the properties of Sobolev spaces are determined by the smoothness or regularity of the boundary. We shall mainly consider one case in this book: Lipschitz polyhedral domains. Because we wish to use this in  $\mathbb{R}^2$  and  $\mathbb{R}^3$ , we make the definition for a domain in  $\mathbb{R}^N$ ,  $N = 2, 3$ .

**Definition 3.1** The boundary  $\partial\Omega$  of a bounded domain  $\Omega$  in  $\mathbb{R}^N$  is *Lipschitz continuous* if for every  $x \in \partial\Omega$  there is an open set  $O \subset \mathbb{R}^N$  with  $x \in O$  and

Fig. 3.1. An example of a simple polyhedral domain that is not Lipschitz. At the point marked • the surface is not the graph of a function.



an orthogonal coordinate system with coordinate  $\zeta = (\zeta_1, \dots, \zeta_N)$  having the following properties. There is a vector  $a \in \mathbb{R}^N$  with

$$\mathcal{O} = \{ \zeta \mid -a_j < \zeta_j < a_j, 1 \leq j \leq N \}$$

and a Lipschitz continuous function  $\varphi$  defined on

$$\mathcal{O}' = \{ \zeta \in \mathbb{R}^{N-1} \mid -a_j < \zeta_j < a_j, 1 \leq j \leq N-1 \}$$

with  $|\varphi(\zeta)| \leq a_N/2$  for all  $\zeta' \in \mathcal{O}'$  such that

$$\begin{aligned} \Omega \cap \mathcal{O} &= \{ \zeta \mid \zeta_N < \varphi(\zeta'), \zeta \in \mathcal{O}' \} \text{ and} \\ \partial \Omega \cap \mathcal{O} &= \{ \zeta \mid \zeta_N = \varphi(\zeta'), \zeta \in \mathcal{O}' \}. \end{aligned}$$

This form of the definition is from [151]. We shall simply say that the domain  $\Omega$  is Lipschitz when we mean that it has a Lipschitz continuous boundary.

The reason for using Lipschitz polyhedral domains is that they can be covered by a mesh of tetrahedra. This makes the presentation of the finite element method easier, but introduces difficulties with respect to the theoretical aspects of existence, uniqueness and regularity of solutions of Maxwell's equations. In particular, Lipschitz polyhedral domains can have reentrant edges and corners that strongly influence the regularity of the solutions of Maxwell's equations. Many common polyhedra are Lipschitz domains, but some rather simple looking domains are not Lipschitz. For example, the crossed bricks shown in Fig. 3.1 is a polyhedron that is not Lipschitz [215].

Sometimes we shall use boundaries in  $C^1$  by which we mean that Definition 3.1 holds with maps  $\varphi \in C^1(\mathcal{O}')$  for each  $\mathcal{O}'$  in the definition. For example, we shall consider a spherical domain which is, of course, smooth having a  $C^\infty$  boundary.

One key property of a Lipschitz domain is that it has a well-defined unit outward normal  $v$  at almost every point on  $\partial\Omega$  [237]. By a unit normal, we mean a normal vector  $v$  such that  $|v| = 1$ . We can now state the following theorem showing the equivalence of  $H^k(\Omega)$  and  $W^{k,2}(\Omega)$ .

**Theorem 3.2** Let  $\Omega$  be a bounded Lipschitz domain in  $\mathbb{R}^N$ . Then the following results hold:

- (1)  $C^\infty(\Omega)$  is dense in  $W^{s,p}(\Omega)$  for  $s \in \mathbb{Z}_+$  and  $p \in \mathbb{R}$ ,  $1 \leq p < \infty$ .
- (2) If  $s \in \mathbb{Z}_+$ ,  $s \geq 1$  and  $1 < p < \infty$  then there exists a continuous linear extension operator  $\Pi$  from  $W^{s,p}(\Omega)$  to  $W^{s,p}(\mathbb{R}^N)$  with the property that

$$(\Pi u)\Big|_{\Omega} = u \text{ for all } u \in W^{s,p}(\Omega).$$

If  $p = 2$ , the operator exists for  $s \geq 0$ .

- (3)  $H(\Omega) = W^{s,2}(\Omega)$ ,  $s \in \mathbb{Z}_+$ , with equivalent norms.

The proof of the first part of this theorem can be found in [237]. The second part, called the Calderon extension theorem, is discussed extensively and proved in [2] (see also Theorem A.4 of [215]). Note that since  $\Pi$  is a continuous linear operator, it is bounded, so there exists a constant  $C$  independent of  $u$  such that

$$\|\Pi u\|_{W^{s,p}(\mathbb{R}^N)} \leq C \|u\|_{W^{s,p}(\Omega)} \text{ for all } u \in W^{s,p}(\Omega).$$

Theorem 3.2(1) asserts the density of  $C^\infty(\Omega)$  in  $W^{s,p}(\Omega)$ . This is important since it sometimes allows us to prove results for smooth functions and extend them by limiting arguments to more general functions. It also allows us to conclude the density of suitable finite element spaces in  $W^{s,p}(\Omega)$ . The next result is a non-standard density result needed for the latter reason, and it is proved in [35].

**Lemma 3.3** Define the space

$$\mathcal{Y} = \left\{ p \in H^1(\Omega) \mid p|_{\partial\Omega} \in H^1(\partial\Omega) \right\}$$

with the graph norm  $\|p\|_{\mathcal{Y}}^2 = \|p\|_{H^1(\Omega)}^2 + \|p\|_{H^1(\partial\Omega)}^2$ . Then  $C^\infty(\Omega)$  is dense in  $\mathcal{Y}$ .

In general,  $C_0^\infty(\Omega)$  is not dense in  $W^{s,p}(\Omega)$  and then  $W_0^{k,p}(\Omega)$  is a proper subset of  $W^{s,p}(\Omega)$ . One useful case when  $C_0^\infty(\Omega)$  is dense is given in the following lemma, which is proved in [298].

**Lemma 3.4** If  $\Omega$  is a bounded Lipschitz domain,  $C_0^\infty(\Omega)$  is dense in  $L^2(\Omega)$ .

The next results are of critical importance in the error analysis of finite element methods. We say that  $W^{s,p}(\Omega)$  is imbedded in a space  $X$  and write  $W^{s,p}(\Omega) \hookrightarrow X$  if  $W^{s,p}(\Omega)$  is a subset of  $X$  and if the identity map  $I$  from  $W^{s,p}(\Omega)$  to  $X$  is continuous. This is equivalent to saying that there exists a constant  $C$  independent of  $u$  such that  $\|Iu\|_x \leq C \|u\|_{W^{s,p}(\Omega)}$  for all  $u \in W^{s,p}(\Omega)$ .

If  $\Omega'$  denotes the intersection of an  $l$ -dimensional hyper-plane with  $\Omega$ , we shall present conditions under which  $W^{m+j,p}(\Omega)$  is imbedded in  $W^{m,p}(\Omega')$ . Here the imbedding has to be interpreted carefully. By Theorem 3.2(1), each element  $u \in W^{m+j,p}(\Omega)$  is a limit of functions  $u_n \in C^\infty(\Omega)$ ,  $n = 1, 2, \dots$ . These functions have a well-defined restriction or trace on  $\Omega'$ . The imbedding result means that the functions  $u_n|_{\Omega'}$  converge to a function in  $W^{m,p}(\Omega')$ .

The imbedding of  $W^{m+j,p}(\Omega)$  in a space of continuous functions is understood in the sense that there is a member of the equivalence class of functions  $u \in W^{m+j,p}(\Omega)$  with the required continuity. The following statement of the famous *Sobolev imbedding theorem* is from [2].

**Theorem 3.5** Let  $\Omega \subset \mathbb{R}^N$  be a bounded domain with Lipschitz continuous boundary and suppose  $m, j$  are non-negative integers and  $1 \leq l \leq N$ . Let  $p \in \mathbb{R}$ , with  $1 \leq p \leq \infty$ . Then the following imbeddings hold:

(1) Suppose  $mp < N$  and  $N - mp < l \leq N$ . Then

$$W^{j+m,p}(\Omega) \hookrightarrow W^{j,q}(\Omega^l), \quad p \leq q \leq lp / (N - mp).$$

(2) Suppose  $mp = N$  then for  $1 \leq l \leq N$  and  $p = q < \infty$ . Then

$$W^{j+m,p}(\Omega) \hookrightarrow W^{j,q}(\Omega^l).$$

(3) Suppose  $mp > N \geq (m - 1)p$ . Then  $W^{j+m,p}(\Omega) \hookrightarrow C(\Omega)$ .

This theorem holds in much greater generality than the above statement. For a detailed discussion, and proof, see [2].

An imbedding is said to be compact if the imbedding operator  $I$  is compact. The following theorem from [2] summarizes some results on when imbeddings are compact.

**Theorem 3.6** Let  $\Omega \subset \mathbb{R}^N$  be a bounded Lipschitz domain and let  $\Omega_0$  be any subdomain of  $\Omega$  (we allow  $\Omega_0 = \Omega$ ). Let  $\Omega_0^l$  denote the intersection of  $\Omega_0$  with an  $l$ -dimensional hyper-plane in  $\mathbb{R}^N$ . Let  $j, m$  be integers with  $m \geq 1$  and  $j \geq 0$  and let  $p \in \mathbb{R}$  with  $1 \leq p < \infty$ . Then the following imbeddings are compact:

(1) If  $mp \leq N$  then (3.6)

$$\begin{aligned} W^{j+m,p}(\Omega) &\hookrightarrow W^{j,q}(\Omega_0^l), \quad 0 < N - mp < l \leq N \text{ and} \\ &1 \leq q \leq lp / (N - mp), \end{aligned} \tag{3.7}$$

$$\begin{aligned} W^{j+m,p}(\Omega) &\hookrightarrow W^{j,q}(\Omega_0^l), \quad mp = N, 1 \leq l \leq N \text{ and} \\ &1 \leq q < \infty, \end{aligned}$$

(2) If  $mp > N$  then  $W^{j+m,p}(\Omega) \hookrightarrow C(\Omega_0)$ .

For a proof of this theorem, and a discussion of its history, see [2]. The special case of (3.6) when  $m = 1, p = 2, j = 0, l = N$  and  $\Omega_0 = \Omega$  states that  $H^1(\Omega)$  is compactly imbedded in  $L^2(\Omega)$  for  $N = 2, 3$ . This observation underlies the analysis of the Helmholtz equation.

Unfortunately, in the analysis of boundary values of functions, and in discussing the regularity of solutions of Maxwell's equations, it is necessary to use Sobolev spaces of fractional order. Following [237], we define the spaces  $W^{s,p}(\Omega)$ ,  $1 \leq p < \infty$ ,  $s \in \mathbb{R}$  and  $s \geq 0$  as follows. Let  $m \in \mathbb{Z}_+$  and suppose  $s = m + \sigma$ ,

where  $\sigma \in \mathbb{R}$  and  $0 < \sigma < 1$ . Then  $W^{s,p}(\Omega)$  is defined to be the space of distributions  $u \in C_0^\infty(\Omega)'$  such that  $u \in W^{m,p}(\Omega)$  and

$$\int_{\Omega} \int_{\Omega} \frac{|\partial^\alpha u(x) - \partial^\alpha u(y)|^p}{|x-y|^{N\sigma p}} dV(x)dV(y) < \infty \text{ for all } |\alpha| = m.$$

The norm for this space is

$$\begin{aligned} \|u\|_{W^{s,p}(\Omega)} = & \left\{ \|u\|_{W^{m,p}(\Omega)}^p \right. \\ & \left. + \sum_{|\alpha|=m} \int_{\Omega} \int_{\Omega} \frac{|\partial^\alpha u(x) - \partial^\alpha u(y)|^p}{|x-y|^{N+\sigma p}} dV(x)dV(y) \right\}^{1/p}. \end{aligned}$$

With this norm, the space  $W^{s,p}(\Omega)$  is a separable, reflexive Banach space for  $1 < p < \infty$  and  $s \in \mathbb{R}$  with  $s \geq 0$  [237]. As in the case of integer values of  $s$ , we define  $W_0^{s,p}(\Omega)$  to be the closure of  $C_0^\infty(\Omega)$  in  $W^{s,p}(\Omega)$ . We still have  $H(\Omega) = W^{s,2}(\Omega)$ ,  $s \geq 0$ .

The imbedding theorems for fractional-order spaces are not as powerful as for integer-order spaces. The following results will be needed for our treatment of Maxwell's equations.

**Theorem 3.7** *Let  $\Omega$  be a bounded Lipschitz domain. Then, if  $0 \leq t < s$  such that  $s - 3/p = t - 3/q$ , the imbedding  $W^{s,p}(\Omega) \hookrightarrow W^{t,q}(\Omega)$  holds. Furthermore, if  $0 \leq t < s < \infty$  and  $p = q = 2$  the imbedding is compact.*

The imbedding result is proved in [237] in a special case and is stated this way in Theorem 1.4.4.1 of [151]. The compact imbedding result is in [298] and also in Theorem 3.27 of [215]. In addition, Theorem 3.2 holds for fractional-order spaces (cf. [151] for a discussion of fractional-order Sobolev spaces on Lipschitz domains). We should point out that the results quoted for fractional-order spaces are a very small selection of the known results (cf. [2]).

We shall denote by  $H^1(\Omega)$  the dual space of  $H_0^1(\Omega)$  with the usual dual norm.

### 3.2.1 Trace spaces

We have one further topic in basic Sobolev space theory to discuss, in particular, the way in which boundary values or traces of functions are handled. First we have to define what we mean by Sobolev spaces on the boundary  $\partial\Omega$  of  $\Omega$ . We follow [151] and recall from Definition 3.1 that the boundary  $\partial\Omega$  of  $\Omega$  is such that for every  $x \in \partial\Omega$  there is a Lipschitz continuous map  $\varphi : O' \subset \mathbb{R}^{N-1} \rightarrow \mathbb{R}$  such that

$$\partial\Omega \cap O = \{ \zeta' = \varphi(\zeta', \varphi(\zeta')) \mid \zeta \in O' \}$$

and thus locally  $\partial\Omega$  is an  $(N-1)$ -dimensional hyper-surface in  $\mathbb{R}^N$ . We define  $\varphi$  via  $\varphi(\zeta) = (\zeta', \varphi(\zeta))$ . Then  $\varphi^{-1}$  exists and is Lipschitz continuous on  $\varphi(O')$ . This motivates the following definition:

**Definition 3.8** Let  $\Omega \subset \mathbb{R}^N$  be a bounded Lipschitz domain with boundary  $\partial\Omega$ . A distribution  $u$  on  $\partial\Omega$  belongs to  $W^{s,p}(\partial\Omega)$  for  $|s| \leq 1$  if the composition  $u \circ \varphi \in W^{s,p}(O' \cap \varphi^{-1}(\partial\Omega \cap O))$  for all possible  $O$  and  $\varphi$  fulfilling the criteria of Definition 3.1.

To define a norm on  $W^{s,p}(\partial\Omega)$ , we let  $(\mathcal{O}_j, \varphi_j)_{j=1}^J$  be any atlas for  $\partial\Omega$  such that the pairs  $(\mathcal{O}_j, \varphi_j)_{j=1}^J$  satisfy the conditions of Definition 3.1. Then

$$\|u\|_{W^{s,p}(\partial\Omega)} = \left( \sum_{j=1}^J \|u \circ \varphi_j\|_{W^{s,p}(\mathcal{O}'_j \cap \varphi_j^{-1}(\partial\Omega \cap \mathcal{O}_j))}^p \right)^{1/p}.$$

In the particular case  $s \in [0, 1)$  and  $\Omega \in \mathbb{R}^N$ , this definition is equivalent to

$$\|u\|_{W^{s,p}(\partial\Omega)} = \left( \int_{\partial\Omega} |u|^p d\sigma + \int_{\partial\Omega} \int_{\partial\Omega} \frac{|u(x) - u(y)|^p}{|x - y|^{N-1+sp}} dA(x) dA(y) \right)^{1/p},$$

where  $dA$  is the surface measure on  $\partial\Omega$ . As usual,  $H^s(\partial\Omega) = W^{s,2}(\partial\Omega)$  for  $0 \leq s \leq 1$ .

The next theorem shows that, provided a function  $u$  is sufficiently smooth, it is possible to define the boundary value of  $u$  on  $\partial\Omega$ . This boundary value is called the trace of  $u$  on  $\partial\Omega$ . Of course, for any function  $u \in C^\infty(\Omega)$ , the evaluation of  $u$  on  $\partial\Omega$  is well-defined. Thus we define the trace operator  $\gamma_0$  for such a function by (3.8)

$$\gamma_0(u) = u / \partial\Omega .$$

**Theorem 3.9** (Trace theorem) *Let  $\Omega$  be a bounded Lipschitz domain. Then, provided  $1/p < s \leq 1$ , the mapping  $\gamma_0$  defined on  $C^\infty(\Omega)$  by (3.8) has a unique continuous extension as a linear operator from  $W^{s,p}(\Omega)$  onto  $W^{s-1/p,p}(\partial\Omega)$ . Moreover, (3.9)*

$$W_0^{1,p}(\Omega) = \left\{ u \in W^{1,p}(\Omega) \mid \gamma_0(u) = 0 \right\}.$$

This theorem is proved for  $s = 1$  in [237] and its extension to general  $s$  is discussed in [151]. Note that (3.9) implies that the space  $W_0^{1,p}(\Omega)$ ,  $p > 1$ , which was defined by density in (3.4) consists of functions that satisfy the homogeneous Dirichlet boundary condition on  $\partial\Omega$ . An alternative definition for  $W_0^{1,p}(\Omega)$ ,  $p > 1$ , is

$$W_0^{1,p}(\Omega) = \left\{ u \in L^p(\Omega) \mid \nabla u \in (L^p(\Omega))^3 \text{ and } \gamma_0(u) = 0 \right\},$$

where  $\nabla$  denotes the gradient, which is the operator from  $C_0^\infty(\Omega)'$  to  $(C_0^\infty(\Omega)')^N$  defined by (3.10)

$$\nabla u = \left( \frac{\partial u}{\partial x_1}, \dots, \frac{\partial u}{\partial x_N} \right)^T.$$

The most important trace spaces for us will be  $H^{1/2}(\partial\Omega) = W^{1/2,2}(\partial\Omega)$  and its dual space  $H^{-1/2}(\partial\Omega)$ . The norm on this space is the usual dual norm. In particular for any Lipschitz surface  $S$  we define

$$\langle f, g \rangle_S = \int_S f \bar{g} dA_1 .$$

The norm on  $H^{1/2}(\partial\Omega)$  can then be written(3.11)

$$\|f\|_{H^{-1/2}(\partial\Omega)} = \sup_{g \in H^{1/2}(\partial\Omega)} \frac{|\langle f, g \rangle_{\partial\Omega}|}{\|g\|_{H^{1/2}(\partial\Omega)}},$$

where we have used the fact that  $H^{1/2}(\partial\Omega)$  can also be characterized as the completion of  $L^2(\partial\Omega)$  in a suitable norm to show that we may identify the duality pairing with the  $L^2(\partial\Omega)$  inner product (see page 98 of McLean [215] for details).

We shall also require, fortunately infrequently, to use trace spaces for  $s > 1$ . In this case the definition is not as natural as the one given previously. Keeping in mind our desire to study boundary value problems, we use the following definition, which agrees with the previous one for  $0 \leq s \leq 1$  [12]. For  $s > 1$  we define the normed space(3.12)

$$H^S(\partial\Omega) = \left\{ u \in L^2(\partial\Omega) \mid u = U|_{\partial\Omega} \text{ for some } U \in H^{s+1/2}(\Omega) \right\},$$

with norm given by

$$\|u\|_{H^S(\partial\Omega)} = \inf_{U \in H^{s+1/2}(\Omega), u = U|_{\partial\Omega}} \|U\|_{H^{s+1/2}(\Omega)} .$$

In particular,  $\|u\|_{H^s(\partial\Omega)} = \|U\|_{H^{s+1/2}(\Omega)}$ , where  $U \in H^{s+1/2}(\Omega)$  satisfies  $U|_{\partial\Omega} = u$  and

$$(U, \varphi)_{H^{s+1/2}(\Omega)} = 0 \text{ for all } \varphi \in H^{s+1/2}(\Omega) \cap H_0^1(\Omega) .$$

This function exists by the Lax–Milgram Lemma 2.21. Thus we can see that  $H^s(\partial\Omega)$  is complete since  $H^{s+1/2}(\Omega)$  is complete, and is, in fact, a Hilbert space.

This definition has the advantage that we know any function in  $H^s(\partial\Omega)$ ,  $s > 1$ , can be extended to a function  $U \in H^{s+1/2}(\Omega)$ , so it is well-suited to the study of boundary value problems. In addition, the boundary data for electromagnetic scattering problems is often given as the trace of a smooth vector field. The main disadvantage of the definition is that it is difficult to determine when a particular function  $g$  defined on  $\partial\Omega$  is in  $H^s(\partial\Omega)$ , since no explicit norm is available. Fortunately, for a Lipschitz polyhedron, we have the following result, which is part of a much more general result in [41]. For a simple proof when  $\partial\Omega$  is a cube, see [167].

**Theorem 3.10** *Let  $\Omega$  be a bounded Lipschitz polyhedron with boundary  $\partial\Omega$ . Suppose  $\partial\Omega$  has faces  $\partial\Omega_j$ ,  $1 \leq j \leq J$ , and suppose  $g \in L^2(\partial\Omega)$  is such that*

- $g \in H^s(\partial\Omega)$  for  $1 < s < \frac{3}{2}$ ;
- if two faces  $\partial\Omega_j$  and  $\partial\Omega_{j'}$  meet at an edge  $e_{j,j'}$  then  $g|_{\partial\Omega_j} = g|_{\partial\Omega_{j'}}$  on  $e_{j,j'}$ .

Then  $g \in H^s(\partial\Omega)$  (i.e. an extension to  $H^{s+1/2}(\Omega)$  exists).

We shall also need the following technical lemma from [143]. To state this lemma we define

$$L^2_{\text{loc}}(\Omega) = \left\{ p \in L^2(\mathcal{O}) \text{ for all compact subdomains } \mathcal{O} \subset \Omega \right\}.$$

**Lemma 3.11** Let  $\Omega$  be a bounded, Lipschitz and connected domain. Suppose  $p \in L^2_{\text{loc}}(\Omega)$  and  $\nabla p \in H^1(\Omega)^3$ . Then  $p \in L^2(\Omega)$ .

### 3.3 Regularity results for elliptic equations

This is a vastly technical subject. All we shall do is summarize some results that will be used later in the analysis of finite element methods for Maxwell's equations. Suppose  $\Omega$  is a bounded Lipschitz domain with boundary  $\partial\Omega = \Gamma_D \cup \Gamma_N$ , where  $\Gamma_N \cap \Gamma_D = \emptyset$ . Let  $\nu$  denote the unit outward normal to  $\Gamma$ . We are interested in conditions under which the problem of finding  $\varphi$  such that

$$\begin{aligned} -\Delta \varphi + c \varphi &= f \quad \text{in } \Omega, \\ \varphi &= \mu_D \quad \text{on } \Gamma_D, \\ \frac{\partial \varphi}{\partial \nu} &= \mu_N \quad \text{on } \Gamma_N, \end{aligned}$$

has a solution. Here  $c$  is a constant and  $\mu_D, \mu_N$  and  $f$  are given functions whose properties will be stated in the upcoming theorems. We say that  $\varphi \in H^1(\Omega)$  is a weak solution of this mixed boundary value problem if

$$(\nabla \varphi, \nabla \xi) + C(\varphi, \xi) = (f, \xi) + (\mu_N, \xi)_{\Gamma_N} \text{ and } \varphi = \mu_D \text{ on } \Gamma_D,$$

for all  $\xi \in H^1(\Omega)$  with  $\xi = 0$  on  $\Gamma_D$  where we recall that  $(\mu_N, \xi)_{\Gamma_N} = \int_{\Gamma_N} \mu_N \xi^- dA$ . In fact, we will only consider several classical special cases of this problem.

The first result is a basic existence result for solutions of elliptic problems. It follows directly from the Lax–Milgram Lemma 2.21 and the Trace Theorem 3.9.

**Theorem 3.12** Let  $\Omega$  be a Lipschitz domain. Let  $\mu \in H^{1/2}(\partial\Omega)$  and  $f \in H^1(\Omega)$ . Then there exists a unique weak solution  $\varphi \in H^1(\Omega)$  of

$$-\nabla \varphi + \varphi = f \text{ in } \Omega \text{ and } \varphi = \mu \text{ on } \partial\Omega,$$

Furthermore, there is a constant  $C$  such that

$$\|\varphi\|_{H^1(\Omega)} \leq C \left( \|\mu\|_{H^{1/2}(\partial\Omega)} + \|f\|_{H^{-1}(\Omega)} \right).$$

We shall also need to know conditions under which the Dirichlet problem for Poisson's equation has a solution. This is essentially the same result as in the previous theorem, but requires a special inequality giving an alternative norm for  $H^1(\Omega)$ . This is called the *Poincaré inequality* and we give a general version from Brenner and Scott [60]

**Lemma 3.13** *There exists a constant  $C > 0$  such that for all  $u \in H^1(\Omega)$*

$$\|u\|_{H^1(\Omega)} \leq C \left( \|\nabla u\|_{L^2(\Omega)} + \left| \int_{\partial\Omega} u dA \right| \right).$$

Note that this result also holds if  $\partial\Omega$  is replaced by a subset of  $\partial\Omega$  of positive measure. In addition, the integral on the right-hand side can be replaced by an integral over all  $\Omega$ .

Using this lemma, and proceeding as for the previous theorem, we can then prove the next theorem.

**Theorem 3.14** *Let  $\Omega$  be a Lipschitz domain. Let  $\mu \in H^{1/2}(\partial\Omega)$  and  $f \in H^1(\Omega)$ . Then there exists a unique weak solution  $\varphi \in H^1(\Omega)$  of*

$$-\nabla \cdot \varphi = f \text{ in } \Omega \text{ and } \varphi = \mu \text{ on } \partial\Omega.$$

Furthermore, there is a constant  $C$  such that

$$\|\varphi\|_{H^1(\Omega)} \leq C \left( \|\mu\|_{H^{1/2}(\partial\Omega)} + \|f\|_{H^{-1}(\Omega)} \right).$$

We shall also need to know that the Neumann problem is well-defined. Proceeding as before, we have the following result.

**Theorem 3.15** *Let  $\Omega$  be a Lipschitz domain with boundary  $\partial\Omega$  and unit outward normal  $\nu$ . Let  $\mu \in H^{1/2}(\partial\Omega)$  and  $f \in H^1(\Omega)'$ . Then there exists a unique weak solution  $\varphi \in H^1(\Omega)$  of*

$$-\Delta \varphi + \varphi = f \text{ in } \Omega \text{ and } \frac{\partial \varphi}{\partial \nu} = \mu \text{ on } \partial\Omega.$$

Furthermore, there is a constant  $C$  such that

$$\|\varphi\|_{H^1(\Omega)} \leq C \left( \|\mu\|_{H^{-1/2}(\partial\Omega)} + \|f\|_{H^1(\Omega)'} \right).$$

For the Neumann problem for Poisson's equation we need to introduce the space

$$H^1(\Omega)/\mathbb{R} = \left\{ u \in H^1(\Omega) \mid \int_{\partial\Omega} u dA = 0 \right\}.$$

Then, using the Poincaré inequality, we have the following theorem, which differs from the previous results in that a compatibility condition must be imposed on the data.

**Theorem 3.16** *Let  $\Omega$  be a Lipschitz domain with boundary  $\partial\Omega$  and unit outward normal  $\nu$ . Let  $\mu \in H^{1/2}(\partial\Omega)$  and  $f \in H^1(\Omega)'$ . Suppose*

$$\int_{\partial\Omega} \mu dA \int_{\Omega} f dV = 0.$$

*Then there exists a unique weak solution  $\varphi \in H^1(\Omega)/\mathbb{R}$  of*

$$-\Delta \varphi = f \text{ in } \Omega \text{ and } \frac{\partial \varphi}{\partial \nu} = \mu \text{ on } \partial \Omega .$$

Furthermore, there is a constant  $C$  such that

$$\|\varphi\|_{H^1(\Omega)} \leq C \left( \|\mu\|_{H^{-1/2}(\partial\Omega)} + \|f\|_{H^1(\Omega)'} \right).$$

The above theorems give the basic existence and uniqueness results we shall need. However, we shall often be in a situation where the data are smoother than assumed above. This can sometimes result in a smoother solution of Poisson's equation. The actual regularity of the solution depends on the data and on the smoothness of the boundary. For an arbitrary Lipschitz domain the following theorem from [102, 175, 176] is known.

**Theorem 3.17** Let  $\Omega$  be a Lipschitz domain. Suppose  $\varphi \in H^1(\Omega)$  is the weak solution of  $\Delta \varphi = 0$  on  $\Omega$  such that  $\varphi|_{\partial\Omega} = \mu \in H^1(\partial\Omega)$ . Then  $\varphi \in H^{3/2}(\Omega)$  and

$$\|\varphi\|_{H^{3/2}(\Omega)} \leq C \|\mu\|_{H^{-1/2}(\partial\Omega)}.$$

Suppose, instead,  $\varphi \in H^1(\Omega)$  satisfies  $\Delta \varphi = 0$  on  $\Omega$  and  $\partial\varphi/\partial\nu = \mu \in L^2(\partial\Omega)$  with  $\langle \mu, 1 \rangle_{\partial\Omega} = 0$ . Then  $\varphi \in H^{3/2}(\Omega)$  and

$$\|\varphi\|_{H^{3/2}(\Omega)} \leq C \|\mu\|_{L^2(\partial\Omega)}.$$

For a Lipschitz polyhedron the situation is improved due to the simpler boundary. Here we use the definition of  $H^s(\partial\Omega)$ ,  $s > 1$ , given in (3.12). This theorem is from [12].

**Theorem 3.18** Let  $\Omega$  be a Lipschitz polyhedral domain. Then there is an exponent  $s_\Omega > 0$  such that if  $\mu \in H^{1+\delta}(\partial\Omega)$ ,  $0 \leq \delta < \min(s_\Omega, 1/2)$ , and  $f \in L^2(\Omega)$  then the weak solution  $\varphi \in H^1(\Omega)$  of

$$-\Delta \varphi = f \text{ in } \Omega \text{ and } \varphi = \mu \text{ on } \partial \Omega$$

is such that  $\varphi \in H^{3/2+\delta}(\Omega)$ .

For the Neumann problem, a similar result holds so that there is an  $\tilde{s}_\Omega > 0$  such that if  $\mu \in H^\delta(\partial\Omega)$ ,  $0 \leq \delta < \min(\tilde{s}_\Omega, 1/2)$ , and  $f \in L^2(\Omega)$ , together with  $(f, 1) + \langle \mu, 1 \rangle_{\partial\Omega} = 0$ , then the weak solution  $\varphi \in H^1(\Omega)$  of

$$-\Delta \varphi = f \text{ in } \Omega \text{ and } \frac{\partial \varphi}{\partial \nu} = \mu \text{ on } \partial \Omega$$

is such that  $\varphi \in H^{3/2+\delta}(\Omega)$ .

Since this theorem is important for our analysis and perhaps not very well-known, we shall sketch a proof of the first part. By the definition of  $H^{1+\delta}(\partial\Omega)$ , there is a function  $\hat{\mu} \in H^{3/2+\delta}(\Omega)$  such that  $\mu = \hat{\mu}|_{\partial\Omega}$ . Then define  $\hat{\varphi} = \varphi - \hat{\mu}$  so that  $\hat{\varphi}$  satisfies

$$-\Delta \hat{\varphi} = f + \Delta \hat{\mu} \text{ in } \Omega \text{ and } \hat{\varphi} = 0 \text{ on } \partial \Omega .$$

The result then follows from Corollary 18.15 of [112].

### 3.4 Differential operators on a surface

Before starting our description of vector Sobolev spaces, we need to define some differential operators related to tangential vector fields on  $\partial\Omega$ . Suppose  $\Omega$  is a bounded domain with  $C^2$  connected boundary  $\partial\Omega$  (i.e. the maps in Definition 3.1 are in  $C^2$ ). In fact, with some additional work of a very non-trivial nature, much of what we present here can be extended to a Lipschitz domain [63]. Let us define the space of surface tangential vector fields in  $L^2(\partial\Omega)$  by(3.13)

$$L_t^2(\partial\Omega) = \left\{ u \in (L^2(\partial\Omega))^3 \mid v \cdot u = 0 \text{ a.e. on } \partial\Omega \right\},$$

where  $v$  is the unit outward normal to  $\Omega$ . The norm on this space is the standard  $(L^2(\partial\Omega))^3$  norm.

We start by defining two fundamental differential operators. For a function  $p \in H^1(\partial\Omega)$  we define the surface gradient  $\nabla_{\partial\Omega} p$  via a parametric representation of  $\partial\Omega$ . Suppose  $x \in \partial\Omega$  can be written as

$$x = (x_1(u_1, u_2), x_2(u_1, u_2), x_3(u_1, u_2))^T$$

for some surface patch of  $\partial\Omega$ . Then, on this patch,  $\nabla_{\partial\Omega} p \in L_t^2(\partial\Omega)$  is defined by

$$\nabla_{\partial\Omega} p = \sum_{i,j=1}^2 g^{ij} \frac{\partial p}{\partial u_i} \frac{\partial x}{\partial u_j}$$

where  $g^{ij}$  is the  $(i, j)$ th entry of the inverse of the matrix  $G$  given by

$$G_{ij} = \frac{\partial x}{\partial u_i} \cdot \frac{\partial x}{\partial u_j}, \quad i, j = 1, 2.$$

In particular, if  $\partial\Omega = \partial B_1$  where  $B_1$  is the unit sphere centered at the origin (i.e.  $\partial\Omega$  is the surface of the unit sphere) and if we use spherical polar coordinates  $(\rho, \theta, \varphi)$  then

$$\nabla_{\partial B_1} p = \frac{\partial p}{\partial \theta} e_\theta + \frac{1}{\sin \theta} \frac{\partial p}{\partial \varphi} e_\varphi.$$

One useful observation is that the surface gradient and volume gradient are related for functions  $p$  that are differentiable in the neighborhood of  $\partial\Omega$  by(3.14)

$$(\nabla p) \Big|_{\partial\Omega} = \nabla_{\partial\Omega} p + \frac{\partial p}{\partial v} v.$$

With this observation, we see that  $(v \times \nabla p) \times v = \nabla_{\partial\Omega} p$  on  $\partial\Omega$ . This important result holds for Lipschitz domains also.

Having defined the surface gradient, we can define the surface divergence  $\nabla_{\partial\Omega} \cdot : L_t^2(\partial\Omega) \rightarrow H^1(\partial\Omega)'$  by duality so that if  $u \in L_t^2(\partial\Omega)$  then  $\nabla_{\partial\Omega} \cdot u \in H^1(\partial\Omega)'$  satisfies

$$\int_{\partial\Omega} \nabla_{\partial\Omega} \cdot u p dA = - \int_{\partial\Omega} \nabla_{\partial\Omega} p \cdot u dA \quad \text{forall } p \in H^1(\partial\Omega).$$

This definition corresponds to the usual definition for the surface divergence given by vector calculus, at least for smooth surfaces, so that if  $v \in (H^1(\partial\Omega))^3 \cap L_t^2(\partial\Omega)$  and if  $v$  has the expansion

$$v = v_1 \frac{\partial x}{\partial u_1} + v_2 \frac{\partial x}{\partial u_2}$$

then

$$\nabla_{\partial\Omega} \cdot v = \frac{1}{\sqrt{g}} \left\{ \frac{\partial}{\partial u_1} (\sqrt{g} v_1) + \frac{\partial}{\partial u_2} (\sqrt{g} v_2) \right\},$$

where  $g = \det(G)$ .

In spherical polar coordinates, if  $v = v_\theta e_\theta + v_\varphi e_\varphi$ , then on the surface of the unit sphere.

$$\int_{\partial\Omega} \nabla_{\partial\Omega} \times v p dA = \int_{\partial\Omega} v \cdot \vec{\nabla}_{\partial\Omega} \times p dA \text{ for all } p \in H^1(\partial\Omega).$$

The operator  $\Delta_{\partial\Omega} : H^1(\partial\Omega) \rightarrow H^1(\partial\Omega)'$  defined for  $p \in H^1(\partial\Omega)$  by  $\Delta_{\partial\Omega} p = \nabla_{\partial\Omega} \cdot (\nabla_{\partial\Omega} p)$  is called the surface Laplacian or *Laplace–Beltrami operator*.

The third fundamental operator is the surface vector curl denoted by  $\vec{\nabla}_{\partial\Omega} \times : H^1(\partial\Omega) \rightarrow L_t^2(\partial\Omega)$  and defined by

$$\vec{\nabla}_{\partial\Omega} \times p = -v \times \nabla_{\partial\Omega} p.$$

Thus  $\vec{\nabla}_{\partial\Omega} \times$  is just the rotated gradient.

One remaining operator, the surface scalar curl denoted by  $\nabla_{\partial\Omega} \times : L_t^2(\partial\Omega) \rightarrow H^1(\partial\Omega)'$ , can be defined via duality using Stokes theorem, so that if  $v \in L_t^2(\partial\Omega)$  then

$$\int_{\partial\Omega} \nabla_{\partial\Omega} \times v p dA = \int_{\partial\Omega} v \cdot \vec{\nabla}_{\partial\Omega} \times p dA \text{ for all } p \in H^1(\partial\Omega).$$

By using the duality definitions, we see that for  $v \in L_t^2(\partial\Omega)$  we have(3.15)

$$\nabla_{\partial\Omega} \times v = -\nabla_{\partial\Omega} \cdot (v \times v) \text{ and } \nabla_{\partial\Omega} \cdot v = \nabla_{\partial\Omega} \times (v \times v).$$

### 3.5 Vector functions with well-defined curl or divergence

The  $L^2(\Omega)$  inner product extends trivially to vector functions. Suppose that  $u = (u_1, u_2, u_3)^\top \in (L^2(\Omega))^3$  and  $v = (v_1, v_2, v_3)^\top \in (L^2(\Omega))^3$ . Then we write the  $(L^2(\Omega))^3$  inner product as(3.16)

$$(u, v) = \int_{\Omega} \sum_{j=1}^3 u_j \overline{v}_j dV.$$

Let us first define the curl and divergence. The curl operator is defined on a three dimensional vector function  $\mathbf{v} \in (C_0^\infty(\Omega))'$ <sup>3</sup> where  $\mathbf{v} = (v_1, v_2, v_3)^\top$  by(3.17)

$$\nabla \times \mathbf{v} = \left( \frac{\partial v_3}{\partial x_2} - \frac{\partial v_2}{\partial x_3}, \frac{\partial v_1}{\partial x_3} - \frac{\partial v_3}{\partial x_1}, \frac{\partial v_2}{\partial x_1} - \frac{\partial v_1}{\partial x_2} \right)$$

where the derivatives are understood in the sense of distributions. In particular, applying (3.1) to each component of the curl we see that(3.18)

$$(\nabla \times \mathbf{v}, \varphi) = (\mathbf{v}, \nabla \times \varphi) \text{ for all } \varphi \in (C_0^\infty(\Omega))'.$$

To define the divergence operator, let  $(C_0^\infty(\Omega))'$ <sup>3</sup> then(3.19)

$$\nabla \cdot \mathbf{v} = \sum_{i=1}^3 \frac{\partial v_i}{\partial x_i}.$$

Applying (3.1) to each component of the divergence, we see that(3.20)

$$(\nabla \cdot \mathbf{v}, \varphi) = -(\mathbf{v}, \nabla \varphi) \text{ for all } \varphi \in C_0^\infty(\Omega)^3.$$

Using the weak definition of derivative, we can then show that(3.21)

$$\begin{aligned} \nabla \times (\nabla p) &= 0 \text{ for all } p \in C_0^\infty(\Omega)', \\ \nabla \cdot (\nabla \times \mathbf{v}) &= 0 \text{ for all } \mathbf{v} \in (C_0^\infty(\Omega))'. \end{aligned} \quad (3.22)$$

For example, to prove (3.21) we use (3.18) with  $\mathbf{v} = \nabla_p$  to show that  $(\nabla \times \nabla_p, \varphi) = (\nabla_p, \nabla \times \varphi)$  for all  $\varphi \in (C_0^\infty(\Omega))'$  and by the distributional definition of the gradient  $(\nabla_p, \nabla \times \varphi) = -(p, \nabla \cdot (\nabla \times \varphi)) = 0$  for all  $\varphi \in (C_0^\infty(\Omega))'$  where the last equality holds since  $\nabla \cdot (\nabla \times \varphi) = 0$  for smooth functions. Now that we have defined the curl and divergence, we can consider suitable function spaces related to these operators. Indeed the goal of the rest of this chapter is to define and investigate such spaces. We shall be particularly interested in density results guaranteeing that functions in an appropriate space can be approximated by smooth functions. This will allow us to establish certain trace theorems and integral identities in the standard way. We start by considering classical integral identities for differentiable functions.

### 3.5.1 Integral identities

Here we recall some basic integral identities for vector functions with sufficiently many classical derivatives. We start with the basic divergence theorem of Gauss, which is proved for Lipschitz domains as Lemma 3.34 of [215].

**Theorem 3.19** (Divergence Theorem) *Let  $\Omega \subset \mathbb{R}^3$ , with boundary  $\partial\Omega$  and unit outward normal  $\nu$ , be a bounded Lipschitz domain. Let  $F : \mathbb{R}^3 \rightarrow \mathbb{R}^3$  be a vector field with  $F \in (C^1(\Omega))^3$ . Then(3.23)*

$$\int_{\Omega} \nabla \cdot F dV = \int_{\partial\Omega} F \cdot \nu dA.$$

*This result also holds for a Lipschitz domain in  $\mathbb{R}^2$  with suitable changes to the integral measures.*

Using this theorem, we can easily prove various important identities.

**Corollary 3.20** *Let  $\Omega \subset \mathbb{R}^3$  be a bounded Lipschitz domain with boundary  $\partial\Omega$  and unit outward normal  $\nu$ .*

(1) *If  $\xi \in C^1(\Omega)$  and  $u \in (C^1(\Omega))^3$  then (3.24)*

$$\int_{\Omega} \nabla \cdot u \xi \, dV = - \int_{\Omega} u \cdot \nabla \xi \, dV + \int_{\partial\Omega} u \cdot u \xi \, dA .$$

(2) (Green's first identity) *If  $\xi \in C^1(\Omega)$  and  $p \in C^2(\Omega)$  then (3.25)*

$$\int_{\Omega} \nabla p \cdot \xi \, dV = - \int_{\Omega} \nabla p \cdot \nabla \xi \, dV + \int_{\partial\Omega} \frac{\partial p}{\partial u} \xi \, dA .$$

(3) (Green's second identity) *If  $\xi \in C^2(\Omega)$  and  $p \in C^2(\Omega)$  then (3.26)*

$$\int_{\Omega} (\nabla p \cdot \xi - p \Delta \xi) \, dV = \int_{\partial\Omega} \left( \frac{\partial p}{\partial u} \xi - \frac{\partial \xi}{\partial u} p \right) \, dA .$$

(4) *Suppose  $u$  and  $\varphi$  are in  $(C^1(\Omega))^3$ . Then (3.27)*

$$\int_{\Omega} \nabla \times u \cdot \varphi \, dV = \int_{\Omega} u \cdot \nabla \times \varphi \, dV + \int_{\partial\Omega} u \times u \cdot \varphi \, dA .$$

For the first identity (3.24), we choose  $F = \xi u$  in (3.23) and use the vector identity (B.3). Identity (3.25) follows from (3.24) by choosing  $u = \nabla p$ . Subtracting (3.25) with the roles of  $\xi$  and  $p$  reversed gives (3.26). Finally, (3.27) is proved by choosing  $F = u \times \varphi$  and using (B.7).

Comparing (3.24) with (3.20), and (3.27) with (3.18) reveals the link between the definitions of the distributional derivative of the divergence and curl, and the corresponding classical integral identities. We shall extend the identities (3.24) and (3.27) to functions in suitable Sobolev spaces in the next two sections.

The last result of this section is a special case of the classical Stokes theorem. Here we use the notation of Section 3.4. Let  $S$  denote a bounded Lipschitz domain in the  $(x_1, x_2)$ -plane. Recall from Section 3.4 that, given a differentiable scalar function  $\varphi = \varphi(x_1, x_2)$ , we have the surface vector curl defined by

$$\vec{\nabla}_S \times \varphi = \left( \frac{\partial \varphi}{\partial x_2}, \frac{\partial \varphi}{\partial x_1} \right)^T$$

and for a vector function  $u = (u_1(x_1, x_2), u_2(x_1, x_2))^T$  we have the surface scalar curl defined by

$$\nabla_S \times u = \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2} .$$

Note that if the unit outward normal to  $\partial S$  in the  $(x_1, x_2)$  plane is  $\nu_p = (\nu_1, \nu_2, 0)^T$  then the corresponding unit tangent vector is  $\tau = (-\nu_2, \nu_1, 0)^T$  (obviously assuming that the plane containing  $S$  is oriented with normal along the positive  $x_3$ -axis and the right-hand rule is in effect). The following can be proved directly by integration by parts or by using (3.23).

**Corollary 3.21** (Stoke's theorem) *Let  $S \subset \mathbb{R}^2$  be a bounded Lipschitz domain with unit tangent  $\tau$  to  $\partial S$ . If  $u \in (C^1(S))^2$  and  $\xi \in C^1(\bar{S})$  then(3.28)*

$$\int_S \nabla_S \times u \cdot \xi \, dA = \int_S u \cdot \vec{\nabla}_S \times \xi \, dA + \int_{\partial S} \tau \cdot u \cdot \xi \, ds.$$

### 3.5.2 Properties of $H(\text{div}; \Omega)$

In this section we state and prove some results concerning the space of vector functions with a square-integrable divergence. The results and proofs are mainly from [143]. The space of functions with square-integrable divergence is denoted by  $H(\text{div}; \Omega)$  and defined by(3.29)

$$H(\text{div}; \Omega) = \left\{ u \in \left(L^2(\Omega)\right)^3 \mid \nabla \cdot u \in L^2(\Omega) \right\}$$

with the associated graph norm(3.30)

$$\|u\|_{H(\text{div}; \Omega)} = \left( \|u\|_{\left(L^2(\Omega)\right)^3}^2 + \|\nabla \cdot u\|_{L^2(\Omega)}^2 \right)^{1/2}.$$

With the obvious inner product,  $H(\text{div}; \Omega)$  is a Hilbert space.

The first result is a basic density result that will be used for the remainder of the proofs.

**Theorem 3.22** *Let  $\Omega$  be a bounded Lipschitz domain in  $\mathbb{R}^3$ . Then*

$$H(\text{div}; \Omega) = \text{closure of } \left(C^\infty(\overline{\Omega})\right)^3 \text{ in the } H(\text{div}; \Omega) \text{ norm.}$$

**Remark 3.23** *This theorem also holds when  $\Omega$  is unbounded provided  $\partial\Omega$  is bounded. The following proof is from [115].*

**Proof of Theorem 3.22** The proof uses the projection theorem (Theorem 2.9). We prove that if  $u \in H(\text{div}; \Omega)$  is such that  $u$  is orthogonal to  $(C^\infty(\Omega))^3$  in the  $H(\text{div}; \Omega)$  inner product then  $u = 0$ . This implies that the orthogonal complement of the closure of  $(C^\infty(\Omega))^3$  in  $H(\text{div}; \Omega)$  contains only the zero vector and the desired result then follows from the projection theorem.

The assumption of orthogonality implies that(3.31)

$$(u, \varphi) + (\nabla \cdot u, \nabla \cdot \varphi) = 0 \text{ for all } \varphi \in \left(C^\infty(\overline{\Omega})\right)^3.$$

Now we define  $Du = \nabla \cdot u$  and define  $\tilde{u}$  and  $D\tilde{u}$  to be the extensions of  $u$  and  $Du$  to  $\mathbb{R}^3$  by zero outside  $\Omega$ . Obviously,  $\tilde{u} \in (L^2(\mathbb{R}^3))^3$  and  $D\tilde{u} \in L^2(\mathbb{R}^3)$  and by (3.31)

$$(\tilde{u}, \varphi)_{\left(L^2(\mathbb{R}^3)\right)^3} + (D\tilde{u}, \nabla \cdot \varphi)_{L^2(\mathbb{R}^3)} = 0 \text{ for all } \varphi \in \left(C_0^\infty(\mathbb{R}^N)\right)^3$$

where  $(\cdot, \cdot)_{\left(L^2(\mathbb{R}^3)\right)^3}$  is the  $(L^2(\mathbb{R}^3))^3$  inner product. But using (3.20), this equality implies that  $\tilde{u} = \nabla D\tilde{u}$  and, since  $\tilde{u} \in (L^2(\mathbb{R}^3))^3$ , we conclude that  $D\tilde{u} \in$

$H^1(\mathbb{R}^3)$ . Hence, by the definition of  $H^1(\Omega)$ , we have that  $D\tilde{u} \in H^1(\Omega)$ . Now let  $O$  denote a ball such that  $\Omega \subset O$ . Then  $O \setminus \Omega$  is a bounded Lipschitz domain and  $D\tilde{u}|_{O \setminus \Omega} = 0$ . Thus by the trace theorem (Theorem 3.9) applied to  $O \setminus \Omega$  we see that  $D\tilde{u} = 0$  on  $\partial\Omega$  and hence  $Du \in H_0^1(\Omega)$ . By definition,  $C_0^\infty(\Omega)$  is dense in  $H_0^1(\Omega)$ , so there is a sequence of functions  $\{\varphi_n\}_{n=1}^\infty \subset C_0^\infty(\Omega)$  such that  $\varphi_n \rightarrow Du$  in  $H_0^1(\Omega)$  as  $n \rightarrow \infty$ . Then, by (3.31) with  $\varphi = \nabla\varphi_n$ , we conclude that

$$\begin{aligned} (u, u) + (\nabla \cdot u, \nabla \cdot u, ) &= (u, \nabla Du) + (\nabla \cdot u, Du) \\ &= \lim_{n \rightarrow \infty} \{ (u, \nabla \varphi_n) + (\nabla \cdot u, \varphi_n) \} = 0. \end{aligned}$$

This shows that  $\|u\|_{H(\text{div}; \Omega)} = 0$  and the theorem is proved.

The next theorem shows that functions in  $H(\text{div}; \Omega)$  have a well-defined normal component on  $\partial\Omega$ . This fact turns out to have implications for the continuity conditions imposed on the electromagnetic field across interfaces between dissimilar materials. (see Section 1.2.2). For a function  $v \in (C^\infty(\Omega))^3$  the normal trace operator  $\gamma_n$  is defined almost everywhere in the classical way by(3.32)

$$\gamma_n(v) = v|_{\partial\Omega} \cdot \nu.$$

**Theorem 3.24** Let  $\Omega \subset \mathbb{R}^3$  be a bounded Lipschitz domain in  $\mathbb{R}^3$  with unit outward normal  $\nu$ . Then

- (1)  $\frac{1}{2}$ the mapping  $\gamma_n$  defined(3.32)on  $(C^\infty(\Omega))^3$ can be extended by continuity to a continuous linear map  $\gamma_n$ from  $H(\text{div}; \Omega)$  onto  $H^{1/2}(\partial\Omega)$ ;
- (2) the following Green's theorem holds for functions  $v \in H(\text{div}; \Omega)$  and  $\varphi \in H^1(\Omega)$ :

(3.33)

$$(v, \nabla \varphi) + (\nabla \cdot v, \varphi) = \langle \varphi, \gamma_n(v) \rangle_{\partial\Omega}.$$

**Proof** The proof is a standard application of the density result in Theorem 3.22. We start using (3.24) so that for  $\varphi \in C^\infty(\Omega)$  and  $v \in (C^\infty(\Omega))^3$  (3.34)

$$(v, \nabla \varphi) + (\nabla \cdot v, \varphi) = \langle \varphi, v \cdot \nu \rangle_{\partial\Omega}.$$

But since  $C^\infty(\Omega)$  is dense in  $H^1(\Omega)$ . (Theorem 3.2) this Green's theorem is also valid for function  $\varphi$  in  $H^1(\Omega)$ . Now using the Cauchy–Schwarz inequality on the left-hand side of (3.34) we conclude that(3.35)

$$|\langle \varphi, v \cdot \nu \rangle_{\partial\Omega}| \leq \|v\|_{H(\text{div}; \Omega)} \|\varphi\|_{H^1(\Omega)}$$

for all  $\varphi \in H^1(\Omega)$  and for all  $v \in (C^\infty(\Omega))^3$ . Let  $\mu \in H^{1/2}(\partial\Omega)$  and define  $\varphi \in H^1(\Omega)$  to be the weak solution of

$$-\Delta \varphi + \varphi = 0 \text{ in } \Omega \text{ and } \varphi = \mu \text{ on } \partial\Omega.$$

By virtue of the regularity result for elliptic problems in Theorem 3.12,

$$\|\varphi\|_{H^1(\Omega)} \leq C \|\mu\|_{H^{1/2}(\partial\Omega)},$$

so we may rewrite (3.35) as

$$|\langle \mu, u \cdot v \rangle_{\partial\Omega}| \leq C \|v\|_{H(\text{div}; \Omega)} \|\mu\|_{H^{1/2}(\partial\Omega)}$$

for all  $\mu \in H^{1/2}(\partial\Omega)$  and for all  $v \in (C^\infty(\Omega))^3$ . This implies, using the definition of the norm on  $H^{1/2}(\partial\Omega)$  given in (3.11), that (3.36)

$$\|u \cdot v\|_{H^{-1/2}(\partial\Omega)} \leq C \|v\|_{H(\text{div}; \Omega)}.$$

Hence  $\gamma_n : v \mapsto v \cdot v|_{\partial\Omega}$  is a bounded, and therefore continuous, linear map from the dense set  $(C^\infty(\Omega))^3 \subset H(\text{div}; \Omega)$  to  $H^{1/2}(\partial\Omega)$ . Hence  $\gamma_n$  can be extended by continuity to a map (still denoted  $\gamma_n$ ) from  $H(\text{div}; \Omega)$  to  $H^{1/2}(\partial\Omega)$ .

It remains to show surjectivity. We accomplish this by showing that for any  $\mu \in H^{1/2}(\partial\Omega)$  there is a  $v \in H(\text{div}; \Omega)$  such that  $\gamma_n(v) = \mu$ . Let  $\varphi \in H^1(\Omega)$  be the solution of the Neumann problem (guaranteed by Theorem 3.14)(3.37)

$$(\nabla \varphi, \nabla \psi) + (\varphi, \psi) = \langle \mu, \psi \rangle_{\partial\Omega} \text{ for all } \psi \in H^1(\Omega).$$

Let  $v = \nabla \varphi \in (L^2(\Omega))^3$ . Then taking  $\psi \in C_0^\infty(\Omega)$  in (3.37) we see that

$$(u, \nabla \psi) + (\varphi, \psi) = 0 \text{ for all } \psi \in C_0^\infty(\Omega),$$

so by the distributional definition of the divergence  $\nabla \cdot v = \varphi \in L^2(\Omega)$ . Thus  $v \in H(\text{div}; \Omega)$  and  $\gamma_n(v) = v \cdot \nabla \varphi = \mu$ , which establishes surjectivity. This completes the proof of the theorem.  $\square$

To solve problems in which the normal component of a vector field is specified on  $\partial\Omega$ , we shall need to consider the subspace of  $H(\text{div}; \Omega)$  on which  $\gamma_n$  vanishes. As in the case of Sobolev spaces, we define this subspace in a roundabout fashion by density. In particular(3.38)

$$H_0(\text{div}; \Omega) = \text{closure of } \left( C_0^\infty(\Omega) \right)^3 \text{ in the } H(\text{div}; \Omega) \text{ norm.}$$

The next theorem tells us that we have made the correct definition

**Theorem 3.25** *Let  $\Omega$  be a bounded Lipschitz domain in  $\mathbb{R}^3$ . Then*

$$H_0(\text{div}; \Omega) = \{ u \in H(\text{div}; \Omega) \mid u \cdot v|_{\partial\Omega} = 0 \}$$

**Proof** Again, we use the projection theorem (Theorem 2.9) to write

$$H(\text{div}; \Omega) = \left( \text{closure} \left( C_0^\infty(\Omega) \right)^3 \right) \oplus \left( \text{closure} \left( C_0^\infty(\Omega) \right)^3 \right)^\perp,$$

where closure is with respect to the  $H(\text{div}; \Omega)$  norm. We then show that if(3.39)

$$u \in \left( \text{closure} \left( C_0^\infty(\Omega) \right)^3 \right)^\perp \text{ and } \gamma_n(u) = 0$$

then  $u = 0$ . Suppose  $u$  satisfies (3.39). Then

$$(u, u) + (\nabla \cdot u, \nabla \cdot u) = 0 \text{ for all } u \in \left(C_0^\infty(\Omega)\right)^3,$$

and hence if we define  $Dv = \nabla \cdot v$  the above equality shows that in the distributional sense  $v = \nabla Dv$ . Since  $v \in (L^2(\Omega))^3$ , we conclude that  $Dv \in H(\Omega)$ . Applying the Green's formula (3.33) with  $v = u$  and  $\varphi = Du$  and using the hypothesis that  $\gamma_n(v) = 0$ , we have that

$$(u, u) + (\nabla \cdot u, \nabla \cdot u) = (u, \nabla Du) + (\nabla \cdot u, \nabla u) = \langle \gamma_n(u), Du \rangle_{\partial\Omega} = 0.$$

Hence  $v = 0$  and we are done.  $\square$

### 3.5.3 Properties of $H(\text{curl};\Omega)$

We define the space of three-dimensional vector functions with curl in  $L^2$  by(3.40)

$$H(\text{curl};\Omega) = \left\{ u \in \left(L^2(\Omega)\right)^3 \mid \nabla \times u \in \left(L^2(\Omega)\right)^3 \right\}$$

with the graph norm(3.41)

$$\|v\|_{H(\text{curl};\Omega)} = \left( \|v\|_{\left(L^2(\Omega)\right)^3}^2 + \|\nabla \times v\|_{\left(L^2(\Omega)\right)^3}^2 \right)^{1/2}.$$

From the point of view of Maxwell's equations the space  $H(\text{curl};\Omega)$  is of central importance since it corresponds to the space of finite-energy solutions.

Corresponding to the definition of higher-order scalar Sobolev spaces, it is also convenient to define, for  $s \geq 0$ ,(3.42)

$$H^S(\text{curl};\Omega) = \left\{ u \in \left(H^S(\Omega)\right)^3 \mid \nabla \times u \in \left(H^S(\Omega)\right)^3 \right\}.$$

The space  $H_0(\text{curl};\Omega)$  is defined by density as follows:

$$H_0(\text{curl};\Omega) = \text{closure of } \left(C_0^\infty(\Omega)\right)^3 \text{ in } H(\text{curl};\Omega).$$

As in the case of the divergence spaces, we start with a density result.

**Theorem 3.26** Suppose  $\Omega$  is a bounded Lipschitz domain in  $\mathbb{R}^3$ . Then the closure of  $(C^\infty(\Omega))^3$  in the  $H(\text{curl};\Omega)$  norm is  $H(\text{curl};\Omega)$ .

To prove Theorem 3.26, we need the following lemma, which gives an alternative characterization of functions in  $H_0(\text{curl};\Omega)$ .

**Lemma 3.27** Let  $\Omega$  be a bounded Lipschitz domain in  $\mathbb{R}^3$  and let  $u \in H(\text{curl};\Omega)$  be such that for every  $\varphi \in (C^\infty(\Omega))^3$ (3.43)

$$(\nabla \times u, \varphi) - (u, \nabla \times \varphi) = 0.$$

Then  $u \in H_0(\text{curl};\Omega)$ .

**Remark 3.28** The Green's theorem in(3.43) holds for  $u \in H_0(\text{curl};\Omega)$  and  $\varphi \in H(\text{curl};\Omega)$ . This follows once we have proved the density of  $(C^\infty(\Omega))^3$  in  $H(\text{curl};\Omega)$ .

**Proof of Lemma 3.27** The proof of this lemma is rather technical and is from [143]. The idea of the proof is to use a convolution to construct a sequence of functions in  $(C_0^\infty(\Omega))^3$  that approach  $\mathbf{u}$  in the  $H(\text{curl}; \Omega)$  norm. In order to do this,  $\Omega$  is first decomposed into a union of simpler subdomains. Then the sequence is constructed on each subdomain. We shall give a complete proof, except that we shall assume a number of properties related to the convolution introduced below (see, e.g. [298] for a complete discussion of the convolution).

Since  $\Omega$  is a bounded Lipschitz domain, there is a finite collection of open sets  $\{\mathcal{O}_j\}_{j=1}^J$  such that  $\Omega \subset \cup_{j=1}^J \mathcal{O}_j$  and such that each  $\Omega_j = O_j \cap \Omega$ ,  $1 \leq j \leq J$ , is a bounded, starlike, Lipschitz domain. By starlike we mean that for each  $j$  there is a  $y_j \in \Omega_j$  such that for any  $x \in \Omega_j$  we have  $y_j + \theta(x - y_j) \in \Omega_j$  for all  $\theta \in [0, 1]$ . Relative to this open covering, there exists a partition of unity which is a set of functions  $\{\alpha_j\}_{j=1}^J$  such that  $\alpha_j \in C_0^\infty(\mathcal{O}_j)$ ,  $1 \leq j \leq J$ , and  $\sum_{j=1}^J \alpha_j(x) = 1$  for all  $x \in \Omega$ . For a discussion of such open coverings and partitions of unity, see, e.g., [298].

Let  $\tilde{u}$  denote the extension of  $u$  by zero to all of  $\mathbb{R}^3$ . From (3.43) it is clear that  $\tilde{u} \in H(\text{curl}; \mathbb{R}^3)$ . Then using the partition of unity

$$\tilde{u} = \sum_{j=1}^J \alpha_j \tilde{u} \text{ in } \Omega$$

and  $\tilde{u}_j = \alpha_j \tilde{u} \in H(\text{curl}; \mathbb{R}^3)$  with  $\text{supp}(\tilde{u}_j) \subset \Omega_j$ .

Now, for each  $\Omega_j$ , we adopt a coordinate system with the origin at the point  $y_j$  (the point about which  $\Omega_j$  is starlike). Then the functions  $\tilde{u}_j^\theta = (x/\theta)$  defined for  $\theta \in (0, 1)$  converge to  $\tilde{u}_j$  in  $H(\text{curl}; \mathbb{R}^3)$  as  $\theta \rightarrow 1$ . Since the set  $\Omega_j$  is starlike, we also have that  $\text{supp}(\tilde{u}_j^\theta) \subset \Omega_j$  for  $0 < \theta < 1$ . Next we construct a sequence in  $(C_0^\infty(\Omega_j))^3$  converging to  $\tilde{u}_j^\theta$ .

Let  $\rho \in C_0^\infty(\mathbb{R}^3)$  be such that

$$\rho \geq 0, \quad \rho(x) = 0 \text{ if } |x| \geq 1 \text{ and } \int_{\mathbb{R}^3} \rho \, dV = 1$$

(for the construction of such a function, see [298]). Then the family of functions  $\varrho_\varepsilon$ ,  $\varepsilon > 0$  defined by  $\varrho_\varepsilon(x) = \rho(x/\varepsilon)/\varepsilon^3$  is such that

$$\varrho_\varepsilon \geq 0, \quad \varrho_\varepsilon(x) = 0 \text{ if } |x| \geq \varepsilon \quad \text{and} \quad \int_{\mathbb{R}^3} \varrho_\varepsilon \, dV = 1.$$

Now for any  $v \in L^2(\mathbb{R}^3)$ , let the convolution  $\varrho_\varepsilon * v$  be defined by

$$\varrho_\varepsilon * v(x) = \int_{\mathbb{R}^3} \varrho_\varepsilon(x-y)v(y) \, dV(y).$$

Then  $\varrho_\varepsilon * v \rightarrow v$  in  $L^2(\mathbb{R}^3)$  as  $\varepsilon \rightarrow 0$ , and  $\varrho_\varepsilon * v \in C_0^\infty(\mathbb{R}^3)$  [298]. The differentiability properties of the convolution imply that if  $v \in (C_0^\infty(\mathbb{R}^3))^3$  then

$\nabla \times (\rho_\varepsilon * v) = \rho_\varepsilon * (\nabla \times v)$  (where the convolution of a vector is defined elementwise). Hence  $\rho_\varepsilon * \tilde{u}_j^\theta \rightarrow \tilde{u}_j^\theta$  as  $\varepsilon \rightarrow 0$  in  $H(\text{curl}; \mathbb{R}^3)$ . Furthermore, since  $\tilde{u}_j^\theta$  has compact support in  $\Omega_j$ , if  $\varepsilon$  is small enough,  $\text{supp}(\rho_\varepsilon * \tilde{u}_j^\theta) \subset \Omega_j$  and hence  $\rho_\varepsilon * \tilde{u}_j^\theta \in (C_0^\infty(\Omega_j))^3$ . As a result, we can find a sequence of values  $\{\theta_k, \varepsilon_k\}_{k=1}^\infty$  such that  $(\theta_k, \varepsilon_k) \rightarrow (1, 0)$  as  $k \rightarrow \infty$  with  $0 < \theta_k < 1$  and  $0 < \varepsilon_k < 1$  such that  $\rho_{\varepsilon_k} * \tilde{u}_j^{\theta_k} \rightarrow \tilde{u}_j$  in  $H(\text{curl}; \Omega_j)$ . The function  $\tilde{u}^{(k)}$  defined by

$$\tilde{u}^{(k)} = \sum_{j=1}^J \rho_{\varepsilon_k} * \tilde{u}_j^{\theta_k}$$

is such that  $\tilde{u}^{(k)} \in (C_0^\infty(\Omega))^3$  for each  $k$  and  $\tilde{u}^{(k)} \rightarrow u$  in  $H(\text{curl}; \Omega)$ . Hence  $u \in H_0(\text{curl}; \Omega)$ , and the proof is complete.  $\square$

**Proof of Theorem 3.26** The proof is from [115]. We use the projection theorem (Theorem 2.9) and consider a function  $u \in H(\text{curl}; \Omega)$  that is orthogonal to all vector functions in  $(C^\infty(\Omega))^3$  so that (3.44)

$$(u, \varphi) + (\nabla \times u, \nabla \times \varphi) = 0 \quad \text{forall } \varphi \in (C^\infty(\overline{\Omega}))^3.$$

Now let  $v = \nabla \times u$ . Then the above equality and (3.18) imply that  $u = -\nabla \times v$ . Since  $u \in H(\text{curl}; \Omega)$ , this implies that  $\nabla \times v \in (L^2(\Omega))^3$  and hence that  $v \in H(\text{curl}; \Omega)$ . Furthermore, (3.44) implies that

$$(u, \nabla \times \varphi) - (\nabla \times u, \varphi) = 0 \quad \text{forall } \varphi \in (C^\infty(\overline{\Omega}))^3.$$

So, Lemma 3.27 is applicable and we conclude that  $v \in H_0(\text{curl}; \Omega)$ . Now since  $(C_0^\infty(\Omega))^3$  is dense in  $H_0(\text{curl}; \Omega)$ , there is a sequence  $\{\varphi_k\}_{k=1}^\infty \subset (C_0^\infty(\Omega))^3$  such that  $\varphi_k \rightarrow v$  in the  $H(\text{curl}; \Omega)$  norm as  $k \rightarrow \infty$ . Applying (3.44) again, we see that

$$(u, u) + (\nabla \times u, \nabla \times u) = \lim_{K \rightarrow \infty} -(u, \nabla \times \varphi_K) + (\nabla \times u, \varphi_K) = 0,$$

so that  $u = 0$ . Thus the orthogonal complement of the closure of  $(C^\infty(\Omega))^3$  is trivial and the theorem is proved.  $\square$

Now we examine the trace properties of functions in  $H(\text{curl}; \Omega)$ . Physically, we know that Maxwell's equations need the tangential trace of the electric field to be well-defined. Thus, if  $H(\text{curl}; \Omega)$  is to be used as the energy space for Maxwell's equations, we must verify that functions in this space have a well-defined tangential trace [143, 8, 63]. This is accomplished next. First we define, for a smooth vector function  $v \in (C^\infty(\Omega))^3$ , the two traces (3.45)

$$\begin{aligned} \gamma_t(v) &= v \times u|_{\partial\Omega}, \\ (3.46) \end{aligned}$$

$$\gamma_T(v) = (u \times u|_{\partial\Omega}) \times v,$$

where as usual  $v$  is the unit outward normal to  $\Omega$ .

**Theorem 3.29** Let  $\Omega$  be a bounded Lipschitz domain in  $\mathbb{R}^3$ . Then the trace map  $\gamma_t$ , which is defined classically via (3.45) on  $(C^\infty(\Omega))^3$  can be extended by continuity to a continuous linear map from  $H(\text{curl};\Omega)$  into  $(H^{1/2}(\partial\Omega))^3$ . Furthermore, the following Green's theorem holds for any  $v \in H(\text{curl};\Omega)$  and  $\varphi \in (H^1(\Omega))^3$ : (3.47)

$$(\nabla \times v, \varphi) - (v, \nabla \times \varphi) = \langle \gamma_t(v), \varphi \rangle_{\partial\Omega}.$$

**Remark 3.30** The map  $\gamma_t : H(\text{curl};\Omega) \rightarrow (H^{1/2}(\partial\Omega))^3$  is not surjective since for any  $v$ , the trace map  $\gamma_t(v)$  is tangential to  $\partial\Omega$ , whereas  $(H^{1/2}(\partial\Omega))^3$  contains vectors that are not tangential to  $\partial\Omega$ . The correct space for the trace is examined in the next theorem and the remark following that theorem.

**Proof of Theorem 3.29** The proof of this theorem closely resembles the proof of Theorem 3.24 using the density result proved in Theorem 3.26. We start with the standard integral identity (3.27) so that for any  $v$  and  $\varphi$  in  $(C^\infty(\Omega))^3$  (3.48)

$$(\nabla \times v, \varphi) - (v, \nabla \times \varphi) = \langle v \times v, \varphi \rangle_{\partial\Omega}.$$

Of course, the right-hand side may be written as  $\langle \gamma_t(v), \varphi \rangle_{\partial\Omega}$ .

Since  $(C^\infty(\Omega))^3$  is dense in  $(H^1(\Omega))^3$ , the identity (3.48) holds for  $\varphi \in (H^1(\Omega))^3$ . Now using the Cauchy–Schwarz inequality on the left-hand side of (3.48) we obtain (3.49)

$$|\langle v \times v, \varphi \rangle_{\partial\Omega}| \leq \|v\|_{H(\text{curl};\Omega)} \|\varphi\|_{(H^1(\Omega))^3}$$

for all  $v \in (C^\infty(\Omega))^3$  and for all  $\varphi \in (H^1(\Omega))^3$ . For given  $\mu \in (H^{1/2}(\Omega))^3$  we choose  $\varphi$  to be the weak solution of  $-\Delta\varphi + \varphi = 0$  in  $\Omega$  and  $\varphi = \mu$  on  $\partial\Omega$ . By Theorem 3.12 applied to each component of  $\varphi$ , we have that  $\|\varphi\|_{(H^1(\Omega))^3} \leq C \|\mu\|_{(H^{1/2}(\partial\Omega))^3}$  and hence, from (3.49),

$$|\langle v \times v, \varphi \rangle_{\partial\Omega}| \leq C \|v\|_{H(\text{curl};\Omega)} \|\mu\|_{(H^{1/2}(\Omega))^3}$$

for all  $v \in (C^\infty(\Omega))^3$  and for all  $\mu \in (H^{1/2}(\partial\Omega))^3$ . It follows from the definition of the  $-1/2$  Sobolev norm in (3.11) that  $\|v \times v\|_{(H^{1/2}(\partial\Omega))^3} \leq C \|v\|_{H(\text{curl};\Omega)}$ . Hence  $\gamma_t$ , which is defined on  $(C^\infty(\Omega))^3$ , is continuous as a map from  $H(\text{curl};\Omega)$  into  $(H^{1/2}(\partial\Omega))^3$ . Since  $(C^\infty(\Omega))^3$  is dense in  $H(\text{curl};\Omega)$ , the map  $\gamma_t$  can be extended by continuity to a map from  $H(\text{curl};\Omega)$  to  $(H^{1/2}(\partial\Omega))^3$ .  $\square$

We would like to prove a similar result about  $\gamma_t$  but this is not valid for Lipschitz domains because, even if  $v \in (H^1(\Omega))^3$ , it is not necessarily true that  $\gamma_t(v) \in (H^{1/2}(\partial\Omega))^3$ . For this reason we follow Chen *et al.* [77] and define the trace space for  $Y(\partial\Omega)$  as follows: (3.50)

$$Y(\partial\Omega) = \left\{ f \in \left(H^{-1/2}(\partial\Omega)\right)^3 \middle| \begin{array}{l} \text{there exists } u \in H(\text{curl};\Omega) \\ \text{with } \gamma_t(u) = f \end{array} \right\},$$

with norm

$$\|f\|_{Y(\partial\Omega)} = \inf_{u \in H(\text{curl};\Omega), \gamma_t(u) = f} \|u\|_{H(\text{curl};\Omega)}.$$

With this norm,  $Y(\partial\Omega)$  is a Banach space.

Obviously, this characterization of the trace space is rather unappealing since we have no intrinsic way to judge if a function is in  $Y(\partial\Omega)$  other than by constructing an extension to  $\Omega$ . Even for Lipschitz domains, it turns out that the space can be characterized completely [63]. However, for our purposes the following theorem is sufficient.

**Theorem 3.31** *The space  $Y(\partial\Omega)$  is a Hilbert space. The trace mapping  $\gamma_t : H(\text{curl};\Omega) \rightarrow Y(\partial\Omega)$  is surjective. The map  $\gamma_T : H(\text{curl};\Omega) \rightarrow Y(\partial\Omega)'$  is well-defined. For any  $v \in H(\text{curl};\Omega)$  and  $\varphi \in H(\text{curl};\Omega)$  (3.51)*

$$(\nabla \times v, \varphi) - (v, \nabla \times \varphi) = \langle \gamma_t(v), \gamma_T(\varphi) \rangle_{\partial\Omega}.$$

**Remark 3.32** *For a Lipschitz domain it is known that  $\gamma_T$  is surjective (although we do not prove that fact here) [63]. Later, in Chapter 14, we shall also consider the case when we wish to define traces of functions in  $H(\text{curl};\Omega)$  on surfaces in the interior of  $\Omega$ .*

*Of course,  $Y(\partial\Omega)$  can be characterized precisely [63]. As a hint of what is involved, let us define*

$$H_t^{-1/2}(\partial\Omega) = \left\{ s \in \left( H^{-1/2}(\partial\Omega) \right)^3 \middle| s \cdot v = 0 \text{ almost everywhere on } \partial\Omega \right\}.$$

*Then we know that  $\gamma(\partial\Omega) \subset H^{-1/2}(\partial\Omega)$ . To see that  $\gamma(v)$  has additional smoothness, note that if we choose  $\varphi = \nabla\xi$  for  $\xi \in H^1(\Omega)$  then, according to (3.51), we have  $\langle \gamma(v), \gamma_T(\nabla\xi) \rangle_{\partial\Omega} = (\nabla \times v, \nabla\xi)$ . Using (3.14), we may write this in terms of the surface gradient as  $\langle v \times \nabla v, \nabla_{\partial\Omega}\xi \rangle_{\partial\Omega} = (\nabla \times v, \nabla\xi)$ . Integrating the right-hand side by parts using (3.33) and using the fact that  $\nabla \cdot \nabla \times v = 0$ , we obtain  $\langle v \times \nabla v, \nabla_{\partial\Omega}\xi \rangle_{\partial\Omega} = \langle v \cdot \nabla \times v, \xi \rangle_{\partial\Omega}$ . The left-hand side is the weak definition of the negative of the surface divergence and we have shown that for any  $v \in H(\text{curl};\Omega)$  (3.52)*

$$\nabla_{\partial\Omega} \cdot (v \times v) = -v \cdot (\nabla \times v) \Big|_{\partial\Omega} \text{ in } H^{-1/2}(\partial\Omega).$$

*Hence it turns out that the surface divergence of  $\gamma(v)$  lies in  $H^{1/2}(\partial\Omega)$ . Thus functions in  $Y(\partial\Omega)$  have a well-defined surface divergence. For a smooth surface it turns out that*

$$Y(\partial\Omega) = \left\{ u \in H_t^{-1/2}(\partial\Omega) \middle| \nabla_{\partial\Omega} \cdot u \in H^{-1/2}(\partial\Omega) \right\}$$

*usually denoted by  $H^{1/2}(\text{Div};\partial\Omega)$ . We shall use this space in Chapter 9 (only for smooth domains). The dual space  $Y(\partial\Omega)'$  is (again for smooth domains) given by (3.53)*

$$H^{-1/2}(\text{Curl};\partial\Omega) = \left\{ u \in H_t^{-1/2}(\partial\Omega) \middle| \nabla_{\partial\Omega} \times u \in H^{-1/2}(\partial\Omega) \right\}.$$

*As remarked above, these spaces can also be defined for Lipschitz domains [63].*

**Proof of Theorem 3.31** We follow Chen *et al.*[77]. To prove that  $Y(\partial\Omega)$  is a Hilbert space we first note that, from (3.47), if  $s \in Y(\partial\Omega)$  is such that  $s = \gamma(v)$  for  $v \in H(\text{curl};\Omega)$  then, for any  $\varphi \in (C^\infty(\Omega))^3$ ,

$$\langle s, \varphi \rangle_{\partial\Omega} = (\nabla \times v, \varphi) - (v, \nabla \times \varphi).$$

The right-hand side is well-defined for any  $v, \varphi \in H(\text{curl};\Omega)$ , and using (3.47) and the density of  $(C^\infty(\Omega))^3$  in  $H(\text{curl};\Omega)$  we conclude that the right-hand side is independent of the choice of  $v$  provided  $s = \gamma(v)$ . Thus, again using the density of  $(C^\infty(\Omega))^3$  in  $H(\text{curl};\Omega)$ , we see that the right-hand side of (3.51) is well-defined for  $\varphi \in H(\text{curl};\Omega)$ .

Now we see that for fixed  $s \in Y(\partial\Omega)$ , the map  $L : H(\text{curl};\Omega) \rightarrow \mathbb{R}$  defined by

$$L(\varphi) = \langle s, \gamma(v) \rangle_{\partial\Omega} = (\nabla \times v, \varphi) - (v, \nabla \times \varphi).$$

is a linear functional on  $H(\text{curl};\Omega)$ ; furthermore,  $L$  is bounded since  $|L(\varphi)| \leq C \|v\|_{H(\text{curl};\Omega)} \|\varphi\|_{H(\text{curl};\Omega)}$ . Since  $L$  is independent of  $v$  (provided  $s = \gamma(v)$ ), we may take the infimum and conclude that  $|L(\varphi)| \leq \|s\|_{Y(\partial\Omega)} \|\varphi\|_{H(\text{curl};\Omega)}$ . Hence, by the Riesz representation Theorem 2.17, we know there is a function  $w \in H(\text{curl};\Omega)$  such that

$$\langle s, \gamma(v) \rangle_{\partial\Omega} = (\nabla \times w, \nabla \times \varphi) + (w, \varphi),$$

and using test functions  $\varphi \in (H_0^1(\partial\Omega))^3$  and then  $\varphi \in (H^1(\Omega))^3$  we see that  $w$  satisfies (3.54)

$$\begin{aligned} \nabla \times (\nabla \times w) + w &= 0 \text{ in } \Omega, \\ (\nabla \times w) \times v &= s \text{ on } \partial\Omega. \end{aligned} \tag{3.55}$$

This implies that  $\nabla \times w \in H(\text{curl};\Omega)$  and, by the definition of the  $Y(\partial\Omega)$  norm and (3.54),

$$\|s\|_{Y(\partial\Omega)} \leq \|\nabla \times w\|_{H(\text{curl};\Omega)} \leq \|w\|_{H(\text{curl};\Omega)}.$$

Combining the above results we see that

$$\|s\|_{Y(\partial\Omega)} \leq \inf_{\varphi \in H(\text{curl};\Omega)} \frac{\langle s, \gamma(v) \rangle_{\partial\Omega}}{\|\varphi\|_{H(\text{curl};\Omega)}}$$

Now we can define an inner product on  $Y(\partial\Omega)$  denoted as usual by  $\langle \cdot, \cdot \rangle_{Y(\partial\Omega)}$ . Given  $s_1, s_2 \in Y(\partial\Omega)$  we know that there are functions  $w_1, w_2 \in H(\text{curl};\Omega)$  such that (3.54) and (3.55) hold with  $s$  replaced with  $s_1$  and  $s_2$ , respectively. Then

$$\langle s_1, s_2 \rangle_{Y(\partial\Omega)} = (\nabla \times w_1, \nabla \times w_2) + (w_1, w_2).$$

This verifies that  $Y(\partial\Omega)$  is a Hilbert space. We now see, via (3.51), that  $\gamma_T(\varphi)$  can be interpreted as a function in  $Y(\partial\Omega)'$ .

□

The next theorem gives two alternative characterizations of  $H_0(\text{curl};\Omega)$ .

**Theorem 3.33** Let  $\Omega$  be a bounded Lipschitz domain in  $\mathbb{R}^3$ . Then

$$\begin{aligned} H_0(\text{curl}; \Omega) &= \{ u \in H(\text{curl}; \Omega) \mid \gamma_t(u) = 0 \} \\ &= \{ u \in H(\text{curl}; \Omega) \mid (u, \nabla \times \varphi) = (\nabla \times u, \varphi) \} \\ &\quad \text{forall } \varphi \in \left( C^\infty \left( \frac{\Omega}{\Omega} \right) \right)^3 \}. \end{aligned}$$

**Proof** This theorem follows from Lemma 3.27 and Theorem 3.29. Lemma 3.27 implies that the set

$$\left\{ u \in H(\text{curl}; \Omega) \mid (u, \nabla \times \varphi) = (\nabla \times u, \varphi) \text{ for all } \varphi \in \left( C^\infty(\overline{\Omega}) \right)^3 \right\}$$

is a subset of  $H_0(\text{curl}; \Omega)$ , and the Green's formula in Theorem 3.31 (applied with  $\gamma_t(u) = 0$ ) implies that

$$\begin{aligned} \{ u \in H(\text{curl}; \Omega) \mid \gamma_t(u) = 0 \} &\subset \\ \left\{ u \in H(\text{curl}; \Omega) \mid (u, \nabla \times \varphi) = (\nabla \times u, \varphi) \text{ for all } \varphi \in \left( C^\infty(\overline{\Omega}) \right)^3 \right\}. \end{aligned}$$

Finally, since  $(C_0^\infty(\Omega))^3 \subset \{ u \in H(\text{curl}; \Omega) \mid \gamma_t(u) = 0 \}$  and the set on the right-hand side of this inclusion is closed (because of the continuity of  $\gamma_t$  on  $H(\text{curl}; \Omega)$ ), we conclude that  $H_0(\text{curl}; \Omega) \subset \{ u \in H(\text{curl}; \Omega) \mid \gamma_t(u) = 0 \}$ . These inclusions prove the result.  $\square$

We close with a theorem due to Chen *et al.*[77] (see also [8]) concerning an extension operator for  $H(\text{curl}; \Omega)$ . This is stated without proof.

**Theorem 3.34** Suppose  $\Omega$  is a bounded Lipschitz domain in  $\mathbb{R}^3$  with boundary  $\partial\Omega$  and suppose  $\Omega$  is compactly contained in another domain  $O$ . Then there exists a bounded linear operator  $E : H(\text{curl}; \Omega) \rightarrow H(\text{curl}; \mathbb{R}^3)$  such that  $Eu = u$  in  $\Omega$  and the support of  $Eu$  is contained in  $O$ .

## 3.6 Scalar and vector potentials

We shall need to represent vector functions by a scalar potential, or by a vector potential as appropriate. We need to know when this is possible, and we now present some material mainly from [143] to accomplish this. Classically, it is well-known that a function with vanishing curl can be represented by a scalar potential. More precisely, we have the following theorem.

**Theorem 3.35** Let  $\Omega \subset \mathbb{R}^3$  and suppose  $u \in (C^1(\Omega))^3$  and  $\nabla \times u = 0$  in  $\Omega$  then for every open rectangular parallelepiped  $O \subset \Omega$  there is a scalar function  $\varphi \in C^2(\Omega)$  such that  $u = \nabla \varphi$  in  $O$ .

**Remark 3.36** If  $\Omega$  is a simply connected Lipschitz domain then by taking a union of parallelepipeds we can show that  $u = \nabla \varphi$  in  $\Omega$ . If we require, for example, that  $\varphi$  has zero average value (i.e.  $(\varphi, 1) = 0$ ) then  $\varphi$  is unique. This follows from the Poincaré inequality in Lemma 3.13.

**Proof of Theorem 3.35** This theorem is proved in many classical texts for example [174] and so we only sketch the proof here. The basis of the proof is that one selects a point  $x = (x_1, x_2, x_3)^\top \in \Omega$ . Then for any other point  $y = (y_1, y_2, y_3)^\top$  the function  $\varphi$  is defined by

$$\varphi(y) = \int_{x_1}^{y_1} u_1(r, x_2, x_3) dr + \int_{x_2}^{y_2} u_2(y_1, s, x_3) ds + \int_{x_3}^{y_3} u_3(y_1, y_2, t) dt.$$

In other words,  $\varphi$  is defined by a path integral of  $u$  from  $x$  to  $y$ . From this definition it is clear that  $u_3 = \partial\varphi/\partial y_3$ . But Stokes theorem, and the fact that  $\nabla \times u = 0$ , shows that  $\varphi$  does not depend on the path, thus  $\varphi$  is unchanged if the path from  $x$  to  $y$  first moves in the  $y_2$ -direction, then in the  $y_3$ -direction and finally in the  $y_1$ -direction. Then it is clear that  $u_1 = \partial\varphi/\partial y_1$ . The second component is derived in the same way.  $\square$

Theorem 3.35 is extended to Sobolev spaces in [143] as follows.

**Theorem 3.37** Let  $\Omega$  be a bounded, simply connected Lipschitz domain in  $\mathbb{R}^3$  and suppose that  $u \in (L^2(\Omega))^3$ . Then  $\nabla \times u = 0$  in  $\Omega$  if and only if there exists a scalar potential  $\varphi \in H^1(\Omega)$  such that  $u = \nabla\varphi$  and  $\varphi$  is unique up to an additive constant.

**Proof** Suppose  $\nabla \times u = 0$ , then using the proof from [143] we prove the existence of a scalar potential  $\varphi$  in two steps. First, using convolutions and a special open covering of  $\Omega$  we prove the result on a sequence of strict subdomains of  $\Omega$ . Then we show that this implies the result on  $\Omega$ .

To start, we extend  $u$  to all of  $\mathbb{R}^3$  by zero, and denote the extension  $\tilde{u}$ . Now we smooth  $\tilde{u}$  using convolution. Let  $\rho_\epsilon$  be the smooth function introduced in the proof of Lemma 3.27, and consider the convolution  $\rho_\epsilon * \tilde{u}$ . Then, as in the proof of Lemma 3.27, we have that for  $\epsilon > 0$  (3.56)

$$\rho_\epsilon * \tilde{u} \in \left(C_0^\infty(\mathbb{R}^3)\right)^3, \quad (3.57)$$

$$\lim_{\epsilon \rightarrow \infty} \rho_\epsilon * \tilde{u} = \tilde{u} \text{ in } \left(L^2(\Omega)\right)^3, \quad (3.58)$$

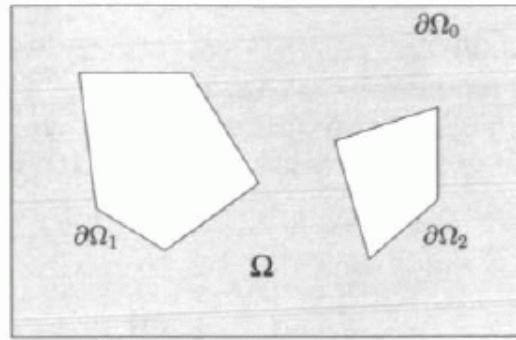
$$\nabla \times \rho_\epsilon * \tilde{u} = \rho_\epsilon * \nabla \times \tilde{u}.$$

Now we use  $\rho_\epsilon * \tilde{u}$  to construct a scalar potential.

Since  $\Omega$  is a bounded, simply connected, Lipschitz domain there is a nested sequence of bounded and simply connected Lipschitz domains  $\{\mathcal{O}_j\}_{j=1}^\infty$  such that  $\bar{\mathcal{O}}_j \subset \Omega$ ,  $\Omega = \cup_{j=1}^\infty \mathcal{O}_j$  and  $\mathcal{O}_j \subset \mathcal{O}_{j+1}$ ,  $1 \leq j < \infty$ . But  $\nabla \times \tilde{u} = \nabla \times u = 0$  in  $\mathcal{O}_j$  and, since  $\rho_\epsilon * \tilde{u}(x)$  is determined by  $\tilde{u}$  in a ball of radius  $\epsilon$  about  $x$ , we have that for fixed  $j$  and  $\epsilon$  small enough  $\nabla \times (\rho_\epsilon * \tilde{u}) = 0$  in  $\mathcal{O}_j$ . Hence, by Theorem 3.35 and the remark following that theorem, there is a continuously differentiable function  $p_\epsilon^j$  such that

$$\rho_\epsilon * \tilde{u} = \nabla p_\epsilon^j \text{ in } \mathcal{O}_j \text{ and } \int_{\mathcal{O}} p_\epsilon^j(x) dx = 0.$$

Fig. 3.2. A diagram of the geometry of  $\Omega$  including labeling of the boundaries. For simplicity this is shown in two dimensions.



The latter integral equality simply fixes  $p_\varepsilon^j$  unambiguously. But, by (3.57),  $\nabla p_\varepsilon^j$  is a convergent in  $(L^2(\Omega))^3$  and hence, by the Poincaré inequality in Lemma 3.13,  $p_\varepsilon^j$  in  $H^1(\Omega)$  as  $\varepsilon \rightarrow 0$ . Thus, on  $\Omega_j$ , we have exhibited a function  $p^j$  such that  $u = \nabla p^j$ .

Now, since  $\Omega_j \subset \Omega_{j+1}$ ,  $\nabla p^j = \nabla p^{j+1}$  on  $\Omega_j$ . So by adjusting  $p^{j+1}$  by a constant, we have that  $p^j = p^{j+1}$  on  $\Omega_j$ . Since this is true for arbitrary  $j$ , we know that there is a function  $p$  on  $\Omega$  such that  $u = \nabla p$  on  $\Omega_j$  and  $p \in L^2(\Omega)$  for all  $j$ . The latter inclusion shows that  $p \in L^2_{\text{loc}}(\Omega)$ . The fact that  $p \in L^2(\Omega)$  is then assured by Lemma 3.11.

On the other hand, if  $u = \nabla \varphi$ , it is immediate that  $\nabla \times u = 0$ .  $\square$

We shall need to use a vector potential to analyze the regularity of solutions of Maxwell's equations. The next theorem gives conditions under which such a potential exists. However, before stating this theorem we need to refine our description of  $\Omega$ . Suppose  $\Omega$  is a bounded, connected, Lipschitz domain. We denote by  $\{\partial\Omega_j\}_{j=0}^J$  the connected components of  $\partial\Omega$ . For definiteness, we define  $\partial\Omega_0$  to denote the component of  $\partial\Omega$  that is the boundary of the unbounded component of  $\mathbb{R}^3 \setminus \Omega$ . For  $j = 0, \dots, J$  let  $\Omega_j$  denote the domain in  $\mathbb{R}^3 \setminus \Omega$  having boundary  $\partial\Omega_j$  (so  $\Omega_0$  is unbounded). Let  $\Omega \subset \mathbb{R}^3$  denote a bounded Lipschitz domain containing  $\Omega$  in its interior. See Fig. 3.2 for a caricature of  $\Omega$  in a special case.

**Theorem 3.38** For any function  $u \in H(\text{div}; \Omega)$  such that

$$\nabla \cdot u = 0 \text{ in } \Omega \text{ and } \langle u \cdot v, 1 \rangle_{\partial\Omega_j} = 0, \quad 0 \leq j \leq J,$$

there exists a vector potential  $A \in (H^1(\Omega))^3$  such that  $u = \nabla \times A$  in  $\Omega$  and  $\nabla \cdot A = 0$  in  $\Omega$ .

**Remark 3.39** The proof uses Fourier transforms (see e.g. [215]). Essentially, the proof is to extend  $u$  to a function  $\tilde{u} \in H(\text{div}; \mathbb{R}^3)$  having compact support. Then  $A$  can be written formally as

$$A = -\Delta^{-1} \nabla \times \tilde{u} \text{ in } \mathbb{R}^3.$$

**Proof of Theorem 3.38** This proof is from [143, 12]. We first note that since  $u \in H(\text{div}; \Omega)$ , the appropriate trace theorem (Theorem 3.9) implies that  $u \cdot v \in H^{1/2}(\partial\Omega)$ . We can thus explicitly extend  $u$  to  $\mathbb{R}^3$  via a scalar potential as follows. On each  $\Omega_j$ ,  $1 \leq j \leq J$ , we define  $p_j \in H^1(\Omega_j)/\mathbb{R}$  to be the solution (see Theorem 3.15) of

$$\begin{aligned} -\Delta p_j &= 0 \quad \text{in } \Omega_j, \\ \frac{\partial p_j}{\partial v} &= u \cdot v \quad \text{on } \partial\Omega_j. \end{aligned}$$

By virtue of the condition  $\langle u \cdot v, 1 \rangle_{\partial\Omega_j} = 0$ , this problem is well-defined and has a unique solution. On  $\Omega_0 \cap \Omega$ , we define  $p_0 \in H^1(\Omega \cap \Omega_0)/\mathbb{R}$  by

$$\begin{aligned} -\Delta p_0 &= 0 \quad \text{in } \Omega \cap \Omega_0, \\ \frac{\partial p_0}{\partial v} &= u \cdot v \quad \text{on } \partial\Omega_0, \\ \frac{\partial p_0}{\partial v} &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

This solution also exists by Theorem 3.15 since  $\langle \partial p / \partial v, 1 \rangle_{\partial(\Omega \cap \Omega_0)} = 0$ . Now we can define  $u \in (L^2(\mathbb{R}^3))^3$  by

$$\tilde{u} = \begin{cases} u & \text{in } \Omega, \\ \nabla p_j & \text{in } \Omega_j \cap \Omega, \quad j = 0, \dots, J, \\ 0 & \text{in } \mathbb{R}^3 \setminus \Omega. \end{cases}$$

Note that  $\nabla \cdot \tilde{u} = 0$  in  $\mathbb{R}^3$  because  $\nabla \cdot \tilde{u} = 0$  on  $\Omega_0 \cap \Omega$ ,  $\mathbb{R}^3 \setminus \Omega$ ,  $\Omega_1, \dots, \Omega_J$ , and  $\Omega$  and the normal component of  $\tilde{u}$  is continuous across common boundaries; see Theorem 5.3. Thus  $\tilde{u} \in H(\text{div}; \mathbb{R}^3)$ .

Now let  $\hat{u}_l$  denote the Fourier transform of the  $l$ th component of  $\tilde{u}$ . If the transform variable is  $\xi$ , let  $\hat{A}$  have components

$$\hat{A}_1 = \frac{\hat{\xi}_3 u_2 - \hat{\xi}_2 u_3}{|\xi|^2}, \quad \hat{A}_2 = \frac{\hat{\xi}_1 u_3 - \hat{\xi}_3 u_1}{|\xi|^2}, \quad \hat{A}_3 = \frac{\hat{\xi}_2 u_1 - \hat{\xi}_1 u_2}{|\xi|^2}.$$

Then it is easy to check, using the fact that  $\tilde{u}$  is divergence free so  $\xi_1 \hat{u}_1 + \xi_2 \hat{u}_2 + \xi_3 \hat{u}_3 = 0$ , that

$$u_1 = \xi_2 \hat{A}_3 - \xi_3 \hat{A}_2, \quad u_2 = \xi_3 \hat{A}_1 - \xi_1 \hat{A}_3, \quad u_3 = \xi_1 \hat{A}_2 - \xi_2 \hat{A}_1.$$

The function  $A$  in the theorem is the inverse Fourier transform of  $\hat{A}$ . We need to show that  $A \in (H^1(\Omega))^3$ . Since

$$\left| \xi_j \hat{A}_l \right| \leq \sum_{l=1}^3 |u_l|,$$

Parseval's identity shows that  $\nabla A_l \in (L^2(\mathbb{R}^3))^3$ ,  $l = 1, 2, 3$ . Now let  $\chi \in C_0^\infty(\mathbb{R}^3)$  be such that  $\chi = 1$  in a neighborhood of the origin. Then writing  $\hat{A}(\xi) =$

$\chi(\xi)\hat{A}(\xi) + (1 - \chi(\xi))\hat{A}(\xi)$  we see that  $\chi\hat{A}$  is of compact support and hence the inverse transform is analytic on  $\mathbf{R}^3$ , and its restriction to  $\Omega$  is in  $(L^2(\Omega))^3$ . The second term  $(1 - \chi)\hat{A}$  vanishes near  $\xi = 0$  and belongs to  $(L^2(\mathbf{R}^3))^3$  and since the Fourier transform preserves the  $(L^2(\mathbf{R}))^3$  norm, we know the inverse Fourier transform of  $(1 - \chi)\hat{A}$  is in  $(L^2(\mathbf{R}^3))^3$  (see Corollary 3.13 of [215]).  $\square$

### 3.7 The Helmholtz decomposition

Frequently, we shall need to write a vector field (e.g. the electric field  $E$ ) in terms of vector and scalar potentials. This decomposition will be termed the *Helmholtz decomposition*. In this section we shall present, without proofs, the related de Rham diagram. As we shall see, this concept is useful when describing interpolation properties of finite element spaces. Because this decomposition is central to our discussion of Maxwell's equations, but we shall not use the full power of the general theory, we delay detailed proofs until we actually use the results.

Before starting our discussion, we need to define more notation related to the domain  $\Omega$ . In this section we do not assume that  $\Omega$  is simply connected, but instead assume that there exist  $L$  open connected surfaces  $\sum_l$ ,  $l = 1, 2, \dots, L$  called *interior cuts* contained in  $\Omega$  such that, for  $1 \leq l \leq L$ ,

- (1) Each surface  $\sum_l$  is an open part of a smooth surface;
- (2)  $\partial\sum_l \subset \partial\Omega$
- (3)  $\sum_l \cap \sum_m = \emptyset$  if  $l \neq m$ ;
- (4) the set  $\Omega^0 = \Omega \setminus \cup_{l=1}^L \Sigma_l$  is simply-connected, and pseudo-Lipschitz by which we mean that for any point  $x \in \partial\Omega$  there is an integer  $r_x$  equal to 1 or 2 and a positive number  $\varrho_0$  such that for all  $\varrho$  with  $0 < \varrho < \varrho_0$  the intersection of  $\Omega$  with the ball with center  $x$  and radius  $\varrho$  has  $r_x$  connected components, each one being a Lipschitz domain.

This definition is from [12]. For a graphical caricature of this geometry see Fig. 3.3 .

It is easy to see that if  $p \in H^1(\Omega)$  then  $\nabla p \in H(\text{curl}; \Omega)$  since  $\nabla \times \nabla p = 0 \in (L^2(\Omega))^3$  and  $\nabla p \in (L^2(\Omega))^3$ . Similarly, if  $u \in H(\text{curl}; \Omega)$ , then  $\nabla \times u \in H(\text{div}; \Omega)$ . These results, and the corresponding result for the divergence applied to functions in  $H(\text{div}; \Omega)$ , can be summarized in the following de Rham diagram:(3.59)

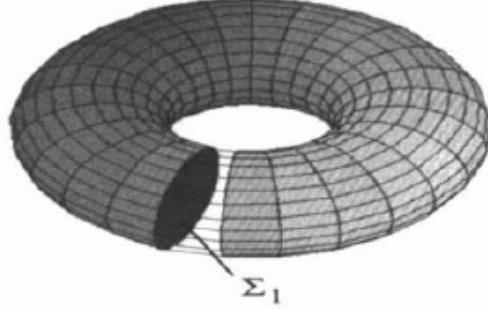
$$H^1(\Omega) / \mathbb{R} \xrightarrow{\nabla} H(\text{curl}; \Omega) \xrightarrow{\nabla \times} H(\text{div}; \Omega) \xrightarrow{\nabla \cdot} L^2(\Omega) .$$

A similar result, with boundary conditions, is(3.60)

$$H_0^1(\Omega) \xrightarrow{\nabla} H_0(\text{curl}; \Omega) \xrightarrow{\nabla \times} H_0(\text{div}; \Omega) \xrightarrow{\nabla \cdot} L^2(\Omega) / \mathbb{R} .$$

Less obvious is that  $\nabla H^1(\Omega)/\mathbb{R}$  is a closed subspace of  $H(\text{curl}; \Omega)$ . We have argued before that  $\nabla H^1(\Omega)/\mathbb{R}$  is contained in the kernel of the curl operator and, moreover, as we shall see, the codimension of  $\nabla H^1(\Omega)/\mathbb{R}$  in this kernel

Fig. 3.3. Example of an interior cut that creates a simply connected pseudo-Lipschitz domain out of a torus. The surface of the torus has partially removed to show the cut  $\Sigma_1$  (marked by darker shading).



is finite dimensional (the codimension of a closed subspace of a Hilbert space is the dimension of its orthogonal complement). Similar results hold for the other spaces and operators in the sequence and are summarized in the following theorem (Theorem 7 of [73]).

**Theorem 3.40** *The diagrams (3.59) and (3.60) have the property that the range of one operator is contained in the kernel of the one following it in the sequence (for example  $\nabla \times (\nabla H^1(\Omega)) = 0$ ). The range space of each operator is a closed subspace of the appropriate kernel with finite codimension.*

Let us now understand the final part of this theorem, and with our applications in mind let us consider diagram (3.60) with boundary conditions. Since  $\nabla H_0^1(\Omega)$  is a closed subspace of the kernel of the curl operator in  $H_0(\text{curl}; \Omega)$ , we may use the projection theorem (Theorem 2.9) to write the null-space of the curl as

$$N(\text{curl}) = \nabla H_0^1(\Omega) \oplus \left( \nabla H_0^1(\Omega) \right)^\perp,$$

with orthogonality in the  $H(\text{curl}; \Omega)$  inner product. Now suppose  $u \in N(\text{curl}) \subset H_0(\text{curl}; \Omega)$  and  $u \in (\nabla H_0^1(\Omega))^\perp$ . Then  $\nabla \times u = 0$  in  $\Omega$  and  $v \times u = 0$  on  $\partial\Omega$ , and

$$\int_{\Omega} u \cdot \nabla \xi \, dV = 0$$

for all  $\xi \in H_0^1(\Omega)$ . Thus  $\nabla \cdot u = 0$  in  $\Omega$ . This argument may also be reversed so that we conclude that

$$\left( \nabla H_0^1(\Omega) \right)^\perp = \{u \in H_0(\text{curl}; \Omega) \mid \nabla \times u = 0, \nabla \cdot u = 0 \text{ in } \Omega\}.$$

The above space is sufficiently important to have its own name. We denote the *normal cohomology space* by  $K_N(\Omega)$  defined by

$$K_N(\Omega) = \{u \in H_0(\text{curl}; \Omega) \mid \nabla \times u = 0, \nabla \cdot u = 0 \text{ in } \Omega\}.$$

A similar analysis shows that the orthogonal complement of  $\nabla \times H_0(\text{curl}; \Omega)$  in the kernel of the divergence is given by the tangential cohomology space  $K_T(\Omega)$  defined by

$$K_T(\Omega) = \{w \in H_0(\text{curl}; \Omega) \mid \nabla \cdot w = 0, \nabla \times w = 0 \text{ in } \Omega\}.$$

The previous theorem shows that  $\dim(K_N(\Omega)) < \infty$  and  $\dim(K_T(\Omega)) < \infty$ .

Note that  $K_N(\Omega)$  and  $K_T(\Omega)$  arise naturally in electrostatics, or magnetostatics, respectively. Suppose we seek a solution of the Maxwell system in a perfectly conducting cavity  $\Omega$  that is independent of time (i.e. a steady state or static solution) when  $\epsilon = 1$ ,  $\sigma = 0$  and  $J = 0$  (i.e. in a vacuum with no source). Then

$$\begin{aligned} \nabla \times E &= 0 \quad \text{in } \Omega \quad (\text{static field}), \\ \nabla \cdot E &= 0 \quad \text{in } \Omega \quad (\text{no sources}), \\ \mathbf{v} \times E &= 0 \quad \text{on } \partial\Omega \quad (\text{perfect conductivity}). \end{aligned}$$

Thus  $E \in K_N(\Omega)$ . Similarly, a source free magnetostatic field lies in  $K_T(\Omega)$ . It is thus desirable to obtain a characterization of these spaces that is amenable to computation. Before doing this, we pause to summarize the situation so far:

**Theorem 3.41** (Theorem 8 of [73])

- (1) If  $u \in H_0(\text{curl}; \Omega)$  is such that  $\nabla \times u = 0$  in  $\Omega$  then there exists a unique scalar potential  $p \in H_0^1(\Omega)$  and function  $f_N \in K_N(\Omega)$  such that

$$u = \nabla p + f_N.$$

- (2) If  $w \in H_0(\text{div}; \Omega)$  is such that  $\nabla \cdot w = 0$  in  $\Omega$  then there is a vector potential  $A \in H_0(\text{curl}; \Omega)$  and a function  $f_T \in K_T(\Omega)$  such that

$$w = \nabla \times A + f_T.$$

The vector potential is unique if we require in addition that  $\nabla \cdot A = 0$  in  $\Omega$  and  $(v \cdot A, 1)_{\Gamma_j} = 0$ ,  $j = 0, 1, \dots, J$ .

Now let us characterize  $K_N(\Omega)$  using the argument from [12]. Let  $u \in K_N(\Omega)$  and define

$$\Theta_0 = \left\{ g \in H^1(\Omega) \mid g|_{\partial\Omega_0} = 0 \text{ and } g \Big|_{\partial\Omega_j} = \text{constant}, 1 \leq j \leq J \right\}.$$

Then obviously  $\nabla \Theta_0 \subset K_N(\Omega)$  and we may define  $p \in \Theta_0$  by requiring that (3.61)

$$\begin{aligned} \square \tilde{u} \cdot v, 1 \square_{\partial\Omega_j} &= \square \tilde{u} \cdot v, \xi \square_{\partial\Omega} \\ &= \square u \cdot v, \xi \square_{\partial\Omega} - \square \frac{\partial p}{\partial v}, \xi \square_{\partial\Omega} \\ &= \square u \cdot v, \xi \square_{\partial\Omega} - \square \nabla p, \nabla \xi \square = 0. \end{aligned}$$

This problem has a unique solution by the Lax–Milgram Lemma 2.21 and the Poincaré inequality in Lemma 3.13. Now consider  $\bar{u} = u - \nabla p$ . Then clearly

$\nabla \times \tilde{u} = 0$  in  $\Omega$  and choosing  $\xi = 1$  on  $\partial\Omega$ , and  $\xi = 0$  on  $\partial\Omega_j, j \neq l$ , in (3.61) we have

$$\begin{aligned} [\tilde{u} \cdot v, 1]_{\partial\Omega_l} &= [\tilde{u} \cdot v, \xi]_{\partial\Omega} \\ &= [u \cdot v, \xi]_{\partial\Omega} - \left[ \frac{\partial p}{\partial v}, \xi \right]_{\partial\Omega} \\ &= [u \cdot v, \xi]_{\partial\Omega} - [\nabla p, \nabla \xi] = 0. \end{aligned}$$

Hence, via Theorem 3.38, there is a vector potential  $A \in (H^1(\Omega))^3$  such that  $\tilde{u} = \nabla \times A$ . Using the volume Stokes theorem (3.51) we have

$$(\tilde{u}, \tilde{u}) = (\tilde{u}, \nabla \times A) = (\nabla \times \tilde{u}, A) - \langle u \times \tilde{u}, A \rangle_{\partial\Omega} = 0$$

Thus  $u = \nabla p$  for some  $p \in H_0^1(\Omega)$  and  $\Delta p = 0$ . Examination of this proof shows that we have proved:

**Theorem 3.42** *The dimension of  $K_N(\Omega)$  is  $J$  and  $K_N(\Omega)$  is spanned by functions  $\nabla p_j, 1 \leq j \leq J$ , where  $p_j \in H^1(\Omega)$  satisfies*

$$\Delta p_j = 0 \text{ in } \Omega, \text{ and } p_j = \delta_{j,s} \text{ on } \partial\Omega_s, \quad 0 \leq s \leq J.$$

**Remark 3.43** *In addition  $\langle \partial p_j / \partial v, 1 \rangle_{\partial\Omega_s} = \delta_{js}$ ,  $1 \leq j \leq J$ , and*

$$\langle \partial p_j / \partial u, 1 \rangle_{\partial\Omega_0} = -1.$$

A similar, but more involved, proof gives the following result (recall that  $\Omega^0 = \Omega \setminus (\cup_{l=1}^L \Sigma_l)$ ) [12]:

**Theorem 3.44** *The dimension of  $K_T(\Omega)$  is  $L$  (the number of interior cuts).  $K_T(\Omega)$  is spanned by  $\nabla p_l, 1 \leq l \leq L$ , where  $p_l \in H^1(\Omega^0)$  and satisfies*

$$\begin{aligned} \Delta p_l &= 0 && \text{in } \Omega^0, \\ \frac{\partial}{\partial v} p_l &= 0 && \text{on } \partial\Omega, \\ [p_l]_{\Sigma_s} &= \text{constant}, & 1 \leq s \leq L, \\ \left[ \frac{\partial p_l}{\partial v} \right]_{\Sigma_s} &= 0, & 1 \leq s \leq L, \\ \left[ \frac{\partial p_l}{\partial v}, 1 \right]_{\Sigma_s} &= \delta_{l,s}, & 1 \leq s \leq L. \end{aligned}$$

*The function  $p_l$  is unique up to a constant.*

From our point of view, the important lesson of this theory is the following Helmholtz or Hodge decomposition (henceforth referred to as the Helmholtz decomposition) of  $(L^2(\Omega))^3$ . This will be fully proved later, after Theorem 4.5, in the form we need for the analysis of Maxwell's equations.

**Theorem 3.45** Every  $u \in (L^2(\Omega))^3$  has the decomposition

$$u = \nabla p + f_N + \nabla \times A$$

for unique  $p \in H_0^1(\Omega)$ ,  $f_N \in K_N(\Omega)$  and

$$\begin{aligned} A \in \{w \in H(\text{curl}; \Omega) \mid & \nabla \cdot w = 0 \text{ in } \Omega, \quad u \cdot w = 0 \text{ on } \partial \Omega \\ & \text{and } \langle w \cdot u, 1 \rangle_{\Sigma_l} = 0 \quad 1 \leq l \leq L\} \end{aligned}$$

**Remark 3.46** A similar decomposition holds with  $p \in H^1(\Omega)$ ,  $f \in K_T(\Omega)$  and  $A \in H_0(\text{curl}; \Omega)$ .

To simplify the presentation in the remainder of the book we are going to assume that  $\Omega$  is simply connected, so  $L = 0$ , and that  $\partial\Omega$  consists of two components  $\partial\Omega_0 = \Sigma$  and  $\partial\Omega_1 = \Gamma$ . Thus  $J = 1$ . At the expense of more complex notation and spaces, we could easily include the case  $J > 1$  in our discussion. More complicated is the case  $L > 0$ . For a discussion of these aspects, see [12, 73, 71, 164].

## 3.8 A function space for the impedance problem

When solving problems involving the impedance boundary condition, we need to use a subspace of  $H(\text{curl}; \Omega)$ . This is the space  $H_{\text{imp}}(\text{curl}; \Omega)$  defined by

$$H_{\text{imp}}(\text{curl}; \Omega) = \left\{ u \in H(\text{curl}; \Omega) \mid u \times u \in L_t^2(\partial \Omega) \right\}$$

with the graph norm (recall that  $L_t^2(\partial \Omega)$  is defined in (3.13))

$$\|u\|_{H_{\text{imp}}(\text{curl}; \Omega)}^2 = \|u\|_{(L^2(\Omega))^3}^2 + \|u\|_{(L^2(\Omega))^3}^2 + \|u\|_{L_t^2(\partial \Omega)^3}^2$$

The choice of boundary data in  $L_t^2(\partial \Omega)$  is justified in [265].

Our next goal is to prove that  $(C^\infty(\Omega))^3$  is dense in this space. To do that we need to establish some more properties of functions in this space. We start with a basic regularity estimate for functions in  $H_{\text{imp}}(\text{curl}; \Omega)$  due to Costabel [102]. For a more complete description of the regularity of the solutions of Maxwell's equations see [44, 106].

**Theorem 3.47** Let  $\Omega$  be a bounded Lipschitz domain in  $\mathbb{R}^3$ . Suppose that  $u \in H(\text{curl}; \Omega) \cap H(\text{div}; \Omega)$ , and  $u \times v \in (L^2(\partial \Omega))^3$ . Then  $u \in (H^{1/2}(\Omega))^3$  and the following norm estimate holds: (3.62)

$$\begin{aligned} \|u\|_{(H^{1/2}(\Omega))^3}^2 \leq & C \left( \|u\|_{(L^2(\Omega))^3}^2 + \|\nabla \times u\|_{(L^2(\Omega))^3}^2 \right. \\ & \left. + \|\nabla \cdot u\|_{L^2(\Omega)} + \|u \times v\|_{(L^2(\partial \Omega))^3} \right). \end{aligned}$$

Similarly, suppose  $u \in H(\text{curl}; \Omega) \cap H(\text{div}; \Omega)$ , and  $u \cdot v \in (L^2(\partial \Omega))^3$ . Then  $u \in (H^{1/2}(\Omega))^3$  and the following norm estimate holds: (3.63)

$$\begin{aligned} \|u\|_{(H^{1/2}(\Omega))^3} &\leq C \left( \|u\|_{(L^2(\Omega))^3} + \|\nabla \times u\|_{(L^2(\Omega))^3} \right. \\ &\quad \left. + \|\nabla \cdot u\|_{L^2(\Omega)} + \|u \cdot v\|_{(L^2(\partial\Omega))^3} \right). \end{aligned}$$

**Remark 3.48** Note that on a convex or smooth domain with  $u \times v = 0$  on the boundary, we can conclude (by the same argument) that  $u \in (H^1(\Omega))^3$  [266], but on general domains this inclusion does not hold. The difficulty with using  $(H^1(\Omega))^3$  elements on general domains is that solutions of Maxwell's equations cannot always be approximated by functions in  $(H^1(\Omega))^3$  (see [105, 114]).

**Proof of Theorem 3.47** The proof is from [143] as modified by [102]. As noted in [207, 12], it suffices to prove the result in the case when  $\Omega$  is simply connected with connected boundary. The general case can be reduced to this case by noting that a general  $\Omega$  can be covered by a finite union of open sets  $O_k$ ,  $k = 1, \dots, K$ , such that  $\Omega_k = O_k \cap \Omega$  is Lipschitz and starlike. Then we can introduce a partition of unity  $\{\chi_k\}_{k=1}^K$  such that the support of  $\chi_k$  is in  $O_k$  (see [215] pp. 83-85). Proving the result for each  $\chi_k u$  on  $\Omega_k$  (now a simply connected Lipschitz domain with connected boundary) and adding the results proves the general case. We shall now prove this result. For simplicity, we shall drop the subscripts on  $\Omega_k$  and  $\chi_k u$ . Thus we assume that  $\Omega$  is bounded and simply connected with connected boundary  $\partial\Omega$ .

The method of proof is to factor  $u$  into terms with well-understood regularity. Let  $f = \nabla \times u \in (L^2(\Omega))^3$ . Then  $\nabla \cdot f = 0$  in  $\Omega$  and so, by Theorem 3.38 (there is now no constraint on the boundary since we have reduced to a connected boundary), there is a vector potential  $w \in (H^1(\Omega))^3$  such that  $\nabla \times w = f$  and  $\nabla \cdot w = 0$  in  $\Omega$ . The construction of  $w$  implies that  $\|w\|_{(H^1(\Omega))^3} \leq C \|\nabla \times u\|_{(L^2(\Omega))^3}$ .

Now let  $\zeta = u - w$ . Then  $\nabla \times \zeta = 0$  and  $\zeta \in (L^2(\Omega))^3$ . Hence, since we have reduced to the case of a simply connected domain, Theorem 3.37 implies that there is a scalar potential  $p \in H^1(\Omega)$  such that  $\zeta = \nabla p$ , and  $p$  satisfies  $\Delta p = \nabla \cdot \zeta$  in  $\Omega$ . Since  $\nabla \cdot \zeta \in L^2(\Omega)$ , we can construct a function  $q \in H^1(\Omega)$  such that  $\Delta q = \nabla \cdot \zeta$  in  $\Omega$ . Such a function can be constructed by convolution of  $\nabla \cdot \zeta$  (extended by zero to  $\mathbb{R}^3$ ) with the fundamental solution of Laplace's equation on all  $\mathbb{R}^3$  in the same way as the existence of a vector potential was proved in Lemma 3.38. Then  $p = q + r$ , where  $\Delta r = 0$  in  $\Omega$  and  $\zeta = \nabla(q + r)$ .

Let us now consider the first estimate of the theorem, so we assume  $v \times u \in (L^2(\partial\Omega))^3$ . By the definition of  $\zeta$ ,

$$z \times u = (u - w) \times u$$

Since  $w \in (H^1(\Omega))^3$ , the trace theorem shows that  $w|_{\partial\Omega} \in (H^{1/2}(\partial\Omega))^3$  and so  $\nabla(q + r) \times v = \zeta \times v \in (L^2(\partial\Omega))^3$ . Hence  $\nabla r \times v \in (L^2(\partial\Omega))^3$  and we see that  $r \in H^1(\partial\Omega)$ . By the regularity result in Theorem 3.17,  $r \in H^{3/2}(\Omega)$ , and recalling that  $u = w + \nabla_q + \nabla_r$  we have (3.64)

$$\begin{aligned} \|r\|_{H^{3/2}(\Omega)} &\leq C \|\nabla r \times v\|_{(L^2(\partial\Omega))^3} \\ &\leq C \left( \|u \times v\|_{(L^2(\partial\Omega))^3} + \|w \times v\|_{(L^2(\partial\Omega))^3} + \|\nabla q \times v\|_{(L^2(\partial\Omega))^3} \right) \\ &\leq C \left( \|u \times v\|_{(L^2(\partial\Omega))^3} + \|w\|_{(H^{1/2}(\partial\Omega))^3} + \|q\|_{(H^1(\partial\Omega))} \right) \\ &\leq C \left( \|u \times v\|_{(L^2(\partial\Omega))^3} + \|\nabla \times u\|_{(L^2(\partial\Omega))^3} + \|\nabla \cdot u\|_{L^2(\Omega)} \right). \end{aligned}$$

Thus,  $u \in (H^{1/2}(\Omega))^3$  and

$$\begin{aligned}\|u\|_{(H^{1/2}(\Omega))^3} &\leq C \left( \|w\|_{(H^{1/2}(\Omega))^3} + \|\nabla q\|_{(H^{1/2}(\Omega))^3} + \|\nabla r\|_{(H^{1/2}(\Omega))^3} \right) \\ &\leq C \left( \|\nabla \times u\|_{L^2(\Omega)}^3 + \|\nabla \cdot u\|_{L^2(\Omega)}^3 + \|\nabla r\|_{(H^{1/2}(\Omega))^3} \right),\end{aligned}$$

and the term in  $\nabla r$  is estimated using (3.64).

In the case  $v \cdot u \in L^2(\partial\Omega)$ , we see that  $z \cdot v = u \cdot v - w \cdot v \in L^2(\partial\Omega)$  and

$$\frac{\partial}{\partial \mathbf{u}}(q+r) = z \cdot \mathbf{u} \in L^2(\partial\Omega).$$

Since  $q \in H^2(\Omega)$ , we know that  $\partial q/\partial v \in L^2(\partial\Omega)$  and can conclude, via Theorem 3.17, that  $r \in H^{3/2}(\Omega)$ . The proof continues as before, replacing  $v \times u$  by  $v \cdot u$  in the appropriate places.  $\square$

The following spaces will be useful for our analysis of  $H_{\text{imp}}(\text{curl}; \Omega)$ : (3.65)

$$\begin{aligned}X_N &= \{u \in H(\text{curl}; \Omega) \cap H(\text{div}; \Omega) \mid \mathbf{v} \times u = 0 \text{ on } \partial\Omega\}, \\ X_N &= \{u \in X_N \mid \nabla \cdot u = 0 \text{ in } \partial\Omega\},\end{aligned}\tag{3.66}$$

$$\begin{aligned}W_N &= \{u \in H(\text{curl}; \Omega) \cap H(\text{div}; \Omega) \mid \nabla \cdot u = 0 \text{ in } \Omega \text{ and} \\ &\quad \mathbf{v} \times u \in L_t^2(\partial\Omega)\},\end{aligned}\tag{3.67}$$

$$\begin{aligned}X_T &= \{u \in H(\text{curl}; \Omega) \cap H(\text{div}; \Omega) \mid \mathbf{v} \cdot u = 0 \text{ on } \partial\Omega\}, \\ X_{T,0} &= \{u \in X_T \mid \nabla \cdot u = 0 \text{ in } \Omega\},\end{aligned}\tag{3.68}$$

$$\begin{aligned}W_T &= \{u \in H(\text{curl}; \Omega) \cap H(\text{div}; \Omega) \mid \nabla \cdot u = 0 \text{ in } \Omega \text{ and} \\ &\quad \mathbf{v} \cdot u \in L^2(\partial\Omega)\},\end{aligned}\tag{3.70}$$

with the obvious graph norms in each case. Now we can prove the next lemma via the previous theorem and the compact imbedding of  $(H^{1/2}(\Omega))^3$  in  $(L^2(\Omega))^3$  (see the Sobolev Imbedding Theorem 3.5).

**Corollary 3.49** *If  $\Omega$  is a bounded Lipschitz domain, the spaces  $X_p$ ,  $X_N$ ,  $W_N$  and  $W_T$  are all compactly imbedded in  $(L^2(\Omega))^3$ .*

For a Lipschitz polyhedron  $\Omega$ , and with homogeneous boundary data, Theorem 3.47 can be improved. The following result is Proposition 3.7 of [12].

**Theorem 3.50** *Let  $\Omega$  be a bounded Lipschitz polyhedron. Suppose  $u \in X_N$  or  $u \in X_T$  then there is a  $\delta > 0$  such that for all  $s$  with  $0 \leq s \leq \delta$ , the function  $u \in (H^{1/2+s}(\Omega))^3$  and the following a priori estimate holds:*

$$\|u\|_{(H^{1/2+s}(\Omega))^3} \leq C \left\{ \|\nabla \times u\|_{(L^2(\Omega))^3}^3 + \|\nabla \cdot u\|_{L^2(\Omega)} + \|u\|_{(L^2(\Omega))^3} \right\}.$$

**Proof** The proof proceeds as in the proof of Theorem 3.47 by first reducing to a simply connected Lipschitz domain with connected (actually polyhedral) boundary. We then introduce a bounded, connected and simply connected Lipschitz domain  $O$  containing  $\Omega$ . Given  $u \in X_N$ , let  $\tilde{z}$  denote the extension of  $\nabla \times u$  by zero to  $O$ . Because  $u \in X_N$ , we know that  $\tilde{z} \in H(\text{div}; O)$  and  $\nabla \cdot \tilde{z} = 0$  in  $O$ . Hence, via Theorem 3.38, we know there exists a vector potential  $w \in (H^1(O))^3$  such that  $\tilde{z} = \nabla \times w$  and  $\nabla \cdot w = 0$  in  $O$ . Clearly  $\nabla \times (u - w) = 0$  in  $O$  so that by Theorem 3.37 there is a scalar  $p \in H^1(O)$  such that  $u - w = \nabla p$  in  $\Omega$  and  $-w = \nabla p$  in  $O \setminus \Omega$ . In  $\Omega$  the potential  $p$  satisfies

$$\begin{aligned}\Delta p &= \nabla \cdot u \quad \text{in } \Omega, \\ \nabla \times \nabla p &= -\nabla \times w \quad \text{on } \partial\Omega.\end{aligned}$$

Since  $w \in (H^1(O))^3$ , we know that  $p \in H^2(O \setminus \Omega)$  and, taking the trace of  $p$  from the exterior of  $\Omega$ , we see that for  $0 \leq s < 1/2$  we have  $p \in H^{1+s}(\partial\Omega)$ . Via Theorem 3.18, we conclude  $p \in H^{3/2+s}(\Omega)$  for  $0 \leq s < s_\Omega$ . Thus  $u = w + \nabla p \in (H^{1/2+s}(\Omega))^3$ ,  $0 \leq s < s_\Omega \leq 1/2$ .

The same argument is made for  $u \in X_p$ , except now  $p$  satisfies a Neumann problem and the use of Theorem 3.18 shows that  $p \in H^{3/2+s}(\Omega)$  for  $0 \leq s < \tilde{s}_\Omega$ . Taking  $\delta$  with  $0 < \delta < \min(s_\Omega, \tilde{s}_\Omega)$  proves the result.  $\square$

The next result shows that, under suitable conditions, the norm  $\|u\|_{(L^2(\Omega))^3}$  of a function in  $W_N$  or  $W_T$  can be estimated from its curl and boundary values. For similar results, see [266, 197, 198] and [143]. For extensions to mixed boundary conditions, see [195]. Results of this type are usually referred to as *Friedrichs inequalities*.

**Corollary 3.51** Suppose that  $\Omega$  is a bounded Lipschitz domain. If  $\Omega$  is simply connected, and has a connected boundary, there is a constant  $C > 0$  such that for every  $u \in W_N$

$$\|u\|_{(L^2(\Omega))^3} \leq C \left( \|\nabla \times u\|_{(L^2(\Omega))^3} + \|\nabla \times u\|_{(L^2(\partial\Omega))^3} \right).$$

The same inequality holds for  $W_T$  with  $\nabla \times u$  replaced by  $\nabla \cdot u$ .

**Remark 3.52** Using this result and Theorem 3.50, we see that if  $\Omega$  is a bounded Lipschitz polyhedron and if  $\Omega$  is simply connected with a connected boundary then there is a  $\delta$  with  $0 < \delta \leq 1/2$  such that for  $u \in X_{N,0}$  or  $u \in X_{T,0}$ , we have  $u \in (H^{s+1/2}(\Omega))^3$  for  $0 \leq s \leq \delta$  and we have

$$\|u\|_{(H^{s+1/2}(\Omega))^3} \leq C \left( \|\nabla \times u\|_{(L^2(\Omega))^3} \right).$$

**Proof of Corollary 3.51** We start with the proof for  $W_N$ . Suppose the result is false. Then there exists a sequence of functions  $\{u_n\}_{n=1}^\infty \subset W_N$  such that

$$\|\nabla \times u_n\|_{(L^2(\Omega))^3} + \|\nabla \times u_n\|_{(L^2(\partial\Omega))^3} \leq 1/n$$

and  $\|u_n\|_{(L^2(\Omega))^3} = 1$  for all  $n$ . By the compactness of the imbedding of  $W_N$  in  $(L^2(\Omega))^3$  (Corollary 3.49) there is a subsequence, still denoted  $\{u_n\}_{n=1}^\infty$ , such

that  $u_n \rightarrow u$  in  $(L^2(\Omega))^3$  as  $n \rightarrow \infty$  (and weakly in  $W_N$ ) for some  $u \in W_N$ . But since  $\|u_n\|_{(L^2(\Omega))^3} \rightarrow 1$  as  $n \rightarrow \infty$ , this implies that  $u \neq 0$ . On the other hand,  $\|\nabla \times u\|_{(L^2(\Omega))^3} = 0$ . And so, since  $\nabla \times u = 0$  in  $\Omega$ , we conclude by Theorem 4.3 that  $u = \nabla p$  for some  $p \in H^1(\Omega)$ . However, since  $\nabla \cdot u = 0$  in  $\Omega$ , this implies that  $\Delta p = 0$  in  $\Omega$ . The fact that  $\nabla \times u = 0$  then shows that  $\nabla_{\partial\Omega} p = 0$  and since there is only one component to  $\partial\Omega$  this implies, perhaps shifting  $p$  by a constant, that  $p = 0$  on  $\partial\Omega$ . The uniqueness of the solution of the Dirichlet problem for Laplace's equation then shows  $p = 0$  in  $\Omega$  and hence  $u = 0$ . This is a contradiction.

The proof for  $W_T$  proceeds similarly. We need only show that if  $\nabla \times u = 0$  in  $\Omega$  and  $u \in W_T$  then  $u = 0$ . Proceeding as before, we see that  $u = \nabla p$  for some  $p \in H^1(\Omega)$  that satisfies  $\Delta p = 0$  in  $\Omega$  and  $\partial p/\partial\nu = 0$  on  $\partial\Omega$ . The uniqueness of the solution of the Neumann problem for Laplace's equation then shows that  $p$  is constant, and we may select  $p = 0$ . Hence  $u = 0$  in this case also.  $\square$

We need one more auxiliary result from [35]. Let us consider the following subspace of  $H_{\text{imp}}(\text{curl};\Omega)$ :

$$\tilde{H}_{\text{imp}}(\text{curl};\Omega) = \{u \in H_{\text{imp}}(\text{curl};\Omega) \mid (\nabla \times u) \cdot \nu = 0 \text{ on } \partial\Omega\}.$$

Note that since  $\nabla \times u \in H(\text{div};\Omega)$ , Theorem 3.24 shows that the trace  $(\nabla \times u) \cdot \nu$  on  $\partial\Omega$  is well-defined.

**Lemma 3.53** *The space  $(C^\infty(\Omega))^3$  is dense in  $\tilde{H}_{\text{imp}}(\text{curl};\Omega)$ .*

**Proof** The proof is from [35]. As in the proof of Theorem 3.47, we may assume that  $\Omega$  is simply connected with connected boundary. Suppose  $u \in \tilde{H}_{\text{imp}}(\text{curl};\Omega)$ . Let  $A \in X_{T,0}$  satisfy (3.71)

$$\int_{\Omega} \nabla \times A \cdot \nabla \times \varphi \, dV = \int_{\Omega} u \cdot \nabla \times \varphi - \nabla \times u \cdot \varphi \, dV$$

for all  $\varphi \in X_{T,0}$ . Using the Friedrichs inequality in Corollary 3.51 and the Lax–Milgram Lemma 2.21 we see that this problem has a unique solution.

Because any function in  $\varphi \in X_T$  may be written as  $\varphi = (\varphi - \nabla \xi) + \nabla \xi$ , where  $\xi \in H^1(\Omega)$  satisfies

$$\nabla \xi = \nabla \cdot \varphi \text{ in } \Omega \text{ and } \frac{\partial \xi}{\partial \nu} = 0 \text{ on } \partial\Omega,$$

we see that  $\varphi = \varphi' + \nabla \xi$ , where  $\varphi' \in X_{T,0}$  and  $\xi \in H^1(\Omega)$ . Using the test function  $\varphi = \nabla \xi$  in both sides of (3.71), we see that the equation also holds for this function. Thus (3.71) holds for all  $\varphi \in X_T$  and hence holds for any test function in  $(C_0^\infty(\Omega))^3$ . Indeed, using such a test function we easily verify that  $\nabla \times (\nabla \times A) = 0$  in  $\Omega$ . In addition, using the definition of  $X_T$  in (3.68) together with (3.51) we conclude that (3.72)

$$A \cdot \nu = 0 \text{ and } u \times (\nabla \times A) = u \times u \text{ on } \partial\Omega.$$

Now let us write  $u = (u - \nabla \times A) + \nabla \times A = 0$  and the boundary of the domain is connected we see, via Theorem 3.38, that  $\nabla \times A = \nabla p$  for some  $p \in H^1(\Omega)$ . But since

$$\nabla_{\partial\Omega} p = (\mathbf{u} \times \nabla p) \times \mathbf{u} = (\mathbf{u} \times (\nabla \times A)) \times \mathbf{u} = (\mathbf{u} \times u) \times \mathbf{u} \in L_t^2(\Omega),$$

we conclude that  $p \in H^1(\partial\Omega)$ . Hence, via the density result in Lemma 3.3, we know that  $p$  may be approximated to arbitrary accuracy by functions in  $C^\infty(\Omega)$ , and so  $\nabla \times A$  can be approximated by functions in  $(C^\infty(\Omega))^3$ .

Now consider  $u - \nabla \times A$ . Note that by (3.72),  $\mathbf{v} \times (u - \nabla \times A) = 0$  so that  $\mathbf{v} - \nabla \times A \in H_0(\text{curl};\Omega)$ , and (3.42) guarantees that it can be approximated to arbitrary accuracy by functions in  $(C^\infty(\Omega))^3$ .  $\square$

Now, at last, we can state and prove the main result of this section

**Theorem 3.54** *The space  $(C^\infty(\Omega))^3$  is dense in  $H_{\text{imp}}(\text{curl};\Omega)$ .*

**Proof** This proof is from [35]. Again, we can assume that  $\Omega$  is simply connected with a connected boundary (see the start of the proof of Theorem 3.47). Let  $u \in H_{\text{imp}}(\text{curl};\Omega)$  and define  $p \in H^1(\Omega)/\mathbb{R}$  by(3.73)

$$\int_{\Omega} \nabla p \cdot \nabla \xi \, dV = \int_{\Omega} \nabla \times u \cdot \nabla \xi \, dV$$

for all  $\xi \in H^1(\Omega)/\mathbb{R}$ . By the Poincaré inequality in Lemma 3.13, and the Lax–Milgram Lemma 2.21 this problem has a unique solution.

Since  $\nabla \cdot (\nabla p) = 0$ , using Theorem 3.38, there is a function  $A \in (H^1(\Omega))^3$  such that  $\nabla \cdot A = 0$  and  $\nabla p = \nabla \times A$  in  $\Omega$ . But  $A$  can be approximated in  $(H^1(\Omega))^3$  to arbitrary accuracy using functions in  $(C^\infty(\Omega))^3$  (see Theorem 3.2) and hence can be approximated by the same functions in  $H_{\text{imp}}(\text{curl};\Omega)$ .

Now we write  $u = A + (u - A)$ . We see that for any  $\xi \in H^1(\Omega)$ , using (3.33) and the definition of  $p$  in (3.73),

$$\begin{aligned} \int_{\partial\Omega} \mathbf{v} \cdot \nabla \times (u - A) \xi \, dA &= \int_{\Omega} \nabla \times (u - A) \cdot \nabla \xi \, dV \\ &= \int_{\Omega} (\nabla \times u - \nabla p) \cdot \nabla \xi \, dV = 0. \end{aligned}$$

Thus  $\mathbf{v} \cdot \nabla \times (u - A) = 0$  on  $\partial\Omega$ , and so  $u - A \in H_{\text{imp}}(\text{curl};\Omega)$ . Use of Lemma 3.53 completes the proof.  $\square$

The space  $X_N$  defined in (3.65) is the appropriate space for the “ellipticized” Maxwell’s equations in a metallic (or perfectly conducting) cavity (see Section 7.4 for more on this approach). This space is, however, rather dangerous. The norm on  $X_N$  is

$$\|u\|_{X_N}^2 = \|\nabla \times u\|_{(L^2(\Omega))^3}^2 + \|\nabla \cdot u\|_{L^2(\Omega)}^2 + \|u\|_{(L^2(\Omega))^3}^2,$$

and if we are to use  $X_N$  we need to know that smooth functions are dense in this space using this norm.

For a smooth function  $u \in (C^\infty(\Omega))^3$  satisfying the boundary conditions characterizing  $X_N$ , integration by parts shows that

$$\|\nabla u\|_{(L^2(\Omega))^3}^2 + \|\nabla \times u\|_{(L^2(\Omega))^3}^2 + \|\nabla \cdot u\|_{L^2(\Omega)}^2.$$

Of course, this is not true for arbitrary  $u \in X_N$ . However, with this observation the following lemma is plausible (although the proof is quite difficult, and can be found in [105, 51, 103] — we do not give it here because it is not central to our study).

**Lemma 3.55** *Let  $\Omega$  be a bounded Lipschitz polyhedral domain and define  $C_N^\infty = (C_0^\infty(\overline{\Omega}))^3 \cap X_N$ . Then  $C_N^\infty$  is dense in  $(H^1(\Omega))^3 \cap X_N$ .*

The difficulty with using  $X_N$  occurs when  $(H^1(\Omega))^3 \cap X_N \neq X_N$ . Unfortunately, this occurs whenever  $\Omega$  is non-convex. To be more precise we follow [112] and define

$$D(\Delta^{\text{Dir}}) = \left\{ \varphi \in H_0^1(\Omega) \mid \Delta \varphi \in L^2(\Omega) \right\}.$$

It is easy to see that  $\nabla D(\Delta^{\text{Dir}})$  is a closed subset of  $X_N$ . But when there are re-entrant corners or edges on  $\partial\Omega$  this space contains singular functions that are not in  $H^1(\Omega)$  and hence which have gradients that are not in  $(H^1(\Omega))^3$ . This would not be a problem if such functions could be approximated by functions in  $(H^1(\Omega))^3 \cap X_N$ , but this is not the case. We consider the decomposition of  $D(\Delta^{\text{Dir}})$  defined by

$$D(\Delta^{\text{Dir}}) = (H^2(\Omega) \cap H_0^1(\Omega)) \oplus K_{\text{Dir}}.$$

Thus we decompose a function in  $D(\Delta^{\text{Dir}})$  into a smooth part in  $H^2(\Omega)$  and a singular part in  $K_{\text{Dir}}$  as is usual when analyzing solutions of Poisson's problem on a polyhedral domain. The functions in  $\nabla(H^2(\Omega) \cap H_0^1(\Omega))$  are contained in  $(H^1(\Omega))^3 \cap X_N$  and in fact the following lemma holds [103].

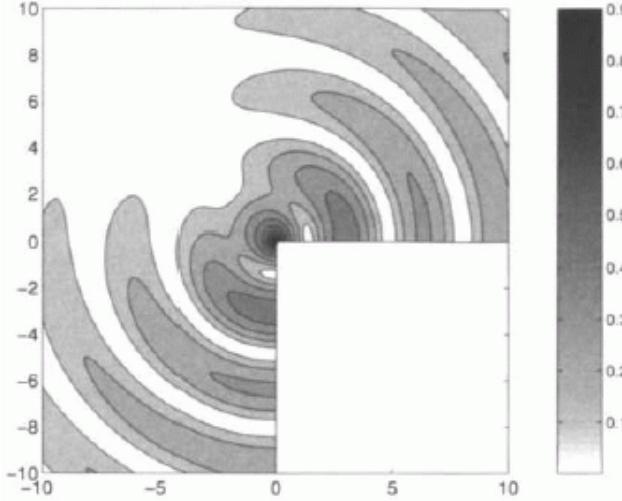
**Lemma 3.56** *Let  $\Omega$  be a bounded Lipschitz polyhedron then*

$$X_N = \left( (H^1(\Omega))^3 \cap X_N \right) \oplus \nabla K_{\text{Dir}}.$$

But  $K_{\text{Dir}}$  is an infinite-dimensional space of singular functions associated with the re-entrant corners and edges of  $\partial\Omega$ . Thus, on a non-convex domain,  $C_N^\infty$  is not dense in  $X_N$  and neither is  $(H^1(\Omega))^3 \cap X_N$ . It follows that the use of finite element functions in  $(H^1(\Omega))^3 \cap X_N$  with convergence in the  $X_N$  norm will fail on a non-convex domain (in particular, care must be taken when using standard vector continuous piecewise polynomial elements). This is discussed further in Section 7.4, where we note that Costabel *et al.*[114] have proposed a remedy using appropriate weighted spaces (i.e. using a weaker norm than the  $X_N$  norm).

To show that this is a concrete concern, we consider the following example from Leis [207]. Consider the L-shaped domain in  $\mathbb{R}^2$  given by  $L = (-10, 10)^2 \setminus ([0, 10] \times (-10, 0])$ , see Fig. 3.4 (similar examples can be constructed in  $\mathbb{R}^3$  using

Fig. 3.4. The L-shaped domain for the example of singular solutions of Maxwell's equations. We plot  $|u|$  so that the singularity is at the origin and the edges of the domain are parallel to the coordinate axes (of course, we have truncated the solution which is infinite at  $(0, 0)$  for graphical purposes).



the exterior of a cone). The domain for the electromagnetic field is  $\Omega = L \times R$ . We seek solutions of Maxwell's equations independent of  $x_3$  lying in the  $(x_1, x_2)$  plane. Hence, if  $u = (E_1, E_2)^\top$ , we see that Maxwell's equations imply that

$$\vec{\nabla}_{\mathcal{L}} \times \vec{\nabla}_{\mathcal{L}} \times u - k^2 u = 0 \quad \text{in } \mathcal{L}$$

where the differential operators are the planar scalar and vector curl defined in Section 3.4. Let us assume that  $u$  satisfies the perfect conducting boundary condition on edges meeting at the origin  $(0, 0)$ , and allow general non-homogeneous boundary data on the remaining edges of  $L$ . Choosing appropriate non-homogeneous boundary data, we can see that one solution of this boundary value problem is

$$u = \vec{\nabla}_{\mathcal{L}} \times U = \begin{pmatrix} \partial U / \partial x_2 \\ -\partial U / \partial x_1 \end{pmatrix},$$

where  $(\varrho, \varphi)$  are the cylindrical coordinates of  $(x_1, x_2)$  and  $U$  is a solution of the scalar Helmholtz equation  $\Delta U + k^2 U = 0$  in  $L$  satisfying zero Neumann data on the edges of the L-shaped domain meeting at  $(0, 0)$  and non-homogeneous Neumann boundary data elsewhere. Leis [207] shows that for  $\alpha = \frac{2}{3}$  one such solution is given by  $U = J_\alpha(\varrho) \cos(\alpha\varphi)$ , where  $J_\alpha$  is the Bessel function of first kind and order  $\alpha$ .

A plot of  $|u|$  for this choice of  $U$  is shown in Fig. 3.4, where the singularity (of course truncated for graphical purposes) at the origin is clearly visible. Note that an asymptotic expansion of  $u$  near  $\varrho = 0$  shows that  $u = O(\varrho^{1/3})$  near  $\varrho = 0$  and hence, since  $|\nabla u| = O(\varrho^{-4/3})$ , we see that  $u \notin (H^1(L))^2$ . This provides an example, which can be generalized to domains exterior to cones in

three dimensions, of the fact that solutions of Maxwell's equation can be rather singular near the boundary.

### 3.9 Curl or divergence conserving transformations

In finite element theory we often wish to transform between different geometric domains. We need to ensure that the transformed function has a well-defined gradient, curl or divergence, as appropriate. Suppose that  $\hat{K}$  and  $K$  are two bounded domains in  $\mathbb{R}^3$  (e.g. the reference tetrahedron and a tetrahedron in the mesh). Suppose  $F_K : \hat{K} \rightarrow K$  is a continuously differentiable, one-to-one and onto map, and  $\det(dF_K)$  is one sign on  $\hat{K}$ .

A scalar function  $\hat{p} \in H(\hat{K})$  is transformed to a scalar function  $p$  on  $K$  by(3.74)

$$p \circ F_K = \hat{p},$$

where  $\circ$  denotes composition of functions. It is then obvious that  $p \in H(K)$  since the chain rule implies that(3.75)

$$\nabla p = (dF_K)^{-T} \hat{\nabla} \hat{p},$$

where  $\nabla$  denotes the gradient with respect to the coordinate system for  $K$ . We shall adopt the same convention for  $\nabla \cdot$  and  $\nabla \times$ . The equality (3.75) is proved in [80].

Vector functions must be transformed in a more careful way to conserve their properties. Suppose  $\hat{u} \in H(\text{curl}; \hat{K})$  and we wish to associate with  $\hat{u}$  a function  $u$  defined on  $K$  in  $H(\text{curl}; K)$ . Since  $\hat{\nabla} \hat{p} \in H(\text{curl}; \hat{K})$  and  $\nabla p \in H(\text{curl}; K)$  (when  $\hat{p}$  and  $p$  are related by (3.74)) we see that we must transform  $\hat{u}$  to  $u$  via the transformation (3.75) so that(3.76)

$$u \circ F_K = (dF_K)^{-T} \hat{u},$$

where  $dF_K$  is the Jacobian matrix defined by

$$(dF_K)_{l,m} = \frac{\partial(F_K)_l}{\partial \hat{x}_m}, \quad 1 \leq l, m \leq 3.$$

Then we have the following result found in the notes of Dubois [133] and Appendix A of [82].

**Lemma 3.57** Suppose  $u$  and  $\hat{u}$  are related by(3.76) where  $F_K : \hat{K} \rightarrow K$  is a continuously differentiable, invertible and surjective mapping. Let  $[\nabla \times u]$  denote the  $3 \times 3$  matrix with

$$[\nabla \times u]_{i,j} = \frac{\partial u_i}{\partial x_j} - \frac{\partial u_j}{\partial x_i}.$$

Then

$$[\nabla \times u] \circ F_K = dF_K^{-T} [\hat{\nabla} \times \hat{u}] dF_K^{-T}.$$

Hence if  $\hat{u} \in H(\text{curl}; \hat{K})$  then  $u \in H(\text{curl}; K)$ .

**Proof** Note that the change of variables can be written

$$u_i = \sum_{k=1}^3 \frac{\partial \hat{x}_k}{\partial x_i} \hat{u}_k.$$

With this expansion

$$\frac{\partial u_i}{\partial x_j} = \frac{\partial}{\partial x_i} \left( \sum_{k=1}^3 \frac{\partial \hat{x}_k}{\partial x_i} \hat{u}_k \right) = \sum_{k=1}^3 \left( \frac{\partial^2 \hat{x}_k}{\partial x_i \partial x_j} \hat{u}_k + \sum_{l=1}^3 \frac{\partial \hat{x}_k}{\partial x_i} \frac{\partial \hat{x}_l}{\partial x_j} \frac{\partial \hat{u}_k}{\partial x_l} \right).$$

Similarly,

$$\frac{\partial u_i}{\partial x_i} = \sum_{k=1}^3 \left( \frac{\partial^2 \hat{x}_k}{\partial x_i \partial x_j} \hat{u}_k + \sum_{l=1}^3 \frac{\partial \hat{x}_k}{\partial x_j} \frac{\partial \hat{x}_l}{\partial x_i} \frac{\partial \hat{u}_k}{\partial x_l} \right).$$

Subtracting these expressions gives

$$\begin{aligned} [\nabla \times u]_{i,j} &= \sum_{k=1}^3 \sum_{l=1}^3 \frac{\partial \hat{x}_k}{\partial x_i} \left( \frac{\partial \hat{u}_k}{\partial x_l} - \frac{\partial \hat{u}_l}{\partial x_k} \right) \frac{\partial \hat{x}_l}{\partial x_j} \\ &= \sum_{k=1}^3 \sum_{l=1}^3 (\mathrm{d}F^{-1})_{k,i} [\hat{\nabla} \times \hat{u}]_{k,l} (\mathrm{d}F^{-1})_{l,j}. \end{aligned}$$

This verifies the change of variable for the curl considered as a matrix and completes the proof.  $\square$

It is now possible, using for example MAPLE, to show that for any skew symmetric matrix  $C$  of the form

$$C = \begin{pmatrix} 0 & c_1 & c_2 \\ -c_1 & 0 & c_3 \\ -c_2 & -c_3 & 0 \end{pmatrix}$$

and any invertible matrix  $B$ , if

$$A = B^{-T} C B^{-1} \text{ and } d = \frac{1}{\det(B)} B \begin{pmatrix} C_3 \\ -C_2 \\ C_1 \end{pmatrix}$$

then  $A_{2,3} = d_1$ ,  $A_{1,3} = -d_2$  and  $A_{1,2} = d_3$ . We have thus verified the following corollary.

**Corollary 3.58** Under the conditions of Lemma 3.57, suppose  $\hat{u} \in H(\mathrm{curl}; \hat{K})$  and that  $u$  and  $\hat{u}$  are related by (3.76). Then  $u \in H(\mathrm{curl}; K)$  and

$$(\nabla \times u) \circ F_K = \frac{1}{\det(\mathrm{d}F_K)} \mathrm{d}F_K \hat{\nabla} \times \hat{u}.$$

Now we wish to prove a similar result for the divergence. In view of the fact that if  $\hat{u} \in H(\text{curl};\hat{K})$  then  $\nabla \times \hat{u} \in H(\text{div};\hat{K})$  we see that we must transform a function in  $\hat{w} \in H(\text{div};\hat{K})$  to  $w \in H(\text{div};K)$  via(3.77)

$$w \circ F_K = \frac{1}{\det(dF_K)} \hat{dF}_K \hat{w}.$$

To verify that this is a good transformation we prove the following lemma:

**Lemma 3.59** Suppose  $w$  and  $\hat{w}$  are differentiable functions related by (3.77), where  $F_K : \hat{K} \rightarrow K$  is a continuously differentiable, invertible and surjective mapping. Then

$$\nabla \cdot w = \frac{1}{\det(dF_K)} \hat{\nabla} \cdot \hat{w}.$$

Hence if  $\hat{w} \in H(\text{div};\hat{K})$  then  $w \in H(\text{div};K)$ .

**Proof** Suppose  $\hat{w} \in H(\text{div};\hat{K})$  and  $w \in H(\text{div};K)$  are related by (3.77). Then transforming the following integral from  $K$  to  $\hat{K}$  using  $F_K$ , for  $p \in C_0^\infty(K)$  we obtain, using (3.74) and (3.20),

$$\begin{aligned} \int_K \nabla \cdot w p dV &= - \int_K w \cdot \nabla p dV \\ &= \int_{\hat{K}} \frac{1}{\det(dF_K)} \left( \hat{dF}_K \hat{w} \right) \cdot \left( \hat{dF}_K^T \hat{\nabla} \hat{p} \right) |\det(dF_K)| d\hat{u} \\ &= - \text{sign}(\det(dF_K)) \int_{\hat{K}} \hat{w} \cdot \hat{\nabla} \hat{p} d\hat{u} \\ &= - \text{sign}(\det(dF_K)) \int_{\hat{K}} \hat{\nabla} \cdot \hat{w} \hat{p} d\hat{u}, \end{aligned}$$

where  $\text{sign}(\det(dF_K)) = \det(dF_K)/|\det(dF_K)|$ . Now we transform back to  $K$  treating  $\nabla \cdot \hat{w}$  as a scalar function using (3.74) to obtain(3.78)

$$\int_K \nabla \cdot w p dV = \int_K \left( \hat{\nabla} \cdot \hat{w} \right) p \frac{1}{\det(dF_K)} dV.$$

Since this holds for all  $p \in C_0^\infty(K)$ , we have the desired result.  $\square$

Some further useful facts that can help in the implementation of a finite element method are as follows. First, let  $\hat{v}$  be the unit outward normal to  $\hat{K}$ . Then if  $\mathcal{O} \in \partial\hat{K}$  and  $v$  is defined by(3.79)

$$v \left( F_K(x) \right) = \frac{\hat{dF}_K^{-T} \hat{v}}{|\hat{dF}_K^{-T} \hat{v}|} \left( x \right),$$

we know that  $v$  is a unit normal to  $K$ . Second, let  $\hat{\tau}$  is any unit vector tangent to  $\partial\hat{K}$  at  $\mathcal{O}$ . Then if  $\tau$  is given by(3.80)

$$\tau \left( F_K(x) \right) = \frac{\hat{dF}_K^{-1} \hat{\tau}}{|\hat{dF}_K^{-1} \hat{\tau}|} \left( x \right)$$

we know that  $\tau$  is a unit vector tangent to  $\partial K$  at  $F_K(\mathcal{O})$  (see p. 265 of [143]).

It is also necessary to know how surface and line integrals transform under  $F_K$ . For example, if we define the surface Jacobian by

$$J_{\hat{\mathbf{u}}}(\hat{\mathbf{x}}) = |\det(dF_K)| \left| (dF_K)^{-T} \hat{\mathbf{u}} \right|,$$

then, if  $p$  and  $\hat{p}$  are related by (3.74), we have [61](3.81)

$$\int_{\partial K} p dA = \int_{\partial \hat{K}} \hat{p} J_{\hat{\mathbf{u}}} d\hat{A}.$$

Here  $\hat{\mathbf{v}}$  is the unit outward normal to  $\hat{K}$ . In addition for  $v$  and  $\hat{v}$  related by (3.77) and  $\varphi$  and  $\hat{\varphi}$  related by (3.74) we have

$$\int_{\partial K} \mathbf{u} \times \mathbf{v} \cdot \varphi dA = \text{sign}(\det(dF_K)) \int_{\partial \hat{K}} \hat{\mathbf{u}} \cdot \hat{\mathbf{v}} \cdot \hat{\varphi} d\hat{A}.$$

This relationship is used to conclude that mapped divergence conforming elements are still divergence conforming on the mapped element (see Chapter 5). Similarly if  $u$  and  $\hat{u}$  are related by (3.76), and if  $v$  and  $\hat{v}$  related in the same way then(3.82)

$$\int_{\partial K} \mathbf{u} \times \mathbf{v} \cdot w_T dA = \text{sign}(\det(dF_K)) \int_{\partial \hat{K}} \hat{\mathbf{u}} \times \hat{\mathbf{v}} \cdot \hat{w}_T d\hat{A}.$$

This latter result follows from Theorem 3.29. Using that theorem, if  $u, v \in H(\text{curl}; K)$  and if (3.76) is used together with the conclusion of Corollary 3.58, we have

$$\begin{aligned} \int_{\partial K} v \times u \cdot w_T dA &= \int_K (\nabla \times u \cdot w - u \cdot \nabla \times w) dV \\ &= \int_{\hat{K}} \frac{|\det(dF_K)|}{\det(dF_K)} \left( (dF_K \hat{\nabla} \times \hat{u}) \cdot (dF_K^{-T} \hat{w}) \right) \\ &\quad - \left( dF_K^{-T} \hat{u} \right) \cdot (dF_K \hat{\nabla} \times \hat{w}) d\hat{V} \\ &= \text{sign}(\det(dF_K)) \int_{\hat{K}} (\hat{\nabla} \times \hat{u} \cdot \hat{w} - \hat{u} \cdot \hat{\nabla} \times \hat{w}) d\hat{V} \\ &= \text{sign}(\det(dF_K)) \int_{\partial \hat{K}} \hat{\mathbf{v}} \times \hat{\mathbf{u}} \cdot \hat{w}_T d\hat{A}, \end{aligned}$$

which completes the derivation of the desired equality.

# 4 VARIATIONAL THEORY FOR THE CAVITY PROBLEM

## 4.1 Introduction

Finite element methods are based on variational or weak formulations of boundary value problems. Before proceeding to discretize the Maxwell system, we, therefore, need to establish a reliable variational formulation. In this chapter we shall develop a standard variational formulation of the cavity problem (1.25) and show that it has a unique solution. Of course, there are many possible variational formulations for this problem, including those based on the first-order Maxwell system (1.10), but the one we shall develop is commonly used as the basis of finite element methods. In order to perform the analysis of this variational problem we will have to expand on the discussion of appropriate function spaces presented in the previous chapter. First we formally derive a variational formulation of the cavity problem using the Galerkin method.

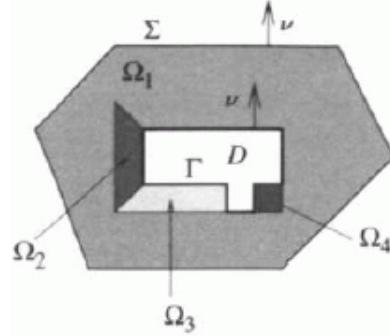
We assume that the domain  $\Omega \subset \mathbb{R}^3$  occupied by the electromagnetic field is bounded and simply connected. In addition, we assume that  $\Omega$  is a Lipschitz polyhedron (see Definition 3.1). In fact, for this section we could drop the assumption that  $\Omega$  is a polyhedron (that assumption comes into play when we discuss finite element methods). The boundary of  $\Omega$  is assumed to consist of at most two connected components, denoted  $\Sigma$  and  $\Gamma$  (either of these may be the empty set if there is only one component to the boundary). For a schematic of the domain, see Fig. 4.1 .

Using the Galerkin method, we can find a variational formulation for the cavity problem as follows. Taking the dot product of (1.25a) by the complex conjugate of a smooth vector function  $\varphi$  (called the test function) and integrating over  $\Omega$ , and then using the integration by parts formula (3.51) we obtain(4.1)

$$\begin{aligned} & \int_{\Omega} \left[ \left( \mu_r^{-1} \nabla \times E \right) \cdot \nabla \times \bar{\varphi} - \kappa^2 (\epsilon_r E) \cdot \bar{\varphi} \right] dV + \int_{\partial\Omega} v \times \left( \mu_r^{-1} \nabla \times E \right) \cdot \bar{\varphi}_T dA \\ &= \int_{\Omega} F \cdot \bar{\varphi} dV, \end{aligned}$$

where  $\partial\Omega$  denotes the boundary of  $\Omega$ ,  $\bar{\varphi}$  is the complex conjugate of  $\varphi$  and  $\varphi_T = (v \times \varphi) \times v$  on  $\partial\Omega$ . Recalling our assumption that  $\partial\Omega = \Sigma \cup \Gamma$ , we now need to take account of the boundary conditions. On  $\Gamma$  the perfect conducting boundary condition (1.25b) gives no information about  $v \times (\mu_r^{-1} \nabla \times E)$ , so we eliminate this portion of the integral by choosing  $\varphi$  such that  $\varphi_T = 0$  or, equivalently,  $v \times \varphi = 0$  on  $\Gamma$ . On  $\Sigma$  the impedance boundary condition (1.25c)

Fig. 4.1. Geometry of the cavity. The impenetrable scatterer occupies the domain  $D$  with boundary  $\Gamma$  and is surrounded, in this case, by a medium made up of materials of differing electromagnetic properties occupying subdomains  $\Omega_j, j = 1, \dots, 4$ . These subdomains are Lipschitz, and on these subdomains the functions  $\mu_r$  and  $\epsilon_r$  are also suitably smooth (see the text). The impedance boundary condition is applied on the boundary component  $\Sigma$ .



gives  $\mu_r^{-1}(\nabla \times E) \times \nu = i\kappa \lambda E_T + g$ . Using this in (4.1) we obtain (taking data terms to the right-hand side)(4.2)

$$\begin{aligned} & \int_{\Omega} \left[ (\mu_r^{-1} \nabla \times E) \cdot \nabla \times \varphi - \kappa^2 (\epsilon_r E) \cdot \varphi \right] dV - i\kappa \int_{\Sigma} \lambda E_T \cdot \varphi_T dA \\ &= \int_{\Omega} F \cdot \varphi dV + \int_{\Sigma} g \cdot \varphi_T dA. \end{aligned}$$

From this we can see that in order for all the integrals to be well defined, we should use the space  $X$  defined by(4.3)

$$X = \left\{ u \in H(\text{curl}; \Omega) \mid \nabla \times u = 0 \text{ on } \Gamma \text{ and } u_T \in (L^2(\Sigma))^3 \text{ on } \Sigma \right\}.$$

Thus, we can state the variational or weak cavity problem as follows. Given  $F \in (L^2(\Omega))^3$  and  $g \in L^2(\Sigma)$ , we wish to find  $E \in X$  such that (4.2) is satisfied for all  $\varphi \in X$ .

In order to simplify notation, we recall the following inner products. For any  $u, v \in (L^2(\Omega))^3$ ,

$$(u, v) = \int_{\Omega} u \cdot \bar{v} dV$$

and for any  $u, v \in (L^2(\Sigma))^3$

$$\langle u, v \rangle = (u, v)_{\Sigma} = \int_{\Sigma} u \cdot \bar{v} dA.$$

Using this notation the variational cavity problem is to find  $E \in X$  such that(4.4)

$$\left( \mu_r^{-1} \nabla \times E, \nabla \times \varphi \right) - \kappa^2 (\epsilon_r E, \varphi) - i\kappa \langle \lambda E_T, \varphi_T \rangle = (F, \varphi) + \langle g, \varphi_T \rangle$$

for all  $\varphi \in X$ . For future reference we define the sesquilinear form  $a: X \times X \rightarrow \mathbb{C}$  as follows.(4.5)

$$a(u, v) = \left( \mu_r^{-1} \nabla \times u, \nabla \times v \right) - \kappa^2 (\epsilon_r u, v) - i\kappa \langle \lambda u_T v_T \rangle$$

for all  $u, v \in X$ .

The remainder of this chapter is devoted to showing that, under appropriate assumptions on the coefficients and domain, problem (4.4) has a unique solution. There are two aspects of Maxwell's equations that make this theory interesting. First, the curl operator has a large null space (the curl of the gradient of a function is zero) and this null space, which is related to the divergence condition (1.11a), must be removed from  $X$  using the Helmholtz decomposition. This will require an understanding of some of the properties of the space  $X$ , which we present in Section 4.3 . Second, the presence of the term  $-\kappa^2(\epsilon_r E, \varphi)$  on the lefthand side of (4.4) means that the left-hand side is not a coercive sesquilinear form. To counter this problem, we can use the Fredholm alternative (Theorem (2.33)) to provide conditions under which we are guaranteed the existence of a solution. This in turn requires a separate proof that (4.4) has at most one solution. The result of this analysis will be a well-understood variational problem suitable for discretization by the finite element method.

## 4.2 Assumptions on the coefficients and data

So far we have been rather lax in specifying our assumptions on the coefficients and data in (4.4). We correct that in this section. Remember our standing assumption that  $\Omega$  is a bounded, simply connected, Lipschitz polyhedral domain with boundary  $\partial\Omega$  consisting of two disjoint connected components  $\Sigma$  and  $\Gamma$ .

The coefficients  $\epsilon_r$  and  $\mu_r$  in (4.4) are assumed to be piecewise smooth. In particular, we assume that  $\Omega$  may be decomposed into  $P$  subdomains denoted  $\Omega_p, p = 1, \dots, P$ , (see Fig. 4.1 ) such that

- (1)  $\Omega = \cup_{p=1}^P \Omega_p$  where  $\Omega$  denotes the closure of  $\Omega$ ;
- (2)  $\Omega_p \cap \Omega_q = \emptyset$ , if  $p \neq q$ ;
- (3) each subdomain  $\Omega_p, p = 1, \dots, P$ , is connected and has a Lipschitz boundary;
- (4) the coefficient  $\mu_r$  is constant on each subdomain (of course, a different constant is allowed from subdomain to subdomain);
- (5) the coefficient  $\epsilon_r$  is assumed to have the following properties:
  - the restriction of  $\epsilon_r$  to  $\Omega_p$  is a function in  $H^3(\Omega_p)$ ,
  - there is a constant  $c > 0$  such that for each  $p, p = 1, \dots, P$ , either  $\Im(\epsilon_r) \geq c$  on  $\Omega_p$  or  $\Im(\epsilon_r) = 0$  on  $\Omega_p$ .

The unusual assumption that  $\epsilon_r|_{\Omega_p} \in H^3(\Omega_p)$  requires some comment! Using the Sobolev Imbedding Theorem 3.5, we can see that this implies  $\epsilon_r \in C^1(\Omega_p)$ . In addition, we shall need to extend such functions outside  $\Omega_p$  while maintaining smoothness. Thanks to the Calderon Extension Theorem 3.2, this is possible.

The assumption that  $\mu_r$  is piecewise constant is used to simplify the proof of uniqueness for the solution of (4.4) and can be avoided as we shall comment in Section 4.6 .

The impedance function  $\lambda$  is assumed to be a strictly-positive real-valued function of position on the boundary  $\Sigma$  with  $\lambda \in L^\infty(\Sigma)$ . Physically, we expect  $\lambda > 0$ , but the case  $\lambda = 0$  is useful for imposing a symmetry boundary condition on the field and will be handled later in Chapter 11 .

The data functions  $\mathbf{F}$  and  $\mathbf{g}$  are assumed to be such that  $\mathbf{g} \in L^2_t(\Sigma)$  and  $\mathbf{F} \in (L^2(\Omega))^3$ . Of course in specific applications we may have more smoothness. In particular, for scattering problems we often have that  $\mathbf{F} = 0$  and  $\mathbf{g} \in (H(\Sigma))^3$  for any  $s \geq 0$ . We could allow  $\mathbf{F}$  to be more general, so that all the arguments of this section are valid if  $\mathbf{F} \in X'$ , where  $X'$  is the dual space of  $X$ . This allows, for example, surface delta functions (i.e. sheets of charge) but not line delta functions (crude models of wires) [73]. However, the presentation of the finite element method would be more complex, and so we generally stick with the easier assumption of square integrability.

## 4.3 The space $X$ and the nullspace of the curl

The space  $X$  was introduced in the previous section as the natural space for posing problem (4.4). Here we will give some of its relevant properties which follow from the properties of  $H_0(\text{curl}; \Omega)$  and  $H_{\text{imp}}(\text{curl}; \Omega)$  introduced in Sections 3.5.3 and 3.8, respectively. Then we continue by identifying the null-space of the curl operator, so we can rewrite (4.4) by removing this null-space.

**Theorem 4.1** *The space  $X$  defined in(4.3)when equipped with the inner product  $(\mathbf{u}, \mathbf{v})_x$  defined, for each  $\mathbf{u}, \mathbf{v} \in X$ , by*

$$(\mathbf{u}, \mathbf{v})_X = (\mathbf{u}, \mathbf{v}) + (\nabla \times \mathbf{u}, \nabla \times \mathbf{v}) + \langle u_T, v_T \rangle$$

*is a Hilbert space. The following space is dense in  $X$ :*

$$\mathcal{X} = \left\{ \mathbf{u} \mid \mathbf{u} = \mathbf{w}|_\Omega \text{ for some } \mathbf{w} \in C_0^\infty(\mathbb{R}^3 \setminus D) \right\},$$

*where  $D$  is the domain exterior to  $\Gamma$  viewed from  $\Omega$ .*

**Remark 4.2** *Of course, the inner product defined in this theorem gives rise to the norm  $\| \cdot \|_x$  defined for every  $\mathbf{u} \in X$  by*

$$\|\mathbf{u}\|_X^2 = \|\mathbf{u}\|_{H(\text{curl}; \Omega)}^2 + \|u_T\|_{(L^2(\Sigma))^3}^2.$$

**Proof of Theorem 4.1** To see that the set  $X$  is well-defined, we need only note that, by Theorem 3.29, for any  $\mathbf{u} \in H(\text{curl}; \Omega)$ , the trace  $\mathbf{v} \times \mathbf{u}$  makes sense as a function in  $(H^{-1/2}(\Sigma))^3$ , hence  $u_T$  is defined.

To see that  $X$  is a Hilbert space, we need to show that it is complete. Note that for any  $\mathbf{u}, \mathbf{v} \in X$ ,  $\langle u_T, v_T \rangle = \langle \mathbf{v} \times \mathbf{u}, \mathbf{u} \times \mathbf{v} \rangle$ , so we can use  $\|\mathbf{v} \times \mathbf{u}\|_{(L^2(\Sigma))^3}$  in place of the corresponding norm on  $u_T$  in the definition of  $X$ . Now suppose we have

a Cauchy sequence  $\{u_n\}_{n=1}^\infty$  in  $X$ . Then this is certainly a Cauchy sequence in  $H(\text{curl}; \Omega)$  and  $\{v \times u_n\}_{n=1}^\infty$  is a Cauchy sequence in  $(L^2(\Sigma))^3$ , so there are functions  $\mathbf{u} \in H(\text{curl}; \Omega)$  and  $v \in (L^2(\Sigma))^3$  such that  $u_n \rightarrow \mathbf{u}$  in  $H(\text{curl}; \Omega)$  and  $v \times u_n \rightarrow v \times \mathbf{u}$  in  $(L^2(\Sigma))^3$ . On the other hand, in  $(H^{-1/2}(\Sigma))^3$ ,  $v \times u_n \rightarrow v \times \mathbf{u}$  since the trace operator is continuous on  $H(\text{curl}; \Omega)$ . Thus  $v \times \mathbf{u} = v$  and we have verified the desired completeness.

To prove the density result, let  $\varphi \in C_0^\infty(\mathbb{R}^3)$  be such that  $\varphi = 1$  in the neighborhood of  $\Gamma$ ,  $\varphi = 0$  in the neighborhood of  $\Sigma$  and  $0 \leq \varphi \leq 1$ . If  $\mathbf{u} \in X$  then  $\varphi \mathbf{u} \in H_0(\text{curl}; \Omega)$ , and so by Theorem 3.26, given  $\varepsilon > 0$ , there is a function  $u_s^{(1)} \in (C_0^\infty(\Omega))^3$  such that  $\|\varphi \mathbf{u} - u_s^{(1)}\| \leq \varepsilon/2$ . On the other hand,  $(1 - \varphi)\mathbf{u} \in H_{\text{imp}}(\text{curl}; \Omega)$ . Hence by Lemma 3.54, for any  $\delta > 0$ , there is a function  $u_s^{(2)} \in (C^\infty(\Omega))^3$  such that  $\|(1 - \varphi)\mathbf{u} - u_s^{(2)}\|_v \leq \delta$ . Now let  $\varphi_1 \in C^\infty(\mathbb{R}^3)$  be such that  $\varphi_1 = 0$  in the neighborhood of  $\Gamma$  and  $\varphi_1 = 1$  in the support of  $(1 - \varphi)$ . Then  $u_s^{(1)} + \varphi_1 u_s^{(2)} \in X$  and

$$\begin{aligned} & \left\| \mathbf{u} - \left( u_s^{(1)} + \varphi_1 u_s^{(2)} \right) \right\|_X \\ & \leq \left\| \varphi \mathbf{u} - u_s^{(1)} \right\|_X + \left\| (1 - \varphi) \mathbf{u} - u_s^{(2)} \right\|_X + \left\| (1 - \varphi_1) u_s^{(2)} \right\|_X \\ & = \left\| \varphi \mathbf{u} - u_s^{(1)} \right\|_X + \left\| (1 - \varphi) \mathbf{u} - u_s^{(2)} \right\|_X + \left\| (1 - \varphi_1) (u_s^{(2)} - (1 - \varphi) \mathbf{u}) \right\|_X \\ & \leq \left\| \varphi \mathbf{u} - u_s^{(1)} \right\|_X + \left\| (1 - \varphi) \mathbf{u} - u_s^{(2)} \right\|_X + \|1 - \varphi_1\|_{W^{1,\infty}(\Omega)} \\ & \quad \times \left\| u_s^{(2)} - (1 - \varphi) \mathbf{u} \right\|_X \\ & \leq \varepsilon/2 + \delta + \|1 - \varphi_1\|_{W^{1,\infty}(\Omega)} \delta. \end{aligned}$$

Picking  $\delta$  such that  $\delta + \|1 - \varphi_1\|_{W^{1,\infty}(\Omega)} \delta < \varepsilon/2$  completes the proof.  $\square$

Next we need to characterize the functions in  $X$  which have vanishing curl, since we will have to handle them carefully in analyzing the variational problem. So far we have not needed to use the strong assumptions about the domain  $\Omega$ . For the next result we need to invoke these standing assumptions.

**Theorem 4.3** Suppose  $\Omega$  is simply connected Lipschitz domain and has a boundary consisting of two disconnected components  $\Sigma$  and  $\Gamma$  (each of which are connected). In addition, suppose  $\mathbf{u} \in X$  is such that  $\mathbf{u}_\tau = 0$  on  $\Sigma$  and  $\nabla \times \mathbf{u} = 0$  in  $\Omega$ . Then there is a scalar potential  $p \in S$  such that  $\mathbf{u} = \nabla p$ , where  $S$  is defined by (4.6)

$$S = \left\{ p \in H^1(\Omega) \mid p = 0 \text{ on } \Sigma \text{ and } p \text{ is constant on } \Sigma \right\}.$$

**Remark 4.4** An equivalent way to state the theorem is that any  $\mathbf{u} \in H_0(\text{curl}; \Omega)$  such that  $\nabla \times \mathbf{u} = 0$  can be represented as the gradient of a scalar potential in  $S$ . Note  $\nabla S = \mathcal{K}_N(\Omega) \oplus \nabla H_0^1(\Omega)$  (see Theorem 3.41).

If the boundary of  $\Omega$  consists of  $N + 1$  connected components, then the above result holds except that the scalar potential is defined up to  $N$  constant boundary values on  $N$  of the surfaces. The proof is obvious.

If the domain  $\Omega$  is not simply connected, the same conclusion holds because of the assumed perfect conducting boundary condition. In this case we must first

reduce to a simply connected domain using “cuts”, and then show that in fact the potential  $p$  is continuous across the cuts. For a detailed discussion of cuts in this context, see [12] and Section 3.7.

**Proof of Theorem 4.3** By Theorem 3.37 we know that  $\mathbf{u} = \nabla p$  for some  $p \in H^1(\Omega)$ . The perfect conducting boundary condition implies that on each component of the boundary  $\mathbf{u}_T = (\mathbf{v} \times \mathbf{u}) \times \mathbf{v} = 0$ . But  $\mathbf{u}_T = (\nabla p)_T = \nabla_{sp}$ , where  $\nabla_{sp}$  is the surface gradient on  $S = \Gamma$  or  $S = \sum$ . Thus  $p$  is constant on each component of the boundary, and we can choose the value on one component (in particular, on  $\Gamma$ ) to vanish.  $\square$

## 4.4 Helmholtz decomposition

Now that we have characterized the null space of the curl operator we can remove, or factor out, this component from  $X$ . The following lemma, which we prove in detail, is essentially a special case of Theorem 3.45 (however, we allow  $\epsilon_r \neq 1$ ).

**Lemma 4.5** (Helmholtz decomposition) *The space  $\nabla S$  is a closed subspace of  $X$ , and we may write (4.7)*

$$X = X_0 \oplus \nabla S,$$

where (4.8)

$$X_0 = \{w \in X \mid (\epsilon_r w, \nabla \xi) = 0 \text{ for all } \xi \in S\}.$$

**Remark 4.6** This is just the Helmholtz decomposition. By the lemma, any  $\mathbf{u} \in X$  can be written uniquely as  $\mathbf{u} = \mathbf{u}_0 + \nabla p$  for some  $\mathbf{u}_0 \in X_0$  and  $p \in S$ . But from the definition of  $X_0$  we see that if  $\xi \in H_0^1(\Omega)$  then

$$(\nabla \cdot (\epsilon_r \mathbf{u}_0), \xi) = -(\epsilon_r \mathbf{u}_0, \nabla \xi) = 0,$$

so  $\nabla \cdot (\epsilon_r \mathbf{u}_0) = 0$ , and we can now conclude that the normal trace  $\mathbf{v} \cdot (\epsilon_r \mathbf{u}_0)$  is well defined. Using this fact, together with the choice of a test function  $\xi \in S$  such that  $\xi = 1$  on  $\sum$ , we see that  $\langle \mathbf{v} \cdot (\epsilon_r \mathbf{u}_0), 1 \rangle = 0$ . Since the divergence vanishes in  $\Omega$ ,

$$0 = (\nabla \cdot (\epsilon_r \mathbf{u}_0), 1) = \langle \mathbf{v} \cdot (\epsilon_r \mathbf{u}_0), 1 \rangle_{\partial\Omega}.$$

We conclude that  $\langle \mathbf{v} \cdot (\epsilon_r \mathbf{u}_0), 1 \rangle_{\Gamma} = 0$  and hence, by Theorem 3.38, there is a vector function  $\psi \in (H^1(\Omega))^3$  such that  $\epsilon_r \mathbf{u}_0 = \nabla \times \psi$ . Therefore, we can write

$$\mathbf{u} = \epsilon_r^{-1} \nabla \times \psi + \nabla p,$$

which reduces to the classical Helmholtz decomposition in Theorem 3.45 when  $\epsilon_r = 1$ .

**Proof of Lemma 4.5** This lemma is proved in [190]. It is entirely classical when  $\epsilon_r$  is real since then the bilinear form  $(\epsilon_r \mathbf{u}, \mathbf{v})$  is an inner product on  $(L^2(\Omega))^3$ , so the result follows from the projection theorem.

The space  $\nabla S$  is closed in  $X$  since  $S$  is closed in  $H^1(\Omega)$ . Now we have to make sense of the decomposition of  $X$ . Let us define the sesquilinear form  $\tilde{a} : X \times X \rightarrow \mathbb{C}$  by

$$\tilde{a}(u, v) = (\nabla \times u, \nabla \times v) + (\boldsymbol{\varepsilon}_r u, v) + \langle u_T, v_T \rangle, \text{ for } u, v \in X.$$

Since  $\boldsymbol{\varepsilon}_r$  is complex symmetric with strictly uniformly positive and bounded real part, the sesquilinear form  $\tilde{a}(\cdot, \cdot)$  has the following properties:

- (i) There exists a constant  $c > 0$  independent of  $\mathbf{u}$  such that

$$|\tilde{a}(u, u)| \geq c \|u\|_X^2 \text{ for all } u \in X.$$

This follows from taking the real part of  $\tilde{a}(u, u)$ .

- (ii) There exists  $C > 0$  independent of  $\mathbf{u}$  and  $v$  such that

$$|\tilde{a}(u, v)| \leq C \|u\|_X \|v\|_X \text{ for all } v, u \in X.$$

Hence, for each  $\mathbf{u} \in X$ , the Lax–Milgram Lemma 2.21 assures us that there exists a unique function  $P\mathbf{u} \in \nabla S$  such that

$$\tilde{a}(P\mathbf{u}, v) = (\boldsymbol{\varepsilon}_r u, v) \text{ for all } v \in \nabla S.$$

It follows that  $P$  is a bounded operator from  $X$  into  $\nabla S$  and obviously  $P\mathbf{u} = \mathbf{u}$  if  $\mathbf{u} \in \nabla S$ , so  $P$  is a projection. Thus we may write any function  $\mathbf{u} \in X$  as  $\mathbf{u} = P\mathbf{u} + (I - P)\mathbf{u}$ . But  $(I - P)\mathbf{u} \in X_0$  since for any  $\xi \in S$

$$(\boldsymbol{\varepsilon}_r(I - P)\mathbf{u}, \nabla \xi) = \tilde{a}((I - P)\mathbf{u}, \nabla \xi) = 0.$$

This completes the proof.  $\square$

#### 4.4.1 Compactness properties of $X_0$

Next we shall prove two useful properties of the space  $X_0$  defined in (4.8). The first, due to Weber [292], is the compact imbedding of  $X_0$  into  $(L^2(\Omega))^3$  (see also [207]). The second result, which follows from the compactness property, is a “Friedrichs inequality” [266, 196] showing that on  $X_0$ , the curl–curl bilinear form is coercive.

Using Theorem 3.47, and a trick from [71], we can prove the following general compactness result (of course, the Sobolev norm estimate in Theorem 3.47 does not hold since, in general, the fields in  $X_0$  are discontinuous). A similar trick is also used in [159].

**Theorem 4.7** Suppose the domain  $\Omega$  and the coefficient  $\boldsymbol{\varepsilon}_r$  satisfy the conditions given in Section 4.2. Then  $X_0$  is compactly imbedded in  $(L^2(\Omega))^3$ .

**Proof** The proof follows closely the proof of Proposition 2.28 in [71]. For this proof we need to distinguish between the space  $X_0$  when  $\varepsilon_r = 1$  and the corresponding space for general  $\varepsilon_r$ . Thus we define

$$\begin{aligned} X_0^{(1)} &= \{u \in X \mid (u, \nabla \xi) = 0 \text{ for all } \xi \in S\}, \\ X_0^{(\varepsilon_r)} &= \{u \in X \mid (\varepsilon_r u, \nabla \xi) = 0 \text{ for all } \xi \in S\}. \end{aligned}$$

Consider a bounded sequence  $\{w_n\}_{n=1}^\infty \subset X_0^{(\varepsilon_r)}$ . Using the Helmholtz decomposition (see Lemma 4.5) when  $\varepsilon_r = 1$  we may write  $w_n = w_n^{(1,0)} + \nabla p_n^{(1)}$ , for some  $w_n^{(1,0)} \in X_0^{(1)}$  and  $p_n^{(1)} \in S$ . Because,

$$(w_n, \nabla p_n^{(1)}) = (\nabla p_n^{(1)}, \nabla p_n^{(1)})$$

we know that  $\|\nabla p_n^{(1)}\|_v \leq \|w_n\|_X$  and thus  $\|w_n^{(1,0)}\|_X \leq C\|w_n\|_X$ . Then by the compactness property when  $\varepsilon_r = 1$  implied by Theorem 3.47 since  $X_0^{(1)} \subset W_N$  (see also Corollary 3.49), there is a function  $w^{(1)} \in X_0^{(1)}$  and a subsequence, still denoted  $\{w_n^{(1,0)}\}_{n=1}^\infty$  such that (4.9)

$$w_n^{(1,0)} \rightarrow w^{(1)} \text{ strongly in } (L^2(\Omega))^3 \text{ as } n \rightarrow \infty.$$

But using the Helmholtz decomposition of  $X_0^{(\varepsilon_r)}$  (i.e. with  $\varepsilon = \varepsilon_r$ ) we have  $w^{(1)} = w^{(\varepsilon_r)} + \nabla p^{(\varepsilon_r)}$ , for some  $w^{(\varepsilon_r)} \in X_0^{(\varepsilon_r)}$ , and  $p^{(\varepsilon_r)} \in S$ . We shall now show that  $w_n \rightarrow w^{(\varepsilon_r)}$  in  $(L^2(\Omega))^3$  as  $n \rightarrow \infty$ . Using the fact that  $\{w_n\}_{n=1}^\infty$  and  $w^{(\varepsilon_r)}$  are in  $X_0^{(\varepsilon_r)}$  we have

$$\begin{aligned} & \left( \varepsilon_r (w^{(\varepsilon_r)} - w_n), (w^{(\varepsilon_r)} - w_n) \right) \\ &= \left( \varepsilon_r (w^{(\varepsilon_r)} - w_n), (w^{(\varepsilon_r)} + \nabla p^{(\varepsilon_r)} - w_n + \nabla p_n^{(1)}) \right) \\ &= \left( \varepsilon_r (w^{(\varepsilon_r)} - w_n), (w^{(1)} - w_n^{(1,0)}) \right). \end{aligned}$$

Hence, due to (4.9),  $\|w^{(\varepsilon_r)} - w_n\|_{(L^2(\Omega))^3} \leq C\|w^{(1)} - w_n^{(1,0)}\|_{(L^2(\Omega))^3} \rightarrow 0$  as  $n \rightarrow 0$ .  $\square$

The next result verifies that we have indeed removed the null-space of the curl from  $X$ . It is very similar to Corollary 3.51, and is referred to as *Friedrichs inequality*.

**Corollary 4.8** Suppose that  $\Omega$  is a bounded, simply connected, Lipschitz domain with boundary consisting of two disjoint connected components  $\Sigma$  and  $\Gamma$ . In addition, suppose that the function  $\varepsilon_r$  satisfies the conditions given in Section 4.2. Then there is a constant  $C$  such that for every  $\mathbf{u} \in X_0$

$$\|\mathbf{u}\|_{(L^2(\Omega))^3} \leq C \left( \|\nabla \times \mathbf{u}\|_{(L^2(\Omega))^3} + \|\mathbf{v} \times \mathbf{u}\|_{(L^2(\Sigma))^3} \right).$$

**Proof** We proceed as in the proof of Corollary 3.51. This requires us to verify that if  $\mathbf{u} \in X_0$  and

$$\|\nabla \times \mathbf{u}\|_{(L^2(\Omega))^3} + \|\mathbf{v} \times \mathbf{u}\|_{(L^2(\Sigma))^3} = 0$$

then  $\mathbf{u} = 0$ . Since  $\nabla \times \mathbf{u} = 0$  in  $\Omega$ ,  $\mathbf{v} \times \mathbf{u} = 0$  on  $\Sigma$  and  $\mathbf{u} \in X_0$ , we conclude by Theorem 4.3 that  $\mathbf{u} = \nabla p$  for some  $p \in S$ . Thus  $\mathbf{u} \in (\nabla S) \cap (\nabla S)^\perp$  and so  $\mathbf{u} = 0$ .  $\square$

## 4.5 The variational problem as an operator equation

Now that we know  $X = X_0 \oplus \nabla S$  (see Lemma 4.5) we can write any solution  $\mathbf{E}$  of (4.4) as  $\mathbf{E} = \mathbf{E}_0 + \nabla p$  for some  $\mathbf{E}_0 \in X_0$  and  $p \in S$ . Substituting this decomposition into (4.4) and using the fact that  $\nabla \times \nabla p = 0$ , and  $(\nabla p) \times v = 0$  on  $\partial\Omega$ , we find that(4.10)

$$\begin{aligned} \left( \mu_r^{-1} \nabla \times E_0, \nabla \times \varphi \right) - \kappa^2 (\epsilon_r(E_0 + \nabla p), \varphi) - i\kappa \langle \lambda E_{0,T}, \varphi_T \rangle \\ = (F, \varphi) + \langle g, \varphi_T \rangle \text{ for all } \varphi \in X. \end{aligned}$$

Now if we choose  $\varphi = \nabla \xi$  for some  $\xi \in S$  and use the fact that  $\nabla \xi \in H_0(\text{curl}; \Omega)$  we obtain  $-\kappa^2(\epsilon_r(\mathbf{E}_0 + \nabla p), \nabla \xi) = (F, \nabla \xi)$ . But since  $\mathbf{E}_0 \in X_0$  we see that  $p \in S$  satisfies(4.11)

$$-\kappa^2(\epsilon_r \nabla p, \nabla \xi) = (F, \nabla \xi) \text{ for all } \xi \in S.$$

Given our assumptions on  $\epsilon_r$ , we can easily show that this variational problem has a unique solution.

**Lemma 4.9** *Assume that  $\kappa > 0$  and that  $\Omega$  and  $\epsilon_r$ satisfy the conditions given in Section 4.2 . Then there exists a unique solution  $p \in S$  to(4.11)and there is a constant  $C$  independent of  $F$ such that(4.12)*

$$\|\nabla p\|_{(L^2(\Omega))^3} \leq C \|F\|_{(L^2(\Omega))^3}.$$

**Proof** This is an easy application of the Lax–Milgram Lemma 2.21. Let  $\tilde{b}: S \times S \rightarrow \mathbb{C}$  be defined by  $\tilde{b}(p, \xi) = -\kappa^2(\epsilon_r p, \nabla \xi)$ . Then we can easily verify that  $\tilde{b}$  is a bounded sesquilinear form on  $S \times S$  using the boundedness of  $\epsilon_r$ . On the other hand, taking  $\xi = p$  and using the fact that the real part of  $\epsilon_r$  is uniformly positive, we can show that  $\tilde{b}$  is coercive.  $\square$

Using Lemma 4.9 we can see that determining  $\mathbf{E}$  or  $\mathbf{E}_0$  is equivalent and so we will study the problem of determining  $\mathbf{E}_0 \in X_0$  such that(4.13)

$$\begin{aligned} \left( \mu_r^{-1} \nabla \times E_0, \nabla \times \varphi \right) - \kappa^2 (\epsilon_r E_0, \varphi) - i\kappa \langle \lambda E_{0,T}, \varphi_T \rangle \\ = (F, \varphi) + \langle g, \varphi_T \rangle + \kappa^2 (\epsilon_r \nabla p, \varphi), \end{aligned}$$

for all  $\varphi \in X_0$ . The restriction to test functions in  $X_0$  is justified since  $X_0$  is a subset of  $X$ .

We shall have to work a good deal harder to analyze (4.13) compared to (4.11). We start by writing (4.13) as an operator equation and then show that the resulting equation is of Fredholm type. Application of the Fredholm alternative then reduces the analysis to proving that (4.13) has at most one solution. The obvious space in which to write the operator equation is  $X_0$ , but because our finite element space will not be a subspace of  $X_0$  (it will be a subspace of  $X$ ) we instead work in  $(L^2(\Omega))^3$ .

We define the sesquilinear form  $a_+: X \times X \rightarrow \mathbb{C}$  by(4.14)

$$a_+(u, v) = \left( \mu_r^{-1} \nabla \times u, \nabla \times v \right) + \kappa^2 (\epsilon_r u, v) - i\kappa \langle \lambda u, v_T \rangle$$

for all  $u, v \in X$ . This form is coercive as the following lemma shows.

**Lemma 4.10** *There exists a constant  $\alpha > 0$  depending on  $\mu_r, \varepsilon_r, \lambda$  and  $\kappa$  such that*

$$|a_+(u, u)| \geq \alpha \|u\|_X^2 \text{ for all } u \in X,$$

where  $\alpha$  is independent of  $u$ .

**Proof** This lemma is obvious if  $\varepsilon_r$  is real valued. From the definition of  $a_+$  we have

$$\begin{aligned} |a_+(u, u)|^2 &= \left( \left\| \mu_r^{-1/2} \nabla \times u \right\|_{L^2(\Omega)}^2 + \kappa^2 \left\| \Re(\varepsilon_r)^{1/2} u \right\|_{L^2(\Omega)}^2 \right)^2 \\ &\quad + \left( \kappa^2 \left\| \Im(\varepsilon_r)^{1/2} u \right\|_{L^2(\Omega)}^2 - \kappa \left\| \lambda^{1/2} u_T \right\|_{L^2(\Sigma)}^2 \right)^2. \end{aligned}$$

Expanding this expression and using the arithmetic-geometric mean inequality we see that, for any  $\delta > 0$ ,

$$\begin{aligned} |a_+(u, u)|^2 &\geq \left\| \mu_r^{-1/2} \nabla \times u \right\|_{L^2(\Omega)}^4 + \kappa^4 \left\| \Re(\varepsilon_r)^{1/2} u \right\|_{L^2(\Omega)}^4 \\ &\quad + \kappa^4 \left( 1 - \frac{1}{\delta} \right) \left\| \Im(\varepsilon_r)^{1/2} u \right\|_{L^2(\Omega)}^4 + \kappa^2 (1 - \delta) \left\| \lambda^{1/2} u_T \right\|_{L^2(\Sigma)}^2. \end{aligned}$$

By the assumptions on the coefficients, there are constants  $\varepsilon_{r,1} > 0$  and  $\varepsilon_{r,2} \geq 0$  such that  $\Re(\varepsilon_r) \geq \varepsilon_{r,1}$  and  $\Im(\varepsilon_r) \leq \varepsilon_{r,2}$  in  $\Omega$ . Then, if we choose  $\delta < 1$ , we may estimate

$$\begin{aligned} \left\| \Re(\varepsilon_r)^{1/2} u \right\|_{L^2(\Omega)}^4 &+ \left( 1 - \frac{1}{\delta} \right) \left\| \Im(\varepsilon_r)^{1/2} u \right\|_{L^2(\Omega)}^4 \\ &\geq \left( \varepsilon_{r,1}^2 + \varepsilon_{r,2}^2 - \frac{1}{\delta} \varepsilon_{r,2}^2 \right) \|u\|_{L^2(\Omega)}^4. \end{aligned}$$

Thus if we choose  $\delta$  such that  $1 > \delta > \varepsilon_{r,2}^2 / (\varepsilon_{r,1}^2 + \varepsilon_{r,2}^2)$  we obtain the desired inequality.  $\square$

Now that we know  $a_+$  is coercive, we can define the map  $K : (L^2(\Omega))^3 \rightarrow (L^2(\Omega))^3$  such that if  $f \in (L^2(\Omega))^3$  then  $Kf \in X_0 \subseteq (L^2(\Omega))^3$  satisfies (4.15)

$$a_+(Kf, \varphi) = -2\kappa^2 \int_{\Omega} \varepsilon_r f \cdot \varphi \, dv \quad \text{for all } \varphi \in X_0.$$

We have the following result.

**Theorem 4.11** *The operator  $K$  is a bounded and compact map from  $(L^2(\Omega))^3$  into  $(L^2(\Omega))^3$ . In addition,*

$$\|Kf\|_X \leq C \|f\|_{L^2(\Omega)}.$$

**Proof** We start by showing that  $K$  is well defined by checking that  $a_+(\mathbf{u}, \mathbf{v})$  satisfies the conditions of the Lax–Milgram Lemma 2.21. First we check boundedness. Using the Cauchy–Schwarz inequality and the boundedness of  $\mu_r$ ,  $\epsilon_r$ , and  $\lambda$ , we see that

$$\begin{aligned} |a_+(\mathbf{u}, \mathbf{v})| &\leq C \left( \|\nabla \times \mathbf{u}\|_{(L^2(\Omega))^3} \|\nabla \times \mathbf{v}\|_{(L^2(\Omega))^3} + \|\mathbf{u}\|_{(L^2(\Omega))^3} \|\mathbf{v}\|_{(L^2(\Omega))^3} \right. \\ &\quad \left. + \|\mathbf{u}_T\|_{(L^2(\Sigma))^3} \|\mathbf{v}_T\|_{(L^2(\Sigma))^3} \right), \end{aligned}$$

where  $C$  depends on  $\kappa$  and the upper and lower bounds on  $\mu_r$ ,  $\epsilon_r$  and  $\lambda$ . Hence  $|a_+(\mathbf{u}, \mathbf{v})| \leq C \|\mathbf{u}\|_X \|\mathbf{v}\|_X$ . Coercivity was proved in Lemma 4.10. Thus  $a_+(\cdot, \cdot)$  satisfies the conditions of the Lax–Milgram Lemma 2.21, so  $Kf$  is well defined and  $\|Kf\|_X \leq C \|f\|_{(L^2(\Omega))^3}$ .

Now we need to show that  $K$  is compact. Suppose  $(f_n)_{n=0}^\infty$  is a bounded sequence in  $(L^2(\Omega))^3$ . Then by the above inequality  $(Kf_n)_{n=0}^\infty$  is a bounded sequence in  $X_0$ . Hence by the Theorem 4.7 there is a subsequence converging strongly in  $(L^2(\Omega))^3$ . This implies that  $K$  is compact and we are done.  $\square$

Next we define a vector  $\mathcal{F} \in (L^2(\Omega))^3$  by requiring  $\mathcal{F} \in X_0$  satisfies(4.16)

$$a_+(\mathcal{F}, \varphi) = (F, \varphi) + \langle g, \varphi_T \rangle + \kappa^2 (\epsilon_r \nabla p, \varphi) \text{ for all } \varphi \in X_0.$$

Again using the Lax–Milgram Lemma 2.21 exactly as before, we have that  $\mathcal{F}$  is well defined and

$$\|\mathcal{F}\|_X \leq C \left( \|F\|_{(L^2(\Omega))^3} + \|g\|_{(L^2(\Sigma))^3} + \|\nabla p\|_{(L^2(\Omega))^3} \right).$$

By Lemma 4.9 this can be rewritten as(4.17)

$$\|\mathcal{F}\|_X \leq C \left( \|F\|_{(L^2(\Omega))^3} + \|g\|_{(L^2(\Sigma))^3} \right).$$

Using the operator  $K$  we can then see that problem (4.13) is equivalent to finding  $\mathbf{E}_0 \in (L^2(\Omega))^3$  such that(4.18)

$$(I + K)\mathbf{E}_0 = \mathcal{F}.$$

Furthermore, since  $K$  is compact, the Fredholm alternative (Theorem 2.33) is applicable. Thus, if we can show that this equation has at most one solution, we will have proved our desired existence result. The proof of the uniqueness of any solution of (4.18) requires a careful analysis and will be undertaken in the next section. Note that, if we can solve (4.18), then we shall have the norm estimate(4.19)

$$\|\mathbf{E}_0\|_{(L^2(\Omega))^3} \leq C \|\mathcal{F}\|_{(L^2(\Omega))^3}.$$

However, rearranging (4.18), we have  $\mathbf{E}_0 = \mathcal{F} - KE_0$ , so  $\mathbf{E}_0 \in X_0$  and, by the *a priori* estimate for  $K$ , we have

$$\|\mathbf{E}_0\|_X \leq C \left( \|\mathcal{F}\|_{(L^2(\Omega))^3} + \|\mathbf{E}_0\|_{(L^2(\Omega))^3} \right)$$

and, using the estimate (4.19) and the *a priori* estimate for  $\mathcal{F}$  in (4.17) provides an estimate for  $\mathbf{E}_0$  in  $X$ .

## 4.6 Uniqueness of the solution

In this section we shall prove that the variational formulation of Maxwell's equations (4.4) has at most one weak solution under the assumptions in Section 4.2 and provided that either (or both) of the following conditions hold:

- The imaginary part of  $\epsilon_r$  is strictly positive on some open subdomain of  $\Omega$ .
- The boundary  $\Sigma$  is non-empty.

Let us suppose there are two solutions  $\mathbf{E}_1, \mathbf{E}_2 \in X$  to (4.4). Let  $e = \mathbf{E}_1 - \mathbf{E}_2$ . Then using the linearity of the system we can subtract (4.4) with  $\mathbf{E} = \mathbf{E}_2$  from (4.4) with  $\mathbf{E} = \mathbf{E}_1$  to show that the difference  $e \in X$  satisfies the homogeneous problem(4.20)

$$\left( \mu_r^{-1} \nabla \times e, \nabla \times \varphi \right) - \kappa^2 (\epsilon_r e, \varphi) - i\kappa \langle \lambda e_T, \varphi_T \rangle = 0 \text{ for all } \varphi \in X.$$

Thus, to prove uniqueness of the solution, it suffices to prove that  $e = 0$  is the only solution of the above homogeneous problem. In fact, we shall prove the following theorem.

**Theorem 4.12** Under the assumption that the data for(4.4) satisfies the conditions in Section 4.2 and, in addition, that either the imaginary part of  $\epsilon_r$  is strictly positive on some non-empty subdomain of  $\Omega$  or that  $\Sigma$  is non-empty, there is at most one solution  $\mathbf{E}$  to (4.4).

The proof proceeds in two steps. First, using the above assumptions on the coefficient  $\epsilon_r$  or boundary  $\Sigma$ , we prove that the solution is unique on either the region where the imaginary part of  $\epsilon_r$  is positive, or on  $\Sigma$ . We then appeal to a unique continuation result to show that the solution is unique everywhere.

We prepare to prove Theorem 4.12 by first proving the unique continuation result. Unfortunately, this result does not seem to follow easily from the variational formulation for Maxwell's equations. Instead, it is proved by using a differential inequality. We state and prove the basic unique continuation result next. This is essentially Theorem 9.3 from Colton and Kress [94].

**Theorem 4.13** Suppose  $\Omega_0$  is an open, connected subdomain of  $\Omega$ . Suppose that  $u, v \in H(\text{curl}; \Omega_0)$  satisfy

$$\left. \begin{aligned} i\kappa \epsilon_r u + \nabla \times v &= 0 \\ i\kappa \mu_r u - \nabla \times v &= 0 \end{aligned} \right\} \text{ in } \Omega_0$$

and that  $u$  vanishes in a ball of non-zero radius contained in  $\Omega_0$ . Suppose, in addition, that  $\epsilon_r$  is a real, continuously differentiable function in  $\Omega_0$  and  $\mu_r$  is real and constant in  $\Omega_0$ . Then  $u = 0$  and  $v = 0$  in  $\Omega_0$ .

**Remark 4.14** This theorem holds in more generality. Vogelsang [287] proves unique continuation under the assumption that  $\epsilon_r$  and  $\mu_r$  are symmetric, real valued, uniformly positive-definite and bounded matrix functions of position in  $(L^\infty(\Omega_0))^\circ$ . However, the proof, although similar in spirit, is much more complicated than the one given here. An even more general case has recently been analyzed in [245].

The proof of this theorem uses a result on differential inequalities. The proof we give is from Colton and Kress [94] who in turn follow Leis [207] and Müller [228].

**Lemma 4.15** *Let  $\Omega_1$  be a connected domain in  $\mathbb{R}^3$  and suppose  $\mathbf{v} \in (H^1(\Omega_1))^3$  where  $v = (v_1, v_2, v_3)^\top$  is a real-valued function that satisfies*

$$|\Delta v| \leq C \sum_{q=1}^3 (|v_q| + |\nabla v_q|)$$

*almost everywhere in  $\Omega_1$ , where  $C$  is a constant. If  $v$  vanishes identically in a neighborhood of a point  $\mathbf{x}_0 \in \Omega_1$ , then  $v$  is identically zero in  $\Omega_1$ .*

**Remark 4.16** *This result is essentially Lemma 8.5 of [94], although our result is stated for  $\mathbf{v} \in (H^2(\Omega_1))^3$ . The proof in Colton and Kress is for  $\mathbf{v} \in (C^2(\Omega_1))^3$  but only relies on the fact that  $\Delta \mathbf{v}$  is well defined in  $L^2(\Omega_1)$ . For the proof of this lemma, see Colton and Kress [94].*

Using this differential inequality we can prove the unique continuation result for Maxwell's equations stated in Theorem 4.13.

**Proof of Theorem 4.13** Using the fact that  $\nabla \times v = i\kappa \epsilon_r \mathbf{u}$  we have

$$\Delta \times \epsilon_r^{-1} \nabla \times v - \kappa^2 \mu_r v = 0 \text{ in } \Omega_0.$$

In addition, since  $\mu_r$  is constant on  $\Omega_0$ , taking the divergence of this equation shows that  $\nabla \cdot v = 0$  in  $\Omega_0$ . But  $\nabla \times \epsilon_r^{-1} \nabla \times v = \epsilon_r^{-1} \nabla \times (\nabla \times v) + (\nabla \epsilon_r^{-1}) \times \nabla \times v$  and using the fact that  $\nabla \times \nabla \times v = -\Delta v + \nabla \nabla \cdot v$  we have

$$\Delta v = \epsilon_r (\nabla \epsilon_r^{-1}) \times \nabla \times v - \kappa^2 \mu_r v.$$

Hence  $\Delta v \in L^2(\Omega_0)$  and, by standard interior elliptic regularity results (see, e.g. Theorem 2.7 of [207]), we have that for any  $\Omega_1$  compactly contained in  $\Omega_0$ ,  $v \in (H^2(\Omega_1))^3$ . Thus we may apply the previous lemma to the real and imaginary parts of  $v$  separately and conclude  $v = 0$  in  $\Omega_1$ . But the subdomain  $\Omega_1$  was arbitrary and hence  $v = 0$  in  $\Omega_0$ .  $\square$

Now that we have prepared the ground, we can prove the desired uniqueness theorem.

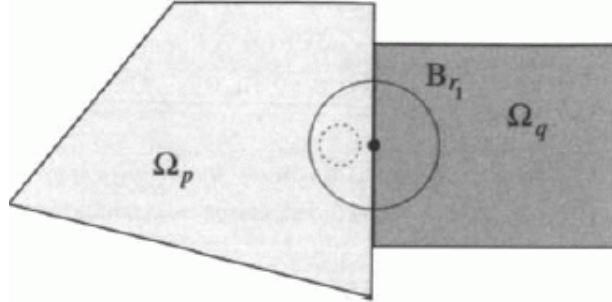
**Proof of Theorem 4.12** Consider the solution  $e \in X$  of (4.20). We now select  $\varphi = e$  in (4.20) and take the imaginary part of the resulting equation to show that (4.21)

$$\kappa^2 (\Im(\epsilon_r) e, e) + \kappa \langle \lambda e_T, e_T \rangle = 0.$$

Since  $\lambda$  is real and positive, this implies  $e_T = 0$  on  $\Sigma$  and  $e = 0$  on any subdomain of  $\Omega$  on which  $\Im(\epsilon_r)$  is strictly positive.

Suppose first that  $\Im(\epsilon_r)$  is strictly positive on some subdomain  $\Omega_p$  of  $\Omega$ . If  $\Omega_p = \Omega$  then we are done. If not, then we can only conclude that  $e = 0$  on all

Fig. 4.2. Geometry of the domains in the continuation proof of uniqueness. The ball  $B_{r_1}(x_0)$  is centered at  $x_0$  (indicated by the • in the diagram) on the common boundary between  $\Omega_p$  and  $\Omega_q$ . The solution  $e$  is zero on  $\Omega_p$  and hence in a small ball (dotted line) inside  $B_{r_1}(x_0)$ .



subdomains such that  $\Im(\varepsilon_r) \neq 0$ . Let  $\Omega_q$  be a subdomain on which  $\Im(\varepsilon_r) = 0$  and that shares a boundary containing an open set with a subdomain  $\Omega_p$  on which we have verified that  $e = 0$ . The pair of domains has the following properties:

- $\Omega_q \cap \Omega_p$  is a Lipschitz surface with non-empty interior;
- $\varepsilon_r$  is real on  $\Omega_q$ , and continuously differentiable there.

Now we extend  $\varepsilon_r$  on  $\Omega_q$  across the boundary  $\Omega_q \cap \Omega_p$  into  $\Omega_p$  such that the extended function  $\tilde{\varepsilon}_r$  is continuously differentiable (this is possible by the Calderon Extension Theorem 3.2 since our assumption is that  $\varepsilon_r \in H^3(\Omega_q)$ ). Also  $\mu_r$  restricted to  $\Omega_q$  is trivially extended as a constant to obtain  $\tilde{\mu}_r$  defined on  $\Omega_p \cup \Omega_q$ . Let  $B_{r_1}(x_0)$  be a ball of sufficiently small radius  $r_1$  centered on a point  $x_0$  in  $\Omega_q \cap \Omega_p$  such that  $B_{r_1}(x_0) \subset \Omega_q \cup \Omega_p$  and  $\tilde{\varepsilon}_r$  is strictly positive in  $\Omega_q \cup B_{r_1}(x_0)$  (see Fig. 4.2 ).

Since  $e$  vanishes on  $\Omega_p$  we have that  $e$  also satisfies

$$\nabla \times \tilde{\mu}_r \tilde{\nabla} \times e - \kappa^2 \tilde{\varepsilon}_r e = 0 \text{ in } \Omega_q \cup B_{r_1}(x_0)$$

and  $e$  vanishes on  $B_{r_1}(x_0) \cap \Omega_p$  and in particular on a ball in this set. Hence, since  $\tilde{\varepsilon}_r$  and  $\tilde{\mu}_r$  are real valued, we conclude by the unique continuation result in Theorem 4.13 that  $e = 0$  in  $\Omega_q \cup B_{r_1}(x_0)$  and hence in  $\Omega_p \cup \Omega_q$ . We can continue in this way jumping from subdomain to subdomain until all the subdomains on which  $\varepsilon_r$  is real have been reached and we conclude that  $e = 0$  on  $\Omega$ .

If  $\varepsilon_r$  is real in all  $\Omega$ , we must invoke the assumption on  $\Sigma$  which is not empty. Using (4.21) we know that  $e_r = 0$  on  $\Sigma$ . Let  $\Omega_q$  be a subdomain such that  $\Omega_q \cap \Sigma$  contains a non-trivial subdomain of  $\Sigma$  and such that  $\varepsilon_r$  is continuously differentiable on  $\Omega_q$ . As before, we can extend  $\varepsilon_r$  on  $\Omega_q$  to a function  $\tilde{\varepsilon}_r$  defined on  $\mathbb{R}^3$  such that  $\tilde{\varepsilon}_r$  is continuously differentiable. Since  $\varepsilon_r$  is positive on  $\Omega_q$ , we may then choose a ball  $B_{r_1}(x_0)$  centered at a point  $x_0$  on  $\Omega_q \cap \Sigma$  such that  $\tilde{\varepsilon}_r$  is positive on  $\Omega_q \cup B_{r_1}(x_0)$  and  $B_{r_1}(x_0) \cap \Omega \subset \Omega_q$ . Now if we extend  $e$  by zero from  $B_{r_1}(x_0) \setminus \Omega_q$  we have that

$$\int_{\Omega_q \cup B_{r_1}(x_0)} \tilde{\mu}_r^{-1} \nabla \times e \cdot \nabla \times \varphi - \kappa^2 \tilde{\epsilon}_r e \cdot \varphi dv = 0$$

for all  $\varphi \in H_0(\text{curl}; \Omega_q \cup B_{r_1}(x_0))$  so that  $e$  is a weak solution of Maxwell's equations there and, furthermore,  $e$  vanishes on  $B_{r_1}(x_0) \setminus \Omega_q$ . Hence, again by the unique continuation result of Theorem 4.13,  $e$  vanishes on  $\Omega_q \cup B_{r_1}(x_0)$  and hence on  $\Omega_p \cup \Omega_q$ . Now we can proceed as before to show that  $e = 0$  on  $\Omega$  by jumping across boundaries between the subdomains on which  $\epsilon_r$  is differentiable.  $\square$

We can now summarize our existence and uniqueness result for the interior problem (4.4).

**Theorem 4.17** Suppose the coefficients, domain and data for problem (4.4) satisfy all the conditions outlined in Section 4.2. Suppose, in addition, that either (or both) of the following conditions hold:

- The imaginary part of  $\epsilon$  is strictly positive on some subdomain of  $\Omega$  containing a ball of non-zero radius.
- The boundary  $\Sigma$  is non-empty.

Then problem (4.4) possesses a unique solution  $\mathbf{E} \in X$  for any value of  $\kappa > 0$ . Furthermore, there is a constant  $C > 0$  independent of  $\mathbf{E}$ ,  $\mathbf{F}$  and  $\mathbf{g}$  but depending on  $\kappa$  such that

$$\|\mathbf{E}\|_X \leq C \left( \|F\|_{(L^2(\Omega))^3} + \|g\|_{(L^2(\Sigma))^3} \right).$$

**Proof** According to Theorem 4.5 we may write  $\mathbf{E} = \mathbf{E}_0 + \nabla_p$  for some  $\mathbf{E}_0 \in X_0$ , and  $p \in S$ . In addition, by Lemma 4.9,  $p$  is uniquely determined and estimate (4.12) holds. By Theorem 4.12, there is at most one solution to (4.4) and hence, by the fact that (4.18) is a Fredholm equation, the Fredholm alternative (Theorem 2.33) implies that (4.18) has a unique solution and (4.19) holds.

Since  $\mathcal{F} \in X_0$  and  $A\mathbf{E}_0 \in X_0$ , we have that  $\mathbf{E}_0 = -K\mathbf{E}_0 + \mathcal{F} \in X_0$  and using (4.19)(4.22)

$$\|\mathbf{E}_0\|_X \leq \|K\mathbf{E}_0\|_X + \|\mathcal{F}\|_X \leq C \left( \|\mathbf{E}_0\|_{(L^2(\Omega))^3} + \|\mathcal{F}\|_X \right) \leq C \|\mathcal{F}\|_X.$$

The norm bound then follows by combining (4.12) and (4.22) and using (4.17).  $\square$

## 4.7 Cavity eigenvalues and resonances

Theorem 4.17 proves the existence and uniqueness of the solution of (4.4) provided  $\Im(\epsilon_r)$  is strictly positive on some region or  $\lambda$  is strictly positive on  $\Sigma \neq \emptyset$ . Suppose now that  $\Sigma = \emptyset$  and  $\Im(\epsilon_r) = 0$ . We are thus trying to solve the problem of finding  $\mathbf{E} \in H_0(\text{curl}; \Omega)$  such that (4.23)

$$\left( \mu_r^{-1} \nabla \times E, \nabla \times \varphi \right) - \kappa^2 (\epsilon_r E, \varphi) = (F, \varphi) \text{ for all } \varphi \in H_0(\text{curl}; \Omega).$$

The term  $\langle g, \varphi_r \rangle$  in (4.4) does not appear since  $\Sigma = \emptyset$ . As before,  $\mu_r$  and  $\epsilon_r$  satisfy the assumptions in Section 4.2 with the additional assumption that  $\Im(\epsilon_r) = 0$ .

In this case, as we shall see, there are values of  $\kappa$  for which we cannot, in general, conclude the existence of a unique solution to (4.23) (existence can be concluded for special choices of  $\mathbf{F}$ ). The values of  $\kappa$  for which (4.23) fails to have a unique solution are called *cavity eigenvalues* or *resonances* of  $\Omega$ . A knowledge of these resonances is useful in the design of microwave devices [271].

Since  $\Im(\mu_r) = \Im(\epsilon_r) = 0$ , we can take real and imaginary parts of  $\mathbf{E}$  and conclude that it suffices to analyze (4.23) when  $\mathbf{E}$ ,  $\mu_r$  and  $\epsilon_r$  are all real-valued functions and  $H_0(\text{curl}; \Omega)$  is a real Hilbert space. Thus for the rest of this section we shall be dealing with real-valued functions and all constants will be real.

To compute the resonances of  $\Omega$ , we consider the problem of finding non-trivial pairs  $\mathbf{E} \in H_0(\text{curl}; \Omega)$  and  $\kappa \in \mathbb{R}$  such that (4.24)

$$\left( \mu_r^{-1} \nabla \times \mathbf{E}, \nabla \times \varphi \right) = \kappa^2 (\epsilon_r \mathbf{E}, \varphi) \text{ for all } \varphi \in H_0(\text{curl}; \Omega).$$

Note that both  $\kappa$  and  $\mathbf{E}$  are unknown, as is usual for eigenvalue problems.

To analyze this problem, we can invoke the Helmholtz decomposition (Lemma 4.5) to write any solution as

$$\mathbf{E} = \mathbf{E}_0 + \nabla p, \text{ where } \mathbf{E}_0 \in X_0, p \in S.$$

Then we see that  $p$  satisfies (taking  $\varphi = \nabla \xi$  in (4.24) for some  $\xi \in S$ )

$$\kappa^2 (\epsilon_r \nabla p, \nabla \xi) = 0 \text{ for all } \xi \in S.$$

Thus either  $\kappa = 0$  or  $(\epsilon_r \nabla p, \nabla \xi) = 0$ . If  $\kappa \neq 0$  then choosing  $\xi = p$ , we have  $\nabla p = 0$  and hence from the vanishing Dirichlet data on  $\Gamma$  we have that  $p = 0$ . When  $\kappa = 0$ , we have from (4.24) that  $\mathbf{E}_0 \in X_0$  satisfies

$$(\epsilon_r \nabla \times \mathbf{E}_0, \nabla \times \varphi) = 0 \text{ for all } \varphi \in X_0.$$

Since  $\mathbf{E}_0 \in X_0$ , the Friedrichs inequality (Corollary 4.8) implies that  $\mathbf{E}_0 = 0$ .

Thus  $\kappa = 0$  is an eigenvalue of infinite multiplicity of (4.24) and the corresponding eigenfunctions are  $\mathbf{E}_0 = \nabla p$ , for  $p \in S$ . These eigenfunctions are not usually considered to be physically relevant since we also need  $\nabla \cdot (\epsilon_r \mathbf{E}) = 0$  because no sources are present. In this case, we see that  $\nabla \cdot (\epsilon_r \nabla p) = 0$  in  $\Omega$  and we again conclude that  $p = 0$ . Thus any scheme for using (4.24) to compute resonances must be able to identify the eigenfunctions corresponding to  $\kappa = 0$  and either only compute those for  $\kappa \neq 0$  or else compute all eigenpairs for (4.24) and reject those for  $\kappa = 0$ .

Now we can assume that  $\kappa \neq 0$  and we see that (4.24) may be written as the problem of finding  $\mathbf{E}_0 \in X_0$ , ( $\mathbf{E}_0 \neq 0$ ), and  $\kappa \in \mathbb{R}$ , such that (4.25)

$$\left( \mu_r^{-1} \nabla \times \mathbf{E}_0, \nabla \times \varphi \right) = \kappa^2 (\epsilon_r \mathbf{E}_0, \varphi) \text{ for all } \varphi \in X_0.$$

By choosing  $\varphi = \mathbf{E}_0$  we see from the Friedrichs inequality (Corollary 4.8) that

$$\kappa^2 = \frac{\left( \mu_r^{-1} \nabla \times \mathbf{E}_0, \nabla \times \mathbf{E}_0 \right)}{(\epsilon_r \mathbf{E}_0, \mathbf{E}_0)} > 0.$$

We may choose  $\kappa > 0$ .

To conclude the existence of eigenvalues and eigenfunctions we apply the Hilbert–Schmidt theory (Theorem 2.36). To this end, we rewrite (4.25) as the problem of finding  $\mathbf{E}_0 \in X_0$  such that

$$\left( \mu_r^{-1} \nabla \times \mathbf{E}_0, \nabla \times \varphi \right) + (\boldsymbol{\varepsilon}_r \mathbf{E}_0, \varphi) = (\kappa^2 - 1)(\boldsymbol{\varepsilon}_r \mathbf{E}_0, \varphi) \text{ for all } \varphi \in X_0.$$

Now we define the operator  $\mathcal{K}: (L^2(\Omega))^3 \rightarrow (L^2(\Omega))^3$  by requiring that if  $\mathbf{f} \in (L^2(\Omega))^3$ , then  $\mathcal{K}\mathbf{f} \in X_0$  satisfies (4.26)

$$\left( \mu_r^{-1} \nabla \times \tilde{\mathcal{K}}\mathbf{f}, \nabla \times \varphi \right) + (\boldsymbol{\varepsilon}_r \tilde{\mathcal{K}}\mathbf{f}, \varphi) = (\boldsymbol{\varepsilon}_r \mathbf{f}, \varphi) \text{ for all } \varphi \in X_0.$$

The proof of Theorem 4.11 shows that  $\mathcal{K}$  is compact. In addition,  $\mathcal{K}$  is self-adjoint, but not in the usual  $(L^2(\Omega))^3$  inner product. We actually need to introduce a new space  $L_{\boldsymbol{\varepsilon}_r}^2(\Omega)$  defined by  $L_{\boldsymbol{\varepsilon}_r}^2(\Omega) = (L^2(\Omega))^3$  but with the inner product (and corresponding norm) given by

$$(u, v)_{L_{\boldsymbol{\varepsilon}_r}^2(\Omega)} = (\boldsymbol{\varepsilon}_r u, v) \text{ for all } u, v \in L_{\boldsymbol{\varepsilon}_r}^2(\Omega).$$

Obviously, the norm on  $L_{\boldsymbol{\varepsilon}_r}^2(\Omega)$  is equivalent to the standard  $(L_2(\Omega))^3$  norm and so  $\mathcal{K}: L_{\boldsymbol{\varepsilon}_r}^2(\Omega) \rightarrow L_{\boldsymbol{\varepsilon}_r}^2(\Omega)$  is well defined and compact. To see that  $\mathcal{K}$  is self-adjoint in this space, for any  $u, v \in L_{\boldsymbol{\varepsilon}_r}^2(\Omega)$  we have, using the definition of  $\mathcal{K}$ ,

$$\begin{aligned} \left( \tilde{u}, \tilde{\mathcal{K}}v \right)_{L_{\boldsymbol{\varepsilon}_r}^2(\Omega)} &= \left( \tilde{\boldsymbol{\varepsilon}}_r u, \tilde{\mathcal{K}}v \right) = \left( \mu_r^{-1} \nabla \times \tilde{K}u, \nabla \times \tilde{\mathcal{K}}v \right) + \left( \boldsymbol{\varepsilon}_r \tilde{K}u, \tilde{\mathcal{K}}v \right) \\ &= \left( \mu_r^{-1} \nabla \times \tilde{K}v, \nabla \times \tilde{K}u \right) + \left( \boldsymbol{\varepsilon}_r \tilde{K}v, \tilde{K}u \right) \\ &= \left( \boldsymbol{\varepsilon}_r v, \tilde{K}u \right) = \left( \tilde{K}u, v \right)_{L_{\boldsymbol{\varepsilon}_r}^2(\Omega)}. \end{aligned}$$

Here we have used the fact that  $\boldsymbol{\varepsilon}_r$  and  $\mu_r$  are real, and hence  $(\cdot, \cdot)_{L_{\boldsymbol{\varepsilon}_r}^2(\Omega)}$  is a symmetric bilinear form.

Since  $\mathcal{K}$  is compact and self-adjoint as a map from  $L_{\boldsymbol{\varepsilon}_r}^2(\Omega)$  to itself, we can apply the Hilbert–Schmidt theory to the operator equation (4.27)

$$\mathcal{K}\mathbf{E}_0 = \mu\mathbf{E}_0,$$

where  $\mu = 1/(1 + \kappa^2)$ . We summarize our discussion (actually a little more work will be needed!) as follows:

**Theorem 4.18** *The solutions of the eigenvalue problem (4.24) have the following properties:*

- (1) *Corresponding to the eigenvalue  $\kappa = 0$  there is an infinite family of eigenfunctions  $\mathbf{E} = \nabla p$  for any  $p \in S$ .*
- (2) *There is an infinite discrete set of eigenvalues  $\kappa_j > 0, j = 1, 2, \dots$  and corresponding eigenfunctions  $\mathbf{E}_j \in X_0, \mathbf{E}_j \neq 0$ , such that*
  - (a) *equation (4.24) is satisfied,*
  - (b)  *$0 < \kappa_1 \leq \kappa_2 \leq \dots$ ,*

- (c)  $\lim_{j \rightarrow \infty} \kappa_j = \infty$ ,
- (d)  $\mathbf{E}_j$  is orthogonal to  $\mathbf{E}_l$  in the  $L^2_{\epsilon_r}(\Omega)$  inner product if  $j \neq l$ .

**Proof** In our discussion preceding the theorem we have already constructed the eigenfunctions corresponding to  $\kappa = 0$ . We have also verified that the Hilbert–Schmidt theory is applicable, so we know the existence of a possibly finite set of  $\gamma_j, j = 1, \dots$  and  $\mathbf{E}_j \neq 0$  such that

$$K \mathbf{E}_j = \gamma_j \mathbf{E}_j, \quad j = 1, 2, \dots.$$

The increasing property of  $\kappa_j = 1/\gamma_j - 1$  then follows from the fact that  $|\gamma_j|$  decreases.

We now need only verify that the set of eigenvalues is unbounded to complete the verification of the theorem. By the definition of  $K$  and recalling that  $N(K)$  denotes the null-space of  $K$ , we see that  $\mathbf{u} \in N(K)$  if and only if

$$\left( \mu_r^{-1} \nabla \times \tilde{K} \mathbf{u}, \nabla \times \varphi \right) + \left( \epsilon_r \tilde{K} \mathbf{u}, \varphi \right) = 0 \text{ for all } \varphi \in X_0,$$

and choosing  $\varphi = \mathbf{u}$  we see that this requires  $(\epsilon_r \mathbf{u}, \mathbf{u}) = 0$  so  $\mathbf{u} = 0$ . Hence  $N(K) = \{0\}$ , so by Theorem 2.36 we see that

$$X_0 = \text{closure}(\text{span}\{\mathbf{E}_1, \mathbf{E}_2, \dots\})$$

and, since  $X_0$  is infinite dimensional, so is  $\{\mathbf{E}_1, \mathbf{E}_2, \dots\}$ .  $\square$

Since the set of Maxwell eigenvalues is discrete, we can solve (4.23) for almost every value of  $\kappa > 0$ . When  $\kappa$  is not an eigenvalue we can apply the arguments of Section 4.5 and the Fredholm alternative to deduce the following result that is essentially a corollary to the previous theorem.

**Corollary 4.19** Suppose  $\Im(\epsilon_r) = 0$  and  $\sum = \emptyset$  and that the conditions on the data, domain and coefficients in Section 4.2 hold. Then, if  $\kappa$  is not a Maxwell eigenvalue (i.e.  $\kappa^2 \neq \kappa_j^2$  for any  $j$  where  $\kappa_j$  is the  $j$ th Maxwell eigenvalue guaranteed by Theorem 4.18, and  $\kappa > 0$ ), problem (4.23) has a unique solution for any  $\mathbf{F} \in (L^2(\Omega))^3$  and the norm estimate of Theorem 4.17 holds.

# 5 FINITE ELEMENTS ON TETRAHEDRA

## 5.1 Introduction

The previous chapter shows that the Sobolev space  $H(\text{curl}; \Omega)$  plays a central role in the variational theory of Maxwell's equations. Thus we need to derive finite elements in this space to obtain a class of finite element spaces suitable for discretizing the Maxwell system. The choice of this class of elements — termed edge elements — is motivated by the desire to design a robust finite element method for the Maxwell system. Edge elements can be used in the presence of geometric complexity (and its inevitable consequences on the regularity of the solution of Maxwell's equations) and in the presence of discontinuous electromagnetic properties which occur when electromagnetic waves propagate through different materials. Besides this practical justification, edge elements possess many fascinating mathematical properties and challenges.

The chapter starts (see Section 5.2) by discussing some general aspects of finite element methods and a characterization of the continuity requirements of the various spaces. Next, in Section 5.3, we describe the first step in a practical finite element computation: mesh generation. The design of efficient and reliable mesh generators is a research area outside the scope of this book (see, e.g. [278, 140, 134]). We shall simply discuss some basic requirements of a suitable mesh. In reality, mesh generation must be linked, either by the experience of the user or, better still, by automatic adaptive software, to the desired solution. Nevertheless, it is convenient to discuss the mesh first in isolation.

After we have described the mesh, we then discuss finite elements built on this mesh, assuming the mesh to consist of tetrahedra (elements on hexahedral grids are discussed in the next chapter). From the previous chapter we know that it would be best to use finite elements that lie in the space  $X_0$  (see (4.8)). Unfortunately, as we shall see, there is currently no good finite element subspace of this space. This is because any finite element subspace must consist of continuous piecewise polynomials (see the remark following Lemma 5.3) and hence be a subspace of  $(H^1(\Omega))^3$ . Use of standard continuous piecewise linear finite elements, which are in  $(H^1(\Omega))^3$ , cannot generally be successful without special modification due to the fact that  $(H^1(\Omega))^3 \cap X_0$  is a proper subset of  $X_0$  when  $\Omega$  has re-entrant corners (see Lemma 3.56). Hence the resulting finite element space is not dense as the mesh size goes to zero [19, 20, 104, 113]. We are thus driven to use subspaces of  $X$  that are not subspaces of  $X_0$ . This non-conformity results in complications for the analysis and implementation of finite element methods for Maxwell's equations.

It turns out to be necessary to present four different finite element spaces! The most obvious requirement is for finite elements suitable for discretizing the basic energy space for electromagnetics:  $X$  or more generally  $H(\text{curl}; \Omega)$ . This leads us to the edge elements of Nédélec [233]. The analysis presented in the previous chapter also requires the use of a scalar potential in the space  $S \subset H^1(\Omega)$ , so we also need to discuss standard spaces of continuous scalar finite elements. To analyze Nédélec's elements in  $H(\text{curl}; \Omega)$ , we shall also need to present his family of elements suitable for discretizing  $H(\text{div}; \Omega)$ , and for completeness we shall present a related finite element family in  $L^2(\Omega)$ . Note that the relevant function spaces are related by the famous de Rham diagram [54] discussed in Section 3.7:

$$H^1(\Omega) \xrightarrow{\nabla} H(\text{curl}; \Omega) \xrightarrow{\nabla \times} H(\text{div}; \Omega) \xrightarrow{\nabla \cdot} L_2(\Omega),$$

which summarizes, for example, the fact that if  $p \in H^1(\Omega)$  then  $\nabla p \in H(\text{curl}; \Omega)$ . Furthermore, the range of the gradient operator is closed and contained in the kernel of the curl operator, or more simply  $\nabla \times (\nabla p) = 0$  (similarly,  $\nabla \cdot (\nabla \times A) = 0$ ). We shall construct finite element spaces  $U_h \subset H^1(\Omega)$ ,  $V_h \subset H(\text{curl}; \Omega)$ ,  $W_h \subset H(\text{div}; \Omega)$  and  $Z_h \subset L^2(\Omega)$  (and from these suitable spaces for discretizing the Maxwell system) which have the same relationship as the continuous spaces. We shall also describe interpolation operators  $\pi_h$ ,  $r_h$ ,  $w_h$  and  $P_{0,h}$  that map from suitable subspaces  $U \subset H^1(\Omega)$ ,  $V \subset H(\text{curl}; \Omega)$ ,  $W \subset H(\text{div}; \Omega)$  and the space  $L^2(\Omega)$  into the appropriate finite element spaces (these operators are used in finite element error analysis and also in implementing boundary conditions). Of central importance to the analysis is that the spaces and interpolation operators are linked by the following commuting diagram called the discrete de Rham diagram:

$$\begin{array}{ccccccc} H^1(\Omega) & \xrightarrow{\nabla} & H(\text{curl}; \Omega) & \xrightarrow{\nabla \times} & H(\text{div}; \Omega) & \xrightarrow{\nabla \cdot} & L^2(\Omega) \\ \cup & & \cup & & \cup & & \\ U & & V & & W & & \\ \pi_h \downarrow & & r_h \downarrow & & w_h \downarrow & & P_{0,h} \downarrow \\ U_h & \xrightarrow{\nabla} & V_h & \xrightarrow{\nabla \times} & W_h & \xrightarrow{\nabla \cdot} & Z_h. \end{array}$$

This diagram implies, for example, that if  $p$  is smooth enough (i.e. in  $U$ ) then

$$\nabla \pi_h p = r_h \nabla p.$$

The interrelationship of the various spaces will be key to our error analysis. It has practical significance in the way that charge conservation is approximated, and in the way that edge finite element methods can be stabilized. One implication of the commuting diagram is the existence of an approximate Helmholtz decomposition of a vector field into the gradient of a scalar potential and an

almost divergence-free vector. This will be made more explicit in Sections 7.2.1 and 7.3.

Although the interrelationship of operators and spaces was implicit in the papers of Nédélec [233, 235], Girault [142] was the first to detail the actual operator  $\pi_b$  involved. The relationship to the de Rham diagram and the central importance of this structure was first noted by Bossavit [53, 54]. Building on this idea, Hiptmair [164] has presented a theory of finite elements for Maxwell's equations from the point of view of differential forms. There is no doubt that this is extremely elegant, and explains some of the rather obscure choices we shall make during the rest of this chapter.<sup>1</sup>

The analysis in this and subsequent chapters is based on the idea that a finite element space is used on a sequence of meshes. Accuracy is obtained by taking a sufficiently fine and well-designed mesh. This is the classical  $h$ -version of the finite element method. Later, in Chapter 8, we will make some observations regarding the  $hp$ -version in which the mesh is refined and the finite element space is also modified to accelerate convergence (see, e.g. [275, 286]).

In Section 5.4 we describe and analyze the finite element spaces of Nédélec in  $H(\text{div}; \Omega)$  [233, 235] which will be used to discretize the magnetic induction. A corresponding analysis for elements in  $H(\text{curl}; \Omega)$  is performed in Section 5.5. These elements will be used to discretize the electric field. We complete our analysis of tetrahedral elements by constructing subspaces of scalar functions in  $H^1(\Omega)$  (Section 5.6) and in  $L^2(\Omega)$  (Section 5.7). Finally, we also need to comment on the trace of finite element functions on the boundary, and we do this in Section 5.8.

Rather than provide every detail of the more complex estimates in this section, full proofs are not provided for all results. For example, some of the approximation results for non-integer-order Sobolev space are not proved, despite the importance of these estimates in later chapters. Instead, the main ideas of such proofs are provided by proving the result for integer-order spaces, and the reader is directed to the appropriate references for the remaining cases. The aim is to make the presentation more readable without sacrificing too much understanding.

## 5.2 Introduction to finite elements

Finite elements are built using piecewise polynomial functions on simple geometrical domains (e.g. piecewise linear functions on tetrahedra). The presentation of the finite element spaces and the interpolation error analysis in this chapter follows the classical approach of Ciarlet [80]. In this classical approach, a finite element is a triple  $(K, P_K, \Sigma_K)$ , where

- $K$  is a geometric domain (e.g. a tetrahedron, hexahedron or prism),

<sup>1</sup> I have elected not to follow this theory because to do so would restrict the potential readership. However, I do think that Hiptmair's approach will ultimately be the "standard" way to work with finite elements for Maxwell's equations and is a very powerful tool in the right hands.

- $P_K$  is a space of functions (usually polynomials) on  $K$ , and
- $\Sigma_K$  is a set of linear functionals on  $P_K$ . These linear functionals are called the *degrees of freedom* of the finite element.

For example, in the rather trivial case of quadratic finite elements in one dimension, we could define(5.1a)

$$K = (a, b), \quad (5.1b)$$

$$P_K = \text{polynomials in one variable of degree at most two}, \quad (5.1c)$$

$$\begin{aligned} \Sigma_K = & \{ l_i, 1 \leq i \leq 3 \mid l_1(u) = u(a), l_2(u) = ((a+b)/2), \\ & l_3(u) = u(b) \} . \end{aligned}$$

For the purpose of analysis and computation it is necessary to chose a particular *reference element*. This element is usually chosen to be a simple shape and unit size and in this book the reference element will always be denoted  $K$ . In one dimension,  $K = (0, 1)$ . We can obtain a finite element on a general geometric element simply by mapping from the reference element to a general element. Let us again consider quadratic finite elements in one dimension. If  $\hat{p} \in P_K$  then we can obtain a corresponding quadratic polynomial on the general element by  $p(a+(b-a)\hat{x}) = \hat{p}(\hat{x})$  for each  $\hat{x} \in K$ . In this case the degrees of freedom on  $K$  are given by the set  $\Sigma_K$  in (5.1c) with  $a = 0$  and  $b = 1$ , which we write in shorthand as  $\{\hat{u}(0), \hat{u}(1/2), \hat{u}(1)\}$ . Using the change of variables, they map directly to the degrees of freedom on  $K$  given by the set  $\{u(a), u((a+b)/2), u(b)\}$ . Note that we shall often write the degrees of freedom in this shorthand form, where each element of the set defines a functional. We shall usually define operations (e.g. numerical quadrature) on the reference element, and then obtain general results by mapping to a given element. This is also how we prefer to program finite elements.

Of course, the three components of the finite element  $(K, P_K, \Sigma_K)$  cannot be chosen at random. The geometric element must be chosen so that  $K$  is nondegenerate (in our simple example,  $b > a$ ), and  $P_K$  must be a finite-dimensional vector space of functions that are convenient to implement (e.g. polynomials). The degrees of freedom  $\Sigma_K$  must be chosen so that if a value is given for each of the degrees of freedom, it uniquely determines a function in  $P_K$ . In this case the finite element is said to be *unisolvant*.

**Definition 5.1** The finite element  $(K, P_K, \Sigma_K)$  is said to be *unisolvant* if specifying a value for each the degrees of freedom in  $\Sigma_K$  uniquely determines a function in  $P_K$ .

In the case of our simple example (5.1), we know that specifying the value of a quadratic polynomial at three distinct points uniquely determines the polynomial, and thus (5.1) is unisolvant. Once the unisolvence of a finite element has been established, we can use the degrees of freedom to construct a basis

for  $P_K$ . The basis functions are often referred to as *shape functions* in the engineering literature. If we have a general finite element with degrees of freedom  $\sum_K = \{l_n, 1 \leq n \leq m\}$  for some  $m \geq 1$  then unisolvence requires that  $P_K$  have dimension  $m$ . We can then define the basis consistent with  $\sum_K$  to be  $\{\phi_j\}_{j=1}^m$ , where, for  $1 \leq n \leq m$ , (5.2)

$$\phi_j \in P_K \text{ and } l_n(\phi_j) = \delta_{n,j} = \begin{cases} 1 & \text{if } n = j, \\ 0 & \text{otherwise.} \end{cases}$$

The fact that the finite element is unisolvent implies that the set  $\{\phi_j\}_{j=1}^m$  is well-defined and is a basis for  $P_K$ . Any function  $p \in P_K$  can be written as

$$p(x) = \sum_{j=1}^m l_j(p) \phi_j(x)$$

so that the degrees of freedom give the expansion coefficients for writing a general finite element function in terms of a convenient basis.

Applying this general philosophy to (5.1), we can compute the standard Lagrange basis for quadratic polynomials. From the definition of  $P_K$  in (5.1) we know that on  $K$  (when  $a = 0$  and  $b = 1$ )

$$\widehat{\phi}_j(\hat{x}) = a_j + b_j \hat{x} + c_j \hat{x}^2, \quad 1 \leq j \leq 3,$$

and by the definition of  $\sum_K$  in (5.1) we have that

$$\begin{aligned} l_1(\widehat{\phi}_j) &= a_j, \\ l_2(\widehat{\phi}_j) &= a_j + b_j / 2 + c_j / 4, \\ l_3(\widehat{\phi}_j) &= a_j + b_j + c_j. \end{aligned}$$

Thus, to construct  $\Phi_1$  we must solve the matrix problem

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{4} \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ b_1 \\ c_1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

Unisolvence is equivalent to the unique invertability of this linear system for an arbitrary right-hand side (and, in particular, the one given here). A simple calculation shows that  $a_1 = 1$ ,  $b_1 = -3$  and  $c_1 = 2$ , so that we get the expected basis function

$$\widehat{\phi}_1 = 2(\hat{x} - 1/2)(\hat{x} - 1).$$

A similar calculation, using the right-hand side  $(0, 1, 0)^T$  yields  $\widehat{\phi}_2(x) = 4(1 - \hat{x}^2)/3$  and, with  $(0, 0, 1)^T$ , we obtain  $\widehat{\phi}_3(x) = 2\hat{x}(\hat{x} - 1/2)$ . Of course, all we have done is compute the standard Lagrange basis for quadratic polynomials on the unit interval.

Note that the degrees of freedom are not unique. We might, for example, use the following alternative finite element, which differs from (5.1) only by the degrees of freedom:(5.3a)

$$K = (a, b), \quad (5.3b)$$

$$P_K = \text{polynomials of degree at most two}, \quad (5.3c)$$

$$\Sigma_K = \left\{ l_i, 1 \leq i \leq 3 \middle| \begin{array}{l} l_1(u) = u(a), \quad l_2(u) = \int_a^b u(s) ds, \\ l_3(u) = u(b) \end{array} \right\}.$$

A quick calculation shows that this finite element is also unisolvant and we can compute a basis corresponding to these degrees of freedom as before. In fact, the basis is exactly the same as for the element given by (5.1)! This does not always happen, but occurs here because Simpson's quadrature rule is exact for quadratic polynomials and the degrees of freedom for (5.1) are at the quadrature points of Simpson's rule.

Associated with the finite element  $(K, P_K, \Sigma_K)$  is an interpolant. For a suitably smooth function  $u$  (in the case of (5.1), a continuous function), we define the interpolant on the interval  $K$  to be the unique function  $\pi_K u \in P_K$  such that

$$l(\pi_K u - u) = 0 \text{ for all } l \in \Sigma_K.$$

The operator  $\pi_K : C(K) \rightarrow P_K$  is referred to as the *interpolation operator*. Different degrees of freedom give rise to different interpolation operators.

Using the definition of the nodal basis we see that

$$\pi_K u(x) = \sum_{j=1}^m l_j(u) \varphi_j(x).$$

In the case of (5.1) on  $K$ , we have just used an elaborate method to arrive at the standard quadratic Lagrange interpolant since on the reference element

$$\pi_{\hat{K}} \hat{u}(\hat{x}) = \hat{u}(0)2(\hat{x}-1)(\hat{x}-1/2) + \hat{u}(1/2)4(1-\hat{x}^2)/3 + \hat{u}(1)2\hat{x}(\hat{x}-1/2).$$

The interpolant for (5.3) is different. Now (still representing the interpolation operator by  $\pi_K$ ) we have

$$\begin{aligned} \pi_{\hat{K}} \hat{u}(\hat{x}) = & \hat{u}(0)2(\hat{x}-1)(\hat{x}-1/2) + \left( \int_0^1 \hat{u}(\hat{s}) d\hat{s} \right) 4(1-\hat{x}^2)/3 \\ & + \hat{u}(1)2\hat{x}(\hat{x}-1/2). \end{aligned}$$

Generally, interpolants involving integral degree of freedom have better approximation properties than those involving point values alone, as we shall see shortly. This is important from the theoretical point of view since the error in a suitable

interpolant can often be used to estimate the error in the solution obtained by the finite element method. The interpolant is also useful in practice, for example, to set boundary conditions.

The next step is to use the element-wise defined finite elements to build a space of functions on the desired domain. Suppose we want to use the finite element (5.1) to approximate functions defined on  $[0, L]$ . We first “mesh” the interval by decomposing  $[0, L]$  into a finite union of non-overlapping intervals so that  $[0, L] = \cup_{i=1}^N [a_i, b_i]$  where  $a_i < b_i$ ,  $a_i = b_{i-1}$ ,  $2 \leq i \leq m$ ,  $a_1 = 0$  and  $b_N = L$ . By applying (5.1) on each interval  $(a_i, b_i)$ , we obtain a function that is piecewise quadratic on  $[0, L]$ . However, the finite element space is not all piecewise quadratics. The properties of the global finite element are governed by the degrees of freedom. By taking the union of all the element degrees of freedom, we obtain a set of global degrees of freedom  $\Sigma$  given by  $\Sigma = \cup_{K \in \tau_b} \Sigma_K$ . By specifying values for all the degrees of freedom in  $\Sigma$ , we in turn specify the degrees of freedom on each element; thus on each element  $K$  we specify a unique function in  $P_K$  (since  $\Sigma_K$  is unisolvant). In computing a global finite element function (e.g. when we approximate a partial differential equation using the finite element method), we must compute values for these degrees of freedom.

In the case of (5.1) or (5.3), the value of the finite element function at the end point of each interval is in  $\Sigma$  and so the global finite element function is continuous (the piecewise quadratic is continuous in each subinterval and the fact that the quadratics agree across inter-element boundaries means that the global function is continuous; see Fig. 5.1). Thus the elements (5.1) or (5.3) give rise to the finite element space consisting of all continuous piecewise quadratics on the given mesh. We define the maximum element size by  $h = \max_{1 \leq i \leq N} |b_i - a_i|$ . Then we can write the finite element space as

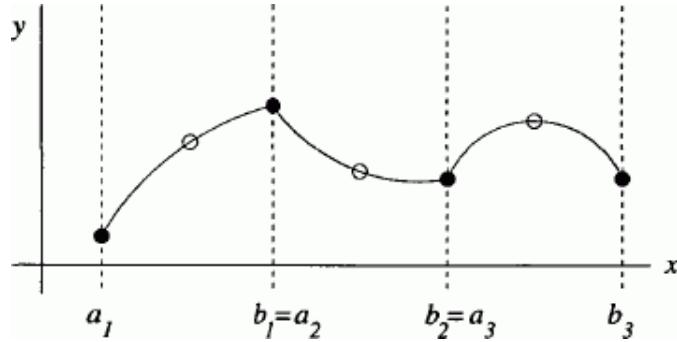
$$S_h = \{u_h \in C(0, L) | u_h|_K \in P_K \text{ for every element } K \text{ in the mesh}\} .$$

This definition does not mention the degrees of freedom explicitly, and shows that the global finite element function space corresponding to (5.1) or (5.3) is the same. This means that if  $S_h$  is used in a numerical algorithm, the accuracy properties will be independent of the interpolant used (provided the interpolant is not used explicitly in the algorithm). However, it is often the case that one set of degrees of freedom is computationally preferable compared to another, since the choice of degrees of freedom effects the conditioning of the matrices occurring in the finite element method [14].

From our definition of  $S_h$  it is obvious that  $S_h \subseteq C(0, 1)$ , and this implies (by virtue of Theorem 5.3 below) that  $S_h \subseteq H^1(0, 1)$ . The key here is that the degrees of freedom line up to guarantee the needed global smoothness. We say that the degrees of freedom  $\Sigma$  are  $H^1$  conforming or more generally:

**Definition 5.2** Let  $W$  be a space of functions. The finite element  $(K, P_K, \Sigma_K)$  is said to be  $W$  *conforming* if the corresponding global finite element space is a subspace of  $W$ .

Fig. 5.1. The degrees of freedom govern global continuity. Between the vertical dashed lines (marking the nodes of the mesh) the function is quadratic on each subinterval. Since the degree of freedom at  $b_1$  is also at  $a_2$ , the corresponding function values must agree, so the piecewise quadratic is continuous there and at other mesh points marked  $\bullet$ . The degrees of freedom marked  $\circ$  are at the mid-point of each subinterval (not at nodes in the mesh) and the polynomial is always continuous within each subinterval.



At this stage we can go ahead and analyze the approximation properties of  $S_h$  by estimating the error in the interpolant. By using the interpolant  $\pi_h u$  on each element  $K$ , we can build a global interpolant  $\pi_h u \in S_h$  of a suitably smooth function  $u$  (in this case, continuous is sufficient). It is possible to prove that for the interpolant given by (5.1) there is a constant  $C$  independent of  $h$  and  $u$  such that(5.4)

$$\|u - \pi_h u\|_{H^s(0, L)} \leq Ch^{t-s}|u|_{H^t(0, L)} \text{ for } 0 \leq s \leq 1 \leq t \leq 3.$$

This result is typical of the estimates we shall derive later in this chapter. Assuming  $u \in H(0, L)$ ,  $t \geq 1$ , this inequality gives the rate at which the interpolation error decreases as the mesh is refined (in other words, as  $h$  is decreased). The exponent of  $h$  is optimal (no higher rate of convergence is possible in the global norms used here). Unfortunately, the constant  $C$  is usually difficult to estimate with any precision. Note that, if the interpolant (5.3) is used instead of (5.1), we obtain the same estimate as above but in addition if  $u \in H_0^1(0, L) \cap H^t(0, L)$  we have

$$\|u - \pi_h u\|_{H^{-1}(0, L)} \leq Ch^{1+t}\|u\|_{H^t(0, L)} \text{ for } 1 \leq t \leq 3,$$

so that the maximum rate of convergence is thus  $O(h^4)$  in the  $H^{-1}(0, L)$  norm provided  $u \in H^3(0, L) \cap H_0^1(0, L)$ .

This simple example has been used to show how locally defined finite elements can be used to build global finite element spaces. When designing a finite element, we need to specify a polynomial space and degrees of freedom that guarantee conformance in a suitable function space. For example, for a subspace of  $H^1(\Omega)$ , it is necessary that the global finite element function be continuous. The only lemma of this section details, in addition, the continuity requirements for subspaces of  $H(\text{curl}; \Omega)$  and  $H(\text{div}; \Omega)$  (see [80, 233]).

**Lemma 5.3** Suppose  $K_1$  and  $K_2$  are two non-overlapping Lipschitz domains meeting at a common surface  $\Sigma$  (with non-zero measure) so that  $K_1 \cap K_2 = \Sigma$ .

(1) Suppose that  $p_1 \in H^1(K_1)$  and  $p_2 \in H^1(K_2)$  and define  $p \in L^2(K_1 \cup K_2 \cup \Sigma)$  by

$$p = \begin{cases} p_1 \text{ on } K_1, \\ p_2 \text{ on } K_2. \end{cases}$$

Then, if  $p_1 = p_2$  on  $\Sigma$ , we have  $p \in H^1(K_1 \cup K_2 \cup \Sigma)$ .

(2) Suppose that  $u_1 \in H(\text{curl}, K_1)$  and  $u_2 \in H(\text{curl}, K_2)$  and define  $u \in (L^2(K_1 \cup K_2 \cup \Sigma))^3$  by (5.5)

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

Then, if  $u_1 \times \nu = u_2 \times \nu$  on  $\Sigma$ , where  $\nu$  is a unit normal to  $\Sigma$ , we have  $u \in H(\text{curl}, K_1 \cup K_2 \cup \Sigma)$ .

(3) Suppose that  $u_1 \in H(\text{div}, K_1)$  and  $u_2 \in H(\text{div}, K_2)$ . Define  $u \in (L^2(K_1 \cup K_2 \cup \Sigma))^3$  by (5.5). Then, if  $u_1 \cdot \nu = u_2 \cdot \nu$  on  $\Sigma$ , where  $\nu$  is a unit normal to  $\Sigma$ , we have  $u \in H(\text{div}, K_1 \cup K_2 \cup \Sigma)$ .

**Remark 5.4** If we wish to use a finite element space in  $X_0$  we would need to impose the continuity across  $\Sigma$  of parts 2. and 3. above. Hence both  $u \cdot \nu$  and  $u \times \nu$  would need to be continuous there. For piecewise polynomials on  $K_1$  and  $K_2$ , this would imply the continuity of  $u$  across  $\Sigma$  and hence the resulting finite element space would be a subspace of  $(H(K_1 \cup K_2 \cup \Sigma))^3$ .

**Proof of Lemma 5.3** The first result is standard for finite element methods (see Theorem 2.1.1 of [80]). We shall only provide a proof of the second part of this lemma (from [235]), since the last part is proved similarly. We shall prove that  $\nabla \times u \in (L^2(K_1 \cup \Sigma \cup K_2))^3$  and

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

Let  $\phi \in (C_0^\infty(K_1 \cup K_2 \cup \Sigma))^3$ . Then, using the definition of  $u$  followed by integration by parts (in fact, (3.51)),

$$\begin{aligned} \int_{K_1 \cup K_2 \cup \Sigma} u \cdot \nabla \times \phi dV &= \int_{K_1} u_1 \cdot \nabla \times \phi dV + \int_{K_2} u_2 \cdot \nabla \times \phi dV \\ &= \int_{K_1} \nabla \times u_1 \cdot \phi dV + \int_{K_2} \nabla \times u_2 \cdot \phi dV \\ &\quad + \sum \int_{\Sigma} (u_1 \times \nu_1 + u_2 \times \nu_2) \cdot \phi dA, \end{aligned}$$

where  $\nu_1$  is the unit outward normal to  $K_1$  and  $\nu_2$  is the unit outward normal to  $K_2$ . Using the distributional definition of the curl, the relationship  $\nu_1 = -\nu_2$  on  $\Sigma$  and the assumed continuity requirement that  $u_1 \times \nu_1 = u_2 \times \nu_1$ , we have

$$\begin{aligned} \int_{K_1 \cup K_2 \cup \sum} \nabla \times u \cdot \phi dV &= \int_{K_1 \cup K_2 \cup \sum} u \cdot \nabla \times \phi dV \\ &= \int_{K_1} \nabla \times u_1 \cdot \phi dV + \int_{K_2} \nabla \times u_2 \cdot \phi dV, \end{aligned}$$

which completes the proof.  $\square$

### 5.2.1 Sets of polynomials

Finite element spaces are built using piecewise polynomial functions. In this section we define some notation and summarize some results for polynomial spaces used throughout the book. Let(5.6)

$$P_k = \{\text{polynomials of maximum total degree } k \text{ in } x_1, x_2, x_3\}, \quad (5.7)$$

$$\tilde{P}_k = \{\text{homogeneous polynomials of total degree exactly } k \text{ in } x_1, x_2, x_3\}$$

More precisely, let the multi-index  $\alpha = (\alpha_1 + \alpha_2 + \alpha_3)^T \in \mathbb{Z}_+^3$  (so that  $\alpha_i \geq 0$ ). Then let  $x^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} x_3^{\alpha_3}$  and  $|\alpha|_1 = \alpha_1 + \alpha_2 + \alpha_3$ . A polynomial  $p \in P_k$  if and only if it can be written as

$$p(x) = \sum_{|\alpha|_1 \leq k} a_\alpha x^\alpha$$

for some choice of coefficients  $a_\alpha \in \mathbb{C}$ . Similarly,  $\tilde{p} \in \tilde{P}_k$  if and only if

$$\tilde{p}(x) = \sum_{|\alpha|_1 = k} a_\alpha x^\alpha$$

for some choice of coefficients  $a_\alpha \in \mathbb{C}$ .

We shall also need to use polynomial spaces defined on planes and lines. We shall write  $P_k(S)$ , where  $S$  is a portion of a plane or line, to denote the space of polynomials of total degree at most  $k$  in two variables (for a plane) or one variable (for a line) using an orthogonal coordinate system in the plane or arc length along the line. Thus, if  $e$  is a segment of a line and  $f$  is a subdomain of plane in  $\mathbb{R}^3$ , then

$$\begin{aligned} P_k(e) &= \{\text{polynomials of maximum total degree } k \text{ in arc length on } e\}, \\ P_k(f) &= \{\text{polynomials of maximum total degree } k \text{ in } \xi_1, \xi_2 \text{ on } f\}, \end{aligned}$$

where  $(\xi_1, \xi_2)$  is an orthogonal coordinate system in the plane containing  $f$ . In particular,

$$P_k(e) = \{p|_e | p \in P_k\} \text{ and } P_k(f) = \{p|_f | p \in P_k\}.$$

Note that in  $\mathbb{R}^3$ (5.8)

$$\dim(P_k) = \binom{N+k}{k},$$

where  $\dim(P_k)$  is the dimension of  $P_k$ . In  $\mathbb{R}^3(5.9)$

$$\dim(\tilde{P}_k) = \frac{1}{2}(k+2)(k+1).$$

A particularly important polynomial in  $P_1$  is the *barycentric coordinate* function or linear shape function. Let  $a_n$ ,  $n = 1, \dots, 4$ , be the vertices of a non-degenerate tetrahedron (i.e. having non-zero volume, or, equivalently, in which the vertices do not lie in a plane). The  $n$ th barycentric coordinate function is the polynomial denoted by  $\lambda_n(x)$  which is the unique function in  $P_1$  such that  $\lambda_n(a_j) = \delta_{n,j}$ ,  $1 \leq j \leq 4$ . The functions  $\{\lambda_n\}_{n=1}^4$  have the property that  $\sum_{n=1}^4 \lambda_n = 1$ . A proof of the existence and uniqueness of  $\lambda_n$  can be found in [80] and proceeds as follows. Suppose the vertices of the tetrahedra are  $a_n$ ,  $n = 1, \dots, 4$ . Then  $\lambda_n = \beta_1^{(n)} + \beta_2^{(n)}x_1 + \beta_3^{(n)}x_2 + \beta_4^{(n)}x_3$  and the coefficients (arranged as a vector  $\beta^{(n)} = (\beta_1^{(n)}, \beta_2^{(n)}, \beta_3^{(n)}, \beta_4^{(n)})^T$ ) satisfy  $A\beta^{(n)} = f^{(n)}$ , where the  $4 \times 4$  matrix  $A$  and vector  $f^{(n)}$  are given block-wise by

$$A = \left( \begin{array}{c|c|c|c} \frac{1}{a_1} & \frac{1}{a_2} & \frac{1}{a_3} & \frac{1}{a_4} \end{array} \right)^T \text{ and } f^{(n)} = e_n,$$

where  $e_n$  is the  $n$ th column of the  $4 \times 4$  identity matrix. The assumption that the elements are non-degenerate implies that  $A$  is non-singular, so the above equation for  $\beta^{(n)}$  has a unique solution.

In order to discuss finite elements built on hexahedra, we also need the following “tensor product” polynomial space:

$$\mathcal{Q}_{l,m,n} = \{\text{polynomials of maximum degree } l \text{ in } x_1, m \text{ in } x_2, \text{ and } n \text{ in } x_3\},$$

and with obvious notation  $\mathcal{Q}_{l,m}$  for polynomials of degree at most  $l$  in  $x_1$  and  $m$  in  $x_2$ . Similar to the definition of  $P_k$ , we define  $\mathcal{Q}_{l,m,n}$  in terms of surface coordinates. Note that

$$\dim(\mathcal{Q}_{l,m,n}) = (l+1)(m+1)(n+1).$$

Other more exotic spaces of polynomials will be introduced as we discuss the various finite element spaces.

We shall need the following result concerning the approximation of a function by polynomials in Sobolev spaces sometimes called the Deny–Lions theorem. Recall that  $\|\cdot\|_{H(K)}$  denotes the full Sobolev norm on  $H(K)$ , whereas  $|\cdot|_{H(K)}$  denotes the semi-norm involving only derivatives of degree exactly  $s$ .

**Theorem 5.5** Suppose  $K$  is a Lipschitz domain. Let  $k \geq 0$  be an integer. Then there exists a constant  $C$  such that

- (1) if  $p \in H(K)$  for some  $s$  with  $0 \leq s \leq k+1$  then (5.10)

$$\inf_{\varphi \in P_k} \|p + \varphi\|_{H^s(K)} \leq C|p|_{H^s(K)};$$

(2) if  $v \in (H^s(K))^3$  for  $0 \leq s \leq k + 1$  then (5.11)

$$\inf_{\varphi \in (P_k)^3} \|v + \varphi\|_{(H^s(K))^3}^3 \leq C|v|_{(H^s(K))^3}^3;$$

(3) if  $v \in (H^s(K))^3$  and  $\nabla \times v \in (H^s(K))^3$  for  $0 \leq s \leq k$  then (5.12)

$$\begin{aligned} & \inf_{\varphi \in (P_{k-1})^3} \left( \|v + \varphi\|_{(H^s(K))^3}^3 + \|\nabla \times (v + \varphi)\|_{(H^s(K))^3}^3 \right) \\ & \leq C \left( |v|_{(H^s(K))^3}^3 + |\nabla \times v|_{(H^s(K))^3}^3 + |\nabla \times v|_{(H^{[s]}(K))^3}^3 \right), \end{aligned}$$

where  $[s]$  is the integer part of  $s$  (see eqn(5.12) of [9]).

In all these estimates,  $C$  is independent of  $p$  or  $v$  but depends on  $K$  and  $s$ .

**Remark 5.6** To understand this theorem suppose a function  $p$  is smooth enough to have a Taylor series up to derivatives of order  $s$  (integer). Then by choosing  $\Phi$  to equal the polynomial formed by the terms up to derivatives of order  $s - 1$  in the Taylor series, and using the remainder term, we have (5.10). This approach can be generalized to the Sobolev space setting [60].

In order to prove this theorem, we need to know that the dual space of  $P_k$  is finite-dimensional. This follows by showing that the unisolvence of the barycentric functions has an analogue for higher-degree polynomials. This lemma is from [239].

**Lemma 5.7** Let  $K$  be a tetrahedron with vertices  $\{a_j\}_{j=1}^4$ . Then for any  $k \geq 1$  a polynomial  $p \in P_k$  is uniquely determined by its values on the principal lattice

$$L_k(K) = \left\{ x \in \mathbb{R}^3 \middle| x = \sum_{j=1}^4 \lambda_j a_j, \text{ where } \sum_{j=1}^4 \lambda_j = 1 \right. \\ \left. \text{and } \lambda_j \in \left\{ 0, \frac{1}{k}, \frac{2}{k}, \dots, \frac{(k-1)}{k}, 1 \right\}, 1 \leq j \leq 4 \right\}.$$

**Remark 5.8** The points in a unisolvent set for  $P_k$  do not have to be exactly on the principal lattice. In some cases it is helpful to perturb them while keeping the unique determination property [160]. This can help improve conditioning.

**Proof of Lemma 5.7** We provide only a sketch of the proof. Since  $\dim(P_k)$  and the number of points in  $L_k(K)$  are identical, we only need show that if  $p(x) = 0$  for all  $x \in L_k(K)$  then  $p = 0$ .

We start by considering the edges of  $K$ . On each edge the polynomial  $p$  is of degree  $k - 1$  in arc length and vanishes at  $k + 1$  points. Hence  $p = 0$  on each edge.

Now consider a face  $f$  which we can assume has vertices  $a_1, a_2$  and  $a_3$ . We know that, if  $k = 1$ , then  $p = 0$  on  $f$  since it vanishes at the four vertices of the tetrahedron and the argument before Theorem 5.5 shows that these uniquely

determine  $p$ . For  $k > 1$  we proceed by induction and assume that the degrees of freedom uniquely determine the polynomial for polynomials of degree  $l$  with  $l \leq k-1$  in  $\mathbb{R}^2$ . Then since  $p = 0$  on each edge of  $f$  we know  $p$  may be represented as  $p = \lambda_1 \lambda_2 \lambda_3 \tilde{p}$ , where  $\lambda_j$  is the barycentric function for node  $a_j$ ,  $1 \leq j \leq 3$ , and  $\tilde{p}$  is of degree  $(k-3)$  (if  $k = 2$  we readily have  $p = 0$ ). Then  $\tilde{p} = 0$  at the points in  $L_k(K)$  interior to  $f$  and hence by the induction hypothesis we have  $\tilde{p} = 0$ , so  $p = 0$  on  $f$ .

Now turning to the tetrahedron  $K$ , we can adopt the same proof by factoring four linear factors from  $p$  corresponding to the four faces of  $K$ . Induction again shows that  $p = 0$  for any  $k$ . This completes the proof of the lemma.  $\square$

Having proved Lemma 5.7, we can now prove the Deny–Lions theorem.

**Proof of Theorem 5.5** We follow the proof of Theorem 3.1.1 of [80]. Note that if  $K$  denotes the largest integer strictly less than  $s$  then

$$\inf_{\varphi \in P_k} \|p + \varphi\|_{H^s(K)} \leq \inf_{\varphi \in P_K} \|p + \varphi\|_{H^s(K)}.$$

Then, if  $N$  is the dimension of  $P_k$ , let  $P_K$  let  $\{f_l\}_{l=1}^N$  denote a basis for the dual space of  $P_k$ . For example,  $\{f_l\}_{l=1}^N$  could consist of point evaluation operators applied at the points in the fundamental lattice in Lemma 5.7. Then by the Hahn–Banach theorem these functionals may be extended to  $H(K)$ , such that for  $\Phi \in P_K(K)$ ,  $f_l(\Phi) = 0$  for all  $l$  if and only if  $\Phi = 0$ .

Now we prove that there is a constant  $C$  such that, for all  $p \in H(K)$ , (5.13)

$$\|p\|_{H^s(K)} \leq C \left( |p|_{H^s(K)} + \sum_{l=1}^N |f_l(p)| \right).$$

If not, there exists a sequence  $\{p_n\}_{n=1}^\infty$  such that  $\|p_n\|_{H^s(K)} = 1$  for all  $n$  and

$$|p_n|_{H^s(K)} + \sum_{l=1}^N |f_l(p_n)| \leq \frac{1}{n}.$$

Since  $\{p_n\}_{n=1}^\infty$  is bounded on  $H(K)$  and Theorem 3.7 shows that  $H(K)$  is compactly embedded in  $H^s(K)$ , we know there is a subsequence still denoted  $\{p_n\}_{n=1}^\infty$  such that  $p_n \rightarrow p \in H^s(K)$  strongly as  $n \rightarrow \infty$ . Since  $H(K)$  is complete, the fact that  $|p_n|_{H^s(K)} \leq 1/n$  and that the series converges in  $H^s(K)$  allows us to conclude that it converges in  $H(K)$ . Since  $|p_n|_{H^s(K)} \rightarrow 0$ , it follows that  $|p|_{H^s(K)} = 0$  and so we know  $p$  is a polynomial of degree less than or equal to  $K$ . The fact that  $f_l(p) = 0$ ,  $1 \leq l \leq N$ , then implies  $p = 0$ , which contradicts  $\|p\|_{H^s(K)} = 1$ .

Next, we apply (5.13) to  $p + \Phi$  for  $\Phi \in P_K$

$$\|(p + \Phi)\|_{H^s(K)} \leq C \left( |p|_{H^s(K)} + \sum_{l=1}^N |f_l(p + \Phi)| \right),$$

where we have used the fact that  $|\Phi|_{H(K)} = 0$ . Choosing  $\Phi$  so that  $f(p + \Phi) = 0$ ,  $1 \leq l \leq N$ , completes the proof of part 1 of the theorem. To prove part 2 we just apply part 1 to every component of  $v$ . Part 3 is proved in the same way as part 1.  $\square$

## 5.3 Meshes and affine maps

The first step in using most finite element software is to generate a finite element mesh covering the domain  $\Omega$  (recall that  $\Omega$  is assumed to be a Lipschitz polyhedron and hence can be meshed by tetrahedra [134] — we shall discuss curved boundaries later in Chapter 8). Abstractly, this means that we find a finite set  $\tau_b = \{K\}$  of subdomains (referred to as elements) such that

- (1)  $\Omega = \bigcup_{K \in \tau_b} K$  where  $\Omega$  denotes the closure of  $\Omega$
- (2) for each  $K \in \tau_b$ ,  $K$  is an open set with positive volume;
- (3) if  $K_1$  and  $K_2$  are distinct elements in  $\tau_b$ , then  $K_1 \cap K_2 = \emptyset$
- (4) each  $K \in \tau_b$  is a Lipschitz domain.

For each element  $K$ , we define the parameters  $b_K$  and  $Q_K$  such that

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

$$\rho_K = \text{diameter of largest sphere contained in } \bar{K},$$

then  $h = \max_{K \in \tau_b} h_K$  so that the index  $b$  denotes the maximum diameter of the elements  $K \in \tau_b$ .

For the purpose of theory, we suppose that there is a family of meshes  $\{\tau_b \mid b > 0\}$  and we analyze the error as  $b$  decreases. This is the standard  $b$ -version of the finite element method where we attempt to obtain convergence by refining the mesh. To be a well-defined finite element mesh we need more geometric constraints. These are simplest to state if every element  $K \in \tau_b$  is a tetrahedron or a hexahedron. Thus we now assume that the mesh consists of tetrahedra or hexahedra. It is necessary that the mesh satisfy the standard finite element geometric constraints, so that if  $K_1 \in \tau_b$  and  $K_2 \in \tau_b$  and if  $K_1 \cap K_2 \neq \emptyset$  then the elements meet in one of the following ways:

- the elements meet at a single point that is a vertex for both elements;
- the elements meet along a common edge and the endpoints of the edge are vertices of the two elements;
- the elements meet at a common face and the vertices of the face are vertices of both elements.

We shall later discuss prismatic elements and curvilinear elements. The generalization of the above geometric constraints to these elements is obvious.

From the point of view of implementation (but not necessarily analysis) the simplest mesh to generate is a tetrahedral mesh of a polyhedral domain. Thus each  $K \in \tau_b$  is a tetrahedron. The reference element is also a tetrahedron and, in this book, it is defined to be the tetrahedron  $\hat{K}$  with vertices  $\hat{a}_1, \dots, \hat{a}_4$  given by  $\hat{a}_1 = (0, 0, 0)^\top$ ,  $\hat{a}_2 = (1, 0, 0)^\top$ ,  $\hat{a}_3 = (0, 1, 0)^\top$ , and  $\hat{a}_4 = (0, 0, 1)^\top$ . Any  $K \in \tau_b$

can be obtained by mapping  $\hat{K}$  using an affine map. By this we mean that for any  $K \in \tau_b$  there is a map  $F_K : \hat{K} \rightarrow K$  such that  $F_K(\hat{K}) = K$  and (5.14)

$$F_K \hat{x} = B_K \hat{x} + b_K,$$

where  $B_K$  is a non-singular  $3 \times 3$  matrix, and  $b_K$  is a vector. The non-singularity of  $B_K$  is a result of the fact that we assumed that  $K$  has a non-empty interior since the volume of  $K$  is  $|\det(B_K)|/6$  (the factor 1/6 comes from the fact that the volume of the reference element is 1/6). In practice, it is easy to compute  $B_K$  and  $b_K$ , since if  $K$  has vertices  $a_1, \dots, a_4$  and if we choose  $F_K$  to satisfy  $F_K(\hat{a}_i) = a_i$  for  $1 \leq i \leq 4$  then  $b_K = a_1$  and  $B_K$  is the matrix with  $j$ th column given by  $a_{j+1} - a_1$ .

We shall be concerned with mapping between functions defined on the reference tetrahedron  $\hat{K}$  and the target tetrahedron  $K$ . For a simple scalar function, we need only change variable in the usual way as discussed in Section 3.9. Thus, if  $\hat{p}$  is a scalar function defined on  $\hat{K}$ , we obtain a corresponding function  $p$  on  $K$  by (5.15)

$$p(F_K(\hat{x})) = \hat{p}(\hat{x})$$

or, equivalently,  $p \circ F_K = \hat{p}$ , where  $\circ$  denotes composition of functions. A simple calculation using the chain rule shows that the gradient transforms as follows: (5.16)

$$(\nabla p) \circ F_K = B_K^{-T} \widehat{\nabla} \hat{p},$$

where  $\widehat{\nabla}$  is the gradient with respect to  $\hat{x}$ . It is clear from these relations that properties of  $B_K$  will determine the effects of the affine map. In particular, keeping track of the change of variables, it is possible to prove the following result (c.f. [80]).

**Lemma 5.9** For each  $m \geq 0$  and real  $p$  with  $1 \leq p < \infty$  the mapping  $\hat{v} \mapsto v = \hat{v} \circ F_K^{-1}$  is an isomorphism from  $W^{m,p}(\hat{K})$  onto  $W^{m,p}(K)$  and the following bounds hold:

$$\begin{aligned} \|\hat{v}\|_{W^{m,p}(\hat{K})} &\leq C_1 |B_K|^m |\det(B_K)|^{-1/p} \|v\|_{W^{m,p}(K)} \quad \text{for all } v \in W^{m,p}(K), \\ \|v\|_{W^{m,p}(K)} &\leq C_1 |B_K^{-1}|^m |\det(B_K)|^{1/p} \|\hat{v}\|_{W^{m,p}(\hat{K})} \quad \text{for all } \hat{v} \in W^{m,p}(\hat{K}), \end{aligned}$$

where  $|B_K|$  denotes the spectral norm of  $B_K$ .

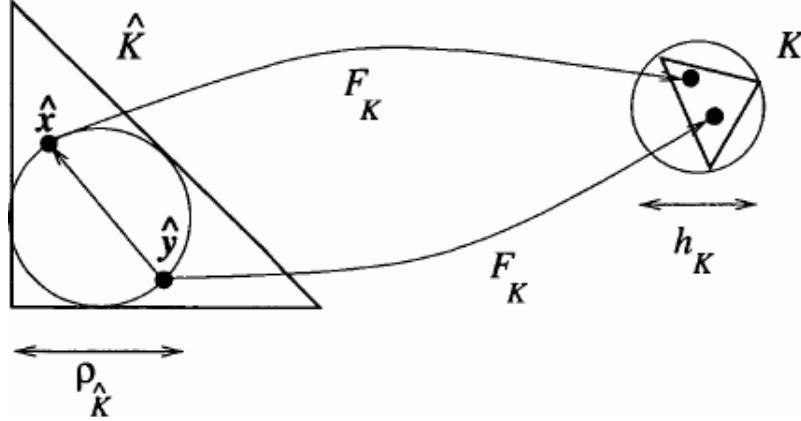
**Proof** For a full proof, see [80]. We shall prove the result in the case  $m = 1$ . Using (5.16) we see that

$$\begin{aligned} \int_K |\nabla v|^p dV &= \int_{\hat{K}} \left| B_K^{-T} \widehat{\nabla} \hat{v} \right|^p |\det(B_K)| d\hat{V} \quad \text{and} \\ \int_{\hat{K}} \left| \widehat{\nabla} \hat{v} \right|^p d\hat{V} &= \int_{\hat{K}} \left| B_K^T \nabla v \right|^p \frac{1}{|\det(B_K)|} dV. \end{aligned}$$

Hence

$$\int_K |\nabla v|^p dV \leq |\det(B_K)| \left| B_K^{-T} \right|^p \int_{\hat{K}} \left| \widehat{\nabla} \hat{v} \right|^p d\hat{V} \quad \text{and}$$

Fig. 5.2. Geometry of the estimate of  $|B_K|$ . For simplicity, we show the two-dimensional case. Suppose two points  $\hat{o}$  and  $\hat{j}$  are on the circle of diameter  $\rho_{\hat{K}}$ . They are mapped into the triangle  $K$  and hence the mapped points are at most  $h_K$  apart.



$$\int_{\hat{K}} |\widehat{\nabla} \hat{v}|^p dV \leq \frac{|B_K^T|^p}{|\det(B_K)|} \int_K |\nabla v|^p dV .$$

Taking the  $p$ th root of both sides completes the proof. Repeated use of (5.16) proves the general case.  $\square$

The previous lemma shows that we must understand how the norm of  $B_K$  and  $B_K^{-1}$  depends upon  $K$ .

**Lemma 5.10** Let  $\hat{K}$  and  $K$  be affine equivalent (by which we mean that there is an invertible affine map  $F_K$  such that  $F_K(\hat{K}) = K$ ). Then

$$|B_K| \leq \frac{h_K}{\rho_{\hat{K}}} \quad \text{and} \quad |B_K^{-1}| \leq \frac{h_{\hat{K}}}{\rho_K} .$$

In addition, there are constants  $C_1 > 0$  and  $C_2$  independent of  $h_K$  and  $\rho_K$  such that

$$C_1 \rho_K^3 \leq |\det(B_K)| \leq C_2 h_K^3 .$$

**Proof** By the definition of the spectral norm norm,

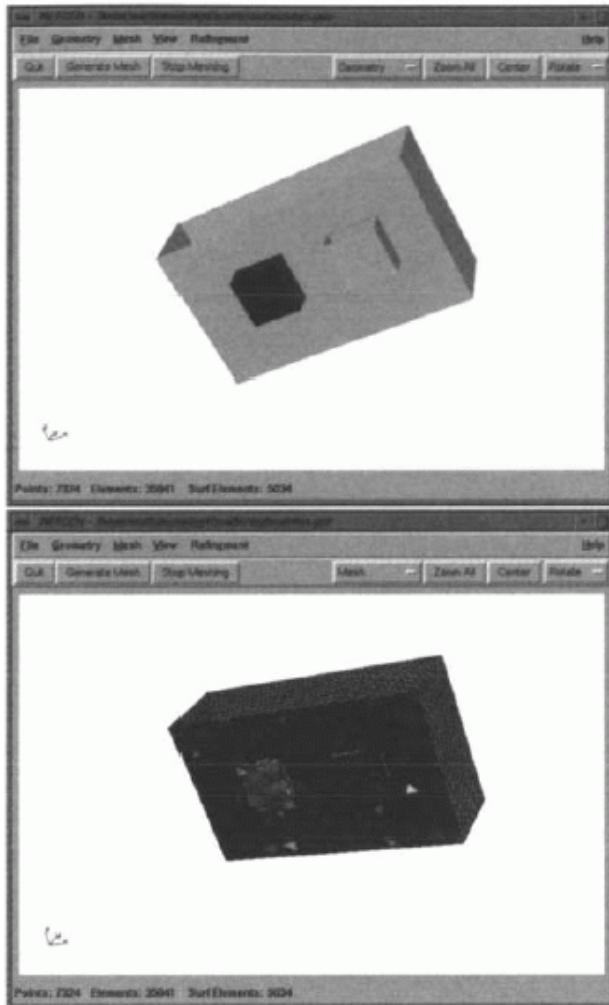
$$|B_K| = \sup_{\|\xi\|} |B_K \xi| = \frac{1}{\rho_{\hat{K}}} \sup_{\|\xi\|=1} |B_K(\rho_{\hat{K}} \xi)| .$$

Hence

$$|B_K| = \frac{1}{\rho_{\hat{K}}} \sup_{\|\xi\|=\rho_{\hat{K}}} |B_K \xi| .$$

However, since a ball of radius  $\rho_K$  is contained in  $\hat{K}$ , if  $|\xi| = \rho_K$  then there are vectors  $\hat{x}, \hat{y} \in \hat{K}$  such that  $\hat{x} - \hat{y} = \xi$  (see Fig. 5.2). Upon applying the affine

Fig. 5.3. An example of a tetrahedral mesh computed by the mesh generator NETGEN [269]. This mesh generator allows the user to define simple surfaces, and then attempts to fill the resulting volume with tetrahedra. *Top*: half of the domain cut by a plane. The domain  $\Omega$  is interior to a parallelepiped and exterior to one of the two cubes, and includes the interior of the other. One cube is a perfect conductor and the other contains a penetrable medium. *Bottom*: a mesh generated by NETGEN showing the tetrahedra. The penetrable cube has been filled with tetrahedra. Some experimentation with meshing parameters such as the desired mesh size and rapidity of change of mesh size within the grid (or granularity of the mesh) is usually required to generate an acceptable mesh.



map  $B_k \xi = B_k(\hat{x} - \hat{y}) = F_k \hat{x} - F_k \hat{y}$ . By the definition of the affine map,  $F_k \hat{x} - F_k \hat{y} \in K$  and hence  $|B_k \xi| \leq |F_k \hat{x} - F_k \hat{y}| \leq b_k$  (again see Fig. 5.2). This completes the proof of the first inequality. The second inequality in the theorem is proved in a similar way by considering the map from  $K$  to  $K$ .

To prove the estimate for  $\det(B_K)$ , we note that  $|\det(B_K)| = \text{vol}(K)/\text{vol}(K)$ .  $\square$

From this theorem, it is clear that it is important to quantify the relationship between  $h_K$  and  $\varrho_K$ . To do this we shall restrict our attention to meshes that form a regular family. By this we mean the following:

**Definition 5.11** Let

$$\sigma_K = h_K / \varrho_K \quad \text{and} \quad \sigma_h = \max_{K \in \mathcal{T}_h} \sigma_K.$$

We say that a family of meshes is *regular* as  $h \rightarrow 0$  if there are constants  $\sigma_{\min} > 0$  and  $h_0 > 0$  such that

$$\sigma_h \geq \sigma_{\min} \quad \text{for all } h \text{ with } 0 < h \leq h_0.$$

In essence, a family of meshes is regular if the tetrahedra do not flatten out during mesh refinement. We note that irregular meshes in which  $\sigma_{\min} \approx 0$  usually result in ill-conditioned matrix problems and can result in poor approximation compared to more regular meshes. Nevertheless, it is sometimes desirable to use rather irregular meshes if the solution is known *a priori* to change more rapidly in one direction than another.

Mesh generators usually require a geometric model of the surface of  $\Omega$ . In practice, this is usually supplied from a CAD package, although most academic mesh generators will allow the user to input a description of the geometry using simple primitives such as spheres, cubes, etc. See, for example, Fig. 5.3. Once the user has supplied a geometric description, the mesh generator will attempt to mesh the domain to the desired value of  $h$  (possibly also refining in selected regions). Output from the mesh generator will at least consist of a list of coordinates of the vertices in the mesh, as well as a list, by tetrahedron, of the vertices belonging to each tetrahedron. There will also be some method for flagging nodes on the boundary of  $\Omega$  and indicating if the tetrahedra are in different subdomains (e.g. each subdomain may represent a different material so that the subdomain label can be used to set different electrical parameters on the various tetrahedra). Mesh generators usually succeed in producing a non-degenerate mesh (although this should be checked before using the mesh!).

Tetrahedral meshes are generally referred to as “unstructured” since the arrangement of elements (see Fig. 5.3), and, in particular, the index of the vertices and neighboring elements cannot be predicted to follow a fixed pattern ahead of mesh generation. By contrast, a “structured” mesh is akin to a finite difference grid, in that it is easy to determine element vertices and neighbors. The unstructured grid imposes a programming overhead, and performance penalty, compared to structured grids, but has the advantage that it can fit a more general geometry.

The most common structured grid is one composed of parallelepipeds (with edges parallel to the coordinate axes). In such a mesh the arrangement of unknowns is (apart from at the boundaries) translation invariant throughout the

mesh and problems such as node numbering and matrix assembly are much easier than for an unstructured mesh. Obviously, structured meshes made of parallelepipeds are only able to cover simple domains, but this disadvantage is often outweighed by easier implementation and greater efficiency compared to unstructured codes.

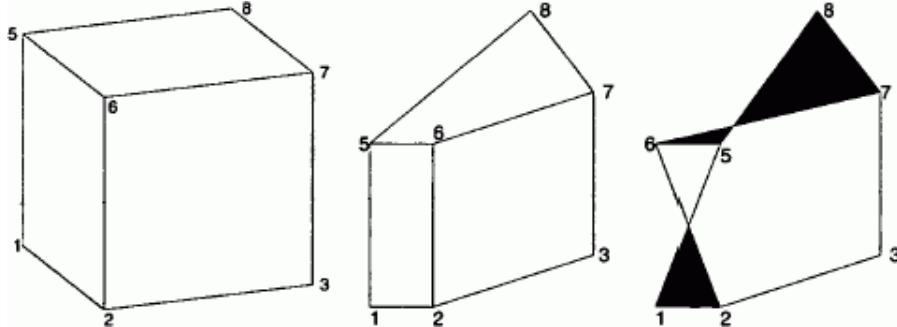
To improve the ability of hexahedral meshes to fit the domain of calculation, it is usual to form grids from general curvilinear hexahedra. In this case the grid  $\tau, b > 0$ , consists of elements  $K$  (of maximum diameter  $b$ ) such that each  $K$  can be obtained by an invertible mapping  $F_K : \hat{K} \rightarrow K$ , where  $\hat{K} = (0,1)^3$  is the reference element.

A typical example of such a mapping is to use a trilinear map (so edges of  $\hat{K}$  map to linear edges in  $K$ , but, of course, in general the faces of  $K$  are curvilinear). If  $\hat{K}$  has vertices  $\hat{a}_1, \dots, \hat{a}_8$  (see Fig. 5.4) and  $K$  has vertices  $a_1, \dots, a_8$  we can take the map to be

$$F_K(\hat{x}) = \sum_{m=1}^8 a_m \phi_m(\hat{x}),$$

where  $\phi_m \in \mathcal{Q}_{1,1,1}$  and  $\phi_m(\hat{a}_l) = \delta_{ml}$ ,  $1 \leq l, m \leq 8$  (we shall prove these functions are well defined in Section 6.4). We need to make sure that the points  $a_1, \dots, a_8$  have the same connectivity with respect to edges as  $\hat{a}_1, \dots, \hat{a}_8$  in  $\hat{K}$  so that the mapped element does not collapse (see Fig. 5.4). In particular, we need that  $\det(dF_K)$  is strictly positive or strictly negative on  $K$ . The trilinear map has the advantage (since  $\mathcal{Q}_{1,1,1}$  with vertex values specified results in a continuous piecewise trilinear function on the mapped mesh — see Section 6.4) that if adjacent elements are obtained in this way, then their common face agrees and no gaps open up between faces of adjacent elements.

Fig. 5.4. The reference hexahedron and its image under a trilinear map. Left: the reference element and node numbering. Center: a mapped element in which the numbering of the vertices in the image is such that the correct connectivity of the vertices is maintained and the result is a curvilinear element. Right: the vertices are connected in the wrong way, and the resulting element is singular!



In high-order codes, particularly the  $hp$  method code of [286], more exotic maps are used to obtain curvilinear hexahedra such that selected faces fit boundaries occurring in the problem (e.g. the surface of a curvilinear scatterer, or a curvilinear boundary between materials of different types). Away from such boundaries, the trilinear map is usually used to save time.

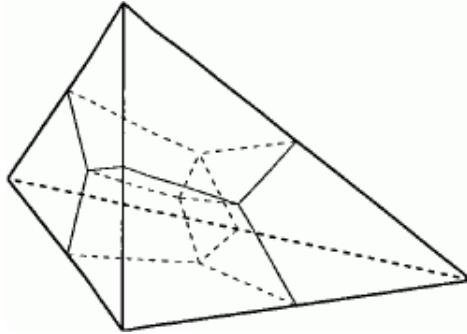
In this book we will generally not discuss the use of arbitrary hexahedra. This is because the theory is not well developed for this general case (but see Section 8.3 for more details on curvilinear elements). In the next chapter, when we discuss hexahedral elements, we shall assume that  $F_k(\hat{x}) = B_k \hat{x} + b_k$ , where  $B_k$  is a diagonal matrix so that all elements in the mesh are parallelepipeds with edges parallel to the coordinate axes.

It should not be thought that mapped hexahedral grids obtained by using the trilinear map are less flexible than tetrahedra for filling space. It is true that tetrahedral mesh generating software appears better developed than hexahedral grid software at the current time. However, Gary Cohen has pointed out that any tetrahedral grid can be converted to a mapped hexahedral grid by the simple expedient of decomposing each tetrahedron into four hexahedra using the centroid of the tetrahedron, the centroid of each face and the mid point of each edge of the tetrahedron as new vertices for the hexahedral mesh (see Fig. 5.5). In this case the hexahedral grid will be unstructured since the arrangement of hexahedra does not follow a simple finite difference lattice. Of course, a mapped hexahedral grid can also be decomposed into curvilinear tetrahedra in the usual way by decomposing each hexahedron using five tetrahedra (preferable) or six tetrahedra (easier, but possibly less accurate for low-order edge elements).

## 5.4 Divergence conforming elements

We start our study of finite element spaces by describing in detail a basic family of elements in  $H(\text{div}; \Omega)$  due to Nédélec [233]. This family extends to three dimensions the classical divergence conforming elements of Raviart and Thomas [261] (in two dimensions elements are often known in the electromagnetics literature as

Fig. 5.5. Decomposition of a tetrahedral element into four hexahedra by adding new vertices at the centroid of each geometric part (edge, face and volume) of the tetrahedron.



the Rao–Wilton–Glisson elements [258]). The three-dimensional family of divergence conforming elements is used in some time-dependent codes for Maxwell's equations, and is a theoretical tool for our later error analysis. At lowest order the degrees of freedom for this element are associated with faces in the mesh (just the average flux across the face) and so these elements are sometimes referred to as “face elements”.

In order to define these elements we shall need to use a special space of vector polynomials. For each  $k > 0$  we define(5.17)

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

Obviously,  $u \in D_1$  if and only if  $u = a + bx$ , where  $a \in C^3$  and  $b \in C$ . This implies that  $D_1$  will need four degrees of freedom in order to be specified uniquely. In general, using (5.8) and (5.9) we have the following result.

**Lemma 5.12** *The dimension of  $D_k$  is  $\frac{1}{2}(k+3)(k+1)k$ .*

We shall also need to know the space containing the divergence of the functions in  $D_k$  and in this regard we have the following lemma:

**Lemma 5.13** *Let  $D_k$  be as defined in (5.17), then  $\nabla \cdot D_k = P_{k-1}$ .*

**Proof** If  $u \in D_k$  then  $u(x) = p(x) + q(x)x$  and since  $q \in P_{k-1}$  we can easily show by direct computation that  $\nabla \cdot (q(x)x) = (k+2)q(x)$ . But  $\nabla \cdot (P_{k-1})^3 = (P_{k-2})^3$  and we have proved the desired result.  $\square$

Next we give the definition of the divergence conforming element on the reference tetrahedron  $K$ . Before doing this we need one more remark about notation. In the definition,  $\hat{f}$  will denote a general face of  $K$  with outward unit normal  $\hat{v}$ .

**Definition 5.14** (*Divergence conforming finite element*) See Fig. 5.6 . The element is defined as follows:

- $\hat{K}$  is the reference tetrahedron;
- $P_K = D_k$ ;
- the degrees of freedom  $\sum_{\hat{f}} M_{\hat{f}}(\hat{u}) \cup M_{\hat{K}}(\hat{u})$ , where the sets of moments  $M_{\hat{f}}(\hat{u})$  and  $M_{\hat{K}}(\hat{u})$  are defined for any sufficiently smooth  $\hat{u}$  as follows:

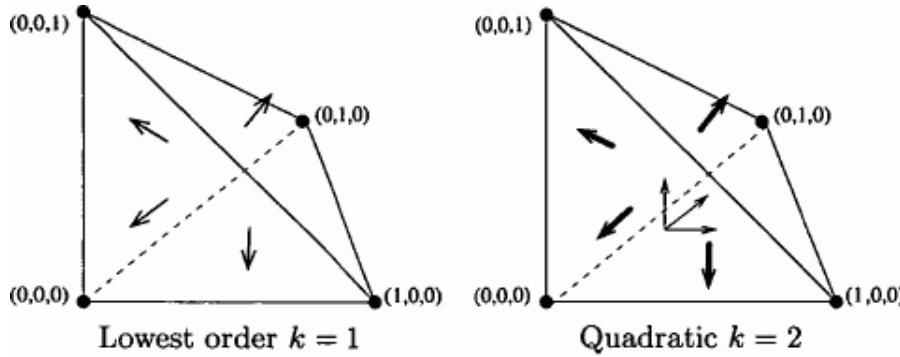
(5.18)

$$M_{\hat{f}}(\hat{u}) = \left\{ \int_{\hat{f}} \hat{u} \cdot \hat{v} \hat{q} d\hat{A} \text{ for all } \hat{q} \in P_{K-1}(\hat{f}) \text{ for each face } \hat{f} \right\}, \quad (5.19)$$

$$M_{\hat{K}}(\hat{u}) = \left\{ \int_{\hat{K}} \hat{u} \cdot \hat{q} d\hat{V} \text{ for all } \hat{q} \in (P_{K-2})^3 \right\},$$

We shall need to know when these degrees of freedom are well defined. Ideally, we would like to require only that  $\hat{u} \in H(\text{div}; \hat{K})$  but this is not possible since the trace of such a function is not sufficiently smooth for the degrees of freedom (5.18) to be well defined. Instead, we shall use a much more stringent criterion:

Fig. 5.6. A graphical representation of the degrees of freedom for the first two divergence conforming elements. *Left:*  $k = 1$ ; the average value of normal component of the finite element vector field is given on each face. *Right:*  $k = 2$ ; there are three normal component degrees of freedom per face (represented by the bold face normal vectors on each face), and, in addition, three interior degrees of freedom represented by the three vectors (not bold face) at the centroid of the tetrahedron.



**Lemma 5.15** The degrees of freedom (5.18) and (5.19) are well defined for any  $\hat{u} \in (H^{1/2 + \delta}(\hat{K}))^3$ ,  $\delta > 0$ .

**Remark 5.16** This is not an optimal result. In [61] it is shown that the degrees of freedom are well defined if  $\hat{u} \in (L^q(\hat{K}))^3$ ,  $q > 2$  and  $\nabla \cdot \hat{u} \in L^2(\hat{K})$ .

**Proof of Lemma 5.15** Since  $\hat{u} \in (L^2(\hat{K}))^3$ , the degrees of freedom  $M_{\hat{K}}$  are defined and bounded (via the Cauchy–Schwarz inequality). By the Trace Theorem 3.9,  $\hat{u}|_J \in (H^{\delta}(\hat{J}))^3 \subset (L^2(\hat{J}))^3$  and hence the degrees of freedom in  $M_J$  are also defined and bounded.  $\square$

We need to extend the divergence conforming element to a general tetrahedron  $K$ . This is done by relating the finite element function on tetrahedron  $K$  to a function on the reference element  $\hat{K}$ . Since the elements are vectorial and we wish to map divergences to divergences, we can no longer use the simple change of variable (5.15). Instead, following Section 3.9, we relate  $u(x)$  on an arbitrary tetrahedron  $K$  to  $\hat{u}(\hat{x})$  on the reference element  $\hat{K}$  by (3.77) which for the affine map  $F_K$  is written as (5.20)

$$u \circ F_K = \frac{1}{\det(B_K)} B_K \hat{u} .$$

Note that if  $\hat{v}$  is the unit outward normal to  $\hat{K}$  then the vector  $v$  on  $K$  given by (5.21)

$$v \circ F_K = \frac{1}{\left| \begin{pmatrix} B_K^{-1} \\ \hat{v} \end{pmatrix}^T \right|} \left( \begin{pmatrix} B_K^{-1} \\ \hat{v} \end{pmatrix}^T \right)^T \hat{v}$$

is a unit normal to  $K$  (it may be inward or outward depending on the sign of determinant of  $B_K$ ). This can be checked by a tedious manipulation or by noting

that the normal is a gradient of a suitable linear function and hence transforms like a gradient.

We now wish to show that the transformed element is well defined. First we show that the space  $D_k$  is invariant under the transformation (5.20).

**Lemma 5.17**  $D_k$  is invariant under the transformation (5.20).

**Proof** If  $\hat{u} \in D_k$  then  $\hat{u} = \hat{p} + \hat{q}\hat{x}$ , where  $\hat{p} \in (P_{k-1})^3$  and  $\hat{q} \in \mathcal{P}_k$ . Hence

$$B_K \hat{u} = B_K \hat{p} + \hat{q} B_K \hat{x} = B_K \hat{p} - \hat{q} b_K + \hat{q} x,$$

where we have used (5.14). But  $B_K \hat{p} \circ F_K^{-1} + \hat{q} \circ F_K^{-1} \in (P_{k-1})^3$  and since  $\hat{q}$  is homogeneous of degree  $k$ ,  $\hat{q} \circ F_K^{-1} = q + q_1$ , where  $q \in \mathcal{P}_{k-1}$  and  $q_1 \in P_{k-2}$ . This completes the verification of the claim once we have written  $u = B_K \hat{u} \circ F_K^{-1} = (B_K \hat{p} \circ F_K^{-1} + \hat{q} \circ F_K^{-1} + q_1 x) + qx$ .  $\square$

Next we need to relate the degrees of freedom on  $K$  and  $\hat{K}$  and show that they are invariant under the transformation (5.20).

**Lemma 5.18** Suppose  $\det(B_k) > 0$  and that the normals  $\nu$  on  $K$  and  $\hat{\nu}$  on  $\hat{K}$  are related by (5.21). Suppose also that the degrees of freedom of a function  $u$  on  $K$  are given by (5.22)

$$M_f(u) = \left\{ \int_f u \cdot v q dA \text{ for all } q \in P_{K-1}(f) \text{ for each face } f \text{ of } K \right\}, \quad (5.23)$$

$$M_K(u) = \left\{ \int_f u \cdot q dV \text{ for all } q \text{ such that } q \circ F_K = B_K^{-T} \hat{q}, \text{ and } \hat{q} \in (P_{K-2})^3 \right\}.$$

Then the degrees of freedom for  $\hat{u}$  on  $\hat{K}$  and for  $u$  on  $K$  (transformed by (5.20)) are identical.

**Remark 5.19** If  $\det(B_k) < 0$ , the degrees of freedom are again identical modulo suitable sign changes.

**Proof of Lemma 5.18** Using (5.20) and the assumed transformation for  $q$  (and canceling the factor  $\det(B_k)$ ), (5.24)

$$\int_K u \cdot q dV = \int_{\hat{K}} \hat{u} \cdot (B_K^\top)(B_K^\top \hat{q}) d\hat{V} = \int_{\hat{K}} \hat{u} \cdot q d\hat{V}.$$

Thus the degrees of freedom in  $M_{\hat{K}}(\hat{u})$  are invariant. In the same way, using (5.20) and transforming the normal vector by (5.21) we can transform the degrees of freedom in (5.22) to obtain

(5.25)

$$\int_f u \cdot v q dA = \int_{\hat{f}} \frac{1}{\det(B_K)} \left| B_K^{-T} \hat{v} \right| \hat{u} \cdot \hat{v} \frac{\text{area}(f)}{\text{area}(\hat{f})} d\hat{A}.$$

Using an orthogonal coordinate system with one axis along  $\hat{v}$ , we can then see that

$$\frac{\text{area}(f)}{\det(B_K) \left| B_K^{-T} \hat{v} \right| \text{area}(\hat{f})} = 1,$$

which shows that the degrees of freedom  $M(u)$  are also invariant (for a more general result see (3.81) which can be used to prove this lemma even for general smooth transformations).  $\square$

Next we prove a lemma that implies (as we shall see soon) that the elements are globally divergence conforming. It will also be used in the proof of unisolvence.

**Lemma 5.20** *If  $u \in D_k$  and if all the degrees of freedom of type (5.22) on a face  $f$  vanish then  $u \cdot v = 0$  on that face.*

**Proof of Lemma 5.20** Since  $u \in D_k$ ,  $u = p + qx$  for some  $p \in (P_{k-1})^3$  and  $q \in P_{k-1}$ . Thus, if  $f$  contains the vertex  $a$  then for all  $x \in f$ ,

$$u \cdot v = p \cdot v + qx \cdot v = p \cdot v + qa \cdot v \in P_{K-1}(f).$$

The choice of  $q = u \cdot v$  in the degrees of freedom of type (5.22) shows that  $u \cdot v = 0$ .  $\square$

Next we show that the element in Definition 5.14 is unisolvant. We first note that there are

$$\begin{aligned} \dim(P_{K-1}(f)) + 3\dim(P_{K-2}) &= \frac{1}{2}k(k+1) + \frac{1}{2}(k+1)k(k-1) \\ &= \frac{1}{2}(k+3)(k+1)k \end{aligned}$$

degrees of freedom which is equal to the dimension of  $D_k$ . So unisolvence can be proved by either proving the existence of a function  $u \in D_k$  corresponding to an arbitrary choice of the degrees of freedom, or by showing that any function consistent with an arbitrary choice of degrees of freedom is unique. We choose to prove uniqueness, and from the linearity of the degrees of freedom, this is equivalent to proving that the only function in  $u \in D_k$  having all degrees of freedom vanish is the function  $u = 0$ . This is proved in the following lemma.

**Lemma 5.21** *If all the degrees of freedom (5.22) and (5.23) of a function  $u \in D_k$  vanish then  $u = 0$ .*

**Proof** First we transform to the reference element and use the invariance of the degrees of freedom (Lemma 5.18) to conclude that all degrees of freedom vanish for  $\hat{u}$ . By Lemma 5.20,  $\hat{u} \cdot \hat{v} = 0$  on  $\partial\hat{K}$ . Using the divergence theorem

(in the form of (3.24)) and the fact that the degrees of freedom (5.23) vanish, shows that for every  $\hat{q} \in P_{k-1}$ ,

$$\int_{\hat{K}} \hat{\nabla} \cdot \hat{u} \hat{q} d\hat{V} = - \int_{\hat{K}} \hat{u} \cdot \hat{\nabla} \hat{q} d\hat{V} = 0.$$

Hence choosing  $\hat{q} = \nabla \cdot \hat{u}$  shows that  $\nabla \cdot \hat{u} = 0$ .

But, using the fact that  $\hat{u} \in D_k$ , we have that  $\hat{u} = \hat{p} + \hat{q}\hat{x}$  for  $\hat{p} \in (P_{k-1})^3$  and  $\hat{q} \in P_{k-1}$ . So  $\nabla \cdot \hat{u} = \nabla \cdot \hat{p} + (k+2)\hat{q}$  (this is a consequence of the fact that  $\hat{q}$  is homogeneous of degree  $k$ ). Since  $\nabla \cdot \hat{u} = 0$ , we conclude that  $\hat{q} = -\nabla \cdot \hat{p}/(k+2) \in P_{k-2}$ , which implies that  $\hat{q} = 0$  and  $\hat{u} = \hat{p} \in (P_{k-1})^3$ . Thus  $\hat{u} = (\hat{x}_{1\phi_1}, \hat{x}_{2\phi_2}, \hat{x}_{3\phi_3})^T$ , where  $\Phi = (\Phi_1, \Phi_2, \Phi_3)^T \in (P_{k-1})^3$ . If  $k=1$  this ends the proof since  $\Phi = 0$ . If  $k > 1$ , the choice of  $\hat{q} = \Phi$  in the degrees of freedom (5.23) shows that  $\Phi = 0$  and hence  $\hat{u} = 0$ . Mapping back to  $K$  shows that  $u = 0$  as claimed.  $\square$

The unisolvence of the element on a given tetrahedron  $K$  implies that there is a well-defined interpolation operator on  $K$  denoted  $w_K$ . By this we mean that if  $u \in (H^{1/2+\delta}(K))^3$ ,  $\delta > 0$  (see Lemma 5.15), then there is a unique finite element function  $w_K u \in D_k$  such that

$$M_f(u - w_K u) = \{0\} \text{ and } M_K(u - w_K u) = \{0\},$$

where  $M_f$  and  $M_K$  are the sets of degrees of freedom in (5.22) and (5.23). Obviously this is the same as requiring that for all faces  $f$  of  $K$  (5.26)

$$\int_f (u - w_K u) \cdot v q dA = 0 \text{ for all } q \in P_{k-1}(f),$$

and (5.27)

$$\int_K (u - w_K u) \cdot q dV = 0 \text{ for } q \circ F_K = B_K^{-T} \hat{q} \text{ for all } \hat{q} \in (P_{k-2})^3.$$

To prove error estimates, we need to prove that the interpolant on a general element  $K$  and interpolation on the reference element  $\hat{K}$  are related.

**Lemma 5.22** *Using the transformation (5.20), provided  $u$  is sufficiently smooth that  $w_K u$  is well defined, we have  $\widehat{w_K u} = w_{\hat{K}} \hat{u}$ .*

**Proof** By the definition of  $w_K u$ , all degrees of freedom of  $u - w_K u$  vanish on  $K$ . Hence by Lemma 5.18 all degrees of freedom for  $\widehat{w_K u} - \hat{u}$  vanish on  $\hat{K}$ . This implies that all degrees of freedom of  $w_{\hat{K}}(\widehat{w_K u} - \hat{u})$  vanish on  $\hat{K}$  and, by the unisolvence of the degrees of freedom guaranteed by Lemma 5.21, we know that  $w_{\hat{K}}(\widehat{w_K u} - \hat{u}) = 0$ . But  $D_k$  is invariant under interpolation and so  $w_{\hat{K}}(\widehat{w_K u} - \hat{u}) = \widehat{w_K u}$  and the lemma is proved.  $\square$

Now we can state a theorem summarizing the unisolvence and conformance of this family of elements. We assume that there is a regular family of meshes of  $\Omega$  denoted  $\{\tau_b\}_{b>0}$ . The global set of degrees of freedom is the union of all element degrees

$$\sum = \bigcup_{K \in \mathcal{T}_h} \sum K,$$

where  $\sum$  is given by Definition 5.14 (of course, the directions of the normals at the faces must be taken consistently).

**Theorem 5.23** *A vector function  $u \in D_k$  defined on tetrahedron  $K$  is uniquely determined by the degrees of freedom (5.22) and (5.23). Moreover, the space  $W_b$  of finite elements on the mesh  $\tau_b$ , defined element-wise by Definition 5.14 is divergence conforming, so that  $W_b \subset H(\text{div}; \Omega)$ .*

**Remark 5.24** *This theorem gives us an alternative characterization of  $W_b$  that is independent of the degrees of freedom:* (5.28)

$$W_h = \{u_h \in H(\text{div}; \Omega) \mid u_h|_K \in D_k \text{ for all } K \in \mathcal{T}_h\}.$$

**Proof of Theorem 5.23** The first part of the theorem just states unisolvence, which is proved in Lemma 5.21. The second part follows from Lemma 5.20, since if two tetrahedra  $K_1$  and  $K_2$  meet on a common face  $f$  with normal  $\nu$  pointing into  $K_2$ , then since the degrees of freedom on  $f$  of type (5.22) agree across the face (taking into account orientation of the normal), we know that the degrees of freedom (5.22) of  $u|_{K_1} - u|_{K_2}$  vanish on  $f$  and hence by Lemma 5.20  $(u|_{K_1} - u|_{K_2}) \cdot \nu = 0$ . By the characterization result in Theorem 5.3, this implies that  $u$  has a well-defined divergence in  $L^2(K_1 \cup K_2)$  and hence since  $K_1$  and  $K_2$  are arbitrary,  $u \in H(\text{div}; \Omega)$ .  $\square$

Now that we know the element is well-defined and divergence conforming, we can define a global interpolation operator  $w_h : (H^{1/2+\delta}(\Omega))^3 \rightarrow W_b$ ,  $\delta > 0$ , which is defined element-wise in terms of the element interpolant by

$$w_h u|_K = w_K u \text{ for each } K \in \mathcal{T}_h.$$

The following theorem proves an error estimate for this interpolant.

**Theorem 5.25** *Suppose  $\{\tau_b\}_{b>0}$  is a regular family of meshes on  $\Omega$  and  $0 < \delta < 1/2$ . Then if  $u \in (H^s(\Omega))^3$ ,  $1/2 + \delta \leq s \leq k$ , there is a constant  $C$  independent of  $b$  and  $u$  such that* (5.29)

$$\|u - w_h u\|_{L^2(\Omega)} \leq C h^s \|u\|_{(H^s(\Omega))^3}, \quad 1/2 + \delta \leq s \leq k.$$

**Remark 5.26** *It is also possible to estimate*

$$\|\nabla \cdot (u - w_h u)\|_{L^2(\Omega)} \leq C h^s \|\nabla \cdot u\|_{H^s(\Omega)} \text{ for } \frac{1}{2} + \delta \leq s \leq k,$$

see (5.60).

**Proof of Theorem 5.25** We shall only prove the result for integer  $s$ . For intermediate values of  $s$  the estimate can then be obtained by interpolation (see [60] for this type of argument). For  $1/2 + \delta \leq s < 1$  the result is proved using the integral characterization of the fractional order Sobolev norm, and essentially the same mapping procedure we shall use. For details the reader can consult [9].

By the definition of the norm and interpolant, we may write

$$\|u - w_h u\|_{(L^2(\Omega))^3}^2 = \sum_{K \in \mathcal{T}_h} \|u - w_h u\|_{(L^2(K))^3}^2.$$

Mapping to the reference element  $\hat{K}$  using (5.20), we find that

$$\begin{aligned} \|u - w_K u\|_{(L^2(K))^3}^2 &= \int_K |u - w_K u|^2 dV \\ &= \int_{\hat{K}} |B_K(\hat{u} - \widehat{w_K u})|^2 \frac{d\hat{V}}{|\det(B_K)|} \\ &\leq \frac{|B_K|^2}{|\det(B_K)|} \|\hat{u} - \widehat{w_K u}\|_{(L^2(\hat{K}))^3}^2. \end{aligned}$$

But by Lemma 5.10, we have

$$\|u - w_K u\|_{(L^2(K))^3} \leq C \frac{h_K}{|\det(B_K)|^{1/2}} \|\hat{u} - \widehat{w_K u}\|_{(L^2(\hat{K}))^3}.$$

Now we use Lemma 5.22, and the fact that  $(I - w_K)\hat{p} = 0$  for all  $\hat{p} \in (P_{k-1})^3$  to write (5.30)

$$\begin{aligned} \|\hat{u} - \widehat{w_K u}\|_{(L^2(\hat{K}))^3} &= \|\hat{u} - w_{\hat{K}} \hat{u}\|_{(L^2(\hat{K}))^3} \\ &= \|(I - w_{\hat{K}})(\hat{u} + \hat{p})\|_{(L^2(\hat{K}))^3} \end{aligned}$$

for all  $\hat{p} \in (P_{k-1})^3$ . By the Sobolev Imbedding Theorem 3.5, the value of each of the degrees of freedom of  $\hat{u}$  on  $\hat{K}$  can be estimated by the  $(H(\hat{K}))^3$  norm of  $\hat{u}$  for  $1/2 + \delta \leq s \leq k$ . Furthermore,  $P_k$  is a finite-dimensional vector space, so all norms are equivalent and hence

$$\|w_{\hat{K}}(\hat{u} + \hat{p})\|_{(L^2(\hat{K}))^3} \leq C \|\hat{u} + \hat{p}\|_{(H^s(\hat{K}))^3}.$$

Using the previous equality, equation (5.30) and Theorem 5.5, we have

$$\|\hat{u} - w_{\hat{K}} u\|_{(L^2(\hat{K}))^3} \leq C \inf_{\hat{p} \in (P_{k-1})^3} \|\hat{u} + \hat{p}\|_{(H^s(\hat{K}))^3} \leq C \|\hat{u}\|_{(H^s(\hat{K}))^3}.$$

Hence, combining the above estimates we obtain

$$\|u - w_K u\|_{(L^2(K))^3} \leq \frac{Ch_K}{|\det(B_K)|^{1/2}} \|\hat{u}\|_{(H^s(\hat{K}))^3}.$$

Mapping back to the reference element (using (5.20)) in the same way as in the proof of Lemma 5.9 we obtain

$$\|u - w_K u\|_{(L^2(K))^3} \leq \frac{Ch_K}{|\det(B_K)|^{1/2}} \frac{|\det(B_K)|^{1/2}}{\rho\kappa} h_K^s |u|_{(H^s(K))^3}.$$

Squaring and adding this estimate over all  $K \in \tau_h$  and using the regularity of the mesh to conclude that  $h_K/\rho_K$  is bounded independent of  $h$  completes the proof of (5.29).  $\square$

To complete our discussion of divergence conforming elements, let us consider in detail the important case  $k = 1$ . In this case, considering the reference element  $K$ , if we label the face opposite vertex  $\hat{a}_i$  as face  $\hat{f}_i$  with unit outward normal  $\hat{v}_i$  then the associated facial element  $\psi_i \in D_1$  satisfies

$$\int_{\hat{f}_j} \hat{\psi}_i \cdot \hat{v}_j dA = \delta_{i,j}.$$

This implies that  $\hat{\psi}_i = 2(\hat{x} - \hat{a}_i)$ . Then using (5.20), we see that

$$\psi_i = \frac{B_K}{\det(B_K)} 2(\hat{x} - \hat{a}_i) = 2 \frac{(\hat{x} - a_i)}{\det(B_K)}.$$

A direct calculation shows that if  $\det(B_K) > 0$

$$\int_{f_j} \psi_i \cdot v_j dA = \frac{6 \text{ vol}(K)}{\det(B_K)} = 1.$$

Thus, when  $\det(B_K) > 0$ ,

$$\int_{f_i} \psi_i \cdot v_i dA = \int_{\hat{f}_i} \hat{\psi}_i \cdot \hat{v}_i d\hat{A} = \delta_{i,i},$$

and so the degrees of freedom are invariant under the affine map as we verified in general.

When  $k = 1$ , the dimension of the space  $W_h$  is exactly the number of faces in the mesh, and the interpolant satisfies the error estimate

$$\|u - w_h u\|_{(L^2(\Omega))^3} \leq Ch^s \|u\|_{(H^s(\Omega))^3}, \quad 1/2 + \delta \leq s \leq 1,$$

so the maximum rate of convergence is  $O(h)$ .

## 5.5 The curl conforming edge elements of Nédélec

The elements we shall define and analyze in this section will be used later to discretize the electric field in Maxwell's equations and are due to Nédélec [233]. The lowest-order space (later  $k = 1$ ) has also been discovered independently by a number of authors [295, 42, 26, 206, 3] (and maybe others!). In particular, Whitney

[295] seems to have been the first to use the basic polynomial space for the lowest-order element. This explains why the lowest-order element is sometimes termed the Whitney element (although Whitney discovered the element in a different context). More generally, elements of the type discussed in this chapter are termed *edge elements* because, at lowest-order (again when  $k = 1$ ), the degrees of freedom are associated with edges of the mesh (see Fig. 5.7).

We shall present and analyze the elements in the manner of [143], which in turn follows Nédélec [233]. Of key importance is a relatively recent paper by Amrouche *et al.* [12] that provides the best characterization to date of the space of vector fields on which the classical edge interpolant is well defined. An alternative construction of edge elements (and the associated scalar spaces) which emphasizes a hierarchical and explicit description of the basis functions is given in [290, 13, 5] and a factorization is given in [149]. This family of elements includes the ones described in this section as well as the second family of edge elements due to Nédélec which we describe in Section 8.2. Higher order elements can also be constructed using differential forms [162, 164].

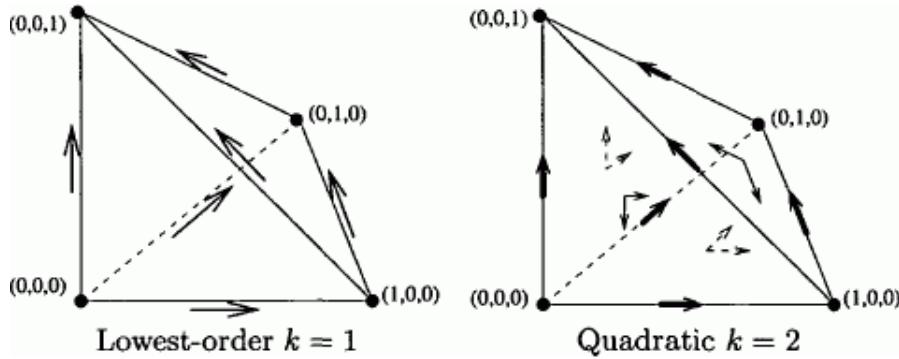
In order to define this family of edge elements, we need to define a special space of polynomials. We start by defining a subspace  $S_k$  of homogeneous vector polynomials of degree  $k$  by (5.31)

$$S_k = \left\{ p \in (\tilde{P}_k)^3 \mid x \cdot p = 0 \right\}.$$

We note that if  $p \in (\tilde{P}_k)^3$ , then  $x \cdot p \in \tilde{P}_{k+1}$  and any polynomial in  $\tilde{P}_{k+1}$  may be written in this way. Thus, using (5.8) and (5.9),

$$\begin{aligned} \dim(S_k) &= 3\dim \tilde{P}_k - \dim \tilde{P}_k + 1 \\ &= \frac{3}{2}(k+2)(k+1) - \frac{1}{2}(k+3)(k+2) = k(k+2). \end{aligned}$$

Fig. 5.7. A graphical representation of the degrees of freedom for the first two curl conforming elements. *Left:*  $k = 1$ ; degrees of freedom are the average value of tangential component of the vector field on each edge. *Right:*  $k = 2$ ; there are two degrees of freedom per edge (represented by the bold face vectors on each edge) and, in addition, two degrees on each face.



Having defined  $S_k$ , we can now define the important space  $R_k$  by (5.32)

$$R_k = (P_{k-1})^3 \oplus S_k.$$

The dimension of  $R_k$  can be calculated from the dimension of  $S_k$  given above and (5.8) as follows:

$$\begin{aligned} \dim(R_k) &= 3\dim(P_{k-1}) + \dim(S_k) \\ &= \frac{1}{2}(k+3)(k+2)k. \end{aligned}$$

In fact,  $R_k$  is quite a natural space of polynomials being part of a Helmholtz decomposition of  $(P_k)^3$ , as is shown in the following lemma:

**Lemma 5.27** *The following algebraic decomposition holds:*

$$(P_k)^3 = R_k \oplus \nabla \tilde{P}_{k+1}.$$

**Proof** Suppose  $u \in R_k \cap \nabla \tilde{P}_{k+1}$ . Then  $u = \nabla p$  for some  $p \in \tilde{P}_{k+1}$ . But, since  $p \in \tilde{P}_{k+1}$ , it is easy to show that  $p = (k+2)x \cdot \nabla p$ , and hence  $x \cdot u = x \cdot \nabla p \neq 0$  unless  $p = 0$ . Thus  $R_k \cap \nabla \tilde{P}_{k+1} = \{0\}$ . But  $\dim(\nabla \tilde{P}_{k+1}) + \dim(\nabla \tilde{P}_{k+1}) = \dim(\nabla \tilde{P}_k^3)$ . This proves the lemma.  $\square$

Edge finite element spaces depend on the use of the vector polynomial space  $R_k$  defined in (5.32). Before defining and analyzing the elements, we need to prove a further auxiliary result concerning a discrete Helmholtz decomposition of  $R_k$ .

**Lemma 5.28** *If  $u \in R_k$  satisfies  $\nabla \times u = 0$  then  $u = \nabla p$  for some  $p \in P_k$ .*

**Remark 5.29** *There is an analogue of  $R_k$  for two-dimensional domains, or surfaces. In this case we define, for a triangle  $f$  in the  $(x_1, x_2)$ -plane,*

$$S_k(f) = \left\{ p \in P_k^2 \mid p(x) \cdot x = 0 \text{ where } x = (x_1, x_2)^\top \right\},$$

and  $R_k(f) = (P_{k-1}(f))^2 \oplus S_k(f)$ . The above lemma remains true for this space, provided we interpret the curl as the scalar surface curl on  $f$  defined by  $\nabla_f \times u = \partial u_2 / \partial x_1 - \partial u_1 / \partial x_2$ .

**Proof of Lemma 5.28** Since  $\nabla \times u = 0$ , we know that  $u = \nabla p$  for some  $p \in H^1(K)$  (see Theorem 3.37). But  $u \in (P_k)^3$ , so we know that  $p \in P_{k+1}$ , and we can write  $p = p_1 + p_2$ , where  $p_1 \in P_k$  and  $p_2 \in \tilde{P}_{k+1}$ . However,  $u \in R_k$  so  $x \cdot \nabla p_2 = 0$  and since  $p_2$  is homogeneous  $x \cdot \nabla p_2 = kp_2$ , which implies that  $p_2 = 0$ .  $\square$

Our analysis of this element follows the same pattern as our analysis of the divergence conforming elements in the previous section. In particular, we make use of the reference element  $K$  and transform between reference and target elements. We first define the curl conforming elements on the reference element:

**Definition 5.30** (Curl conforming element) The curl conforming finite element is defined by

- $\mathcal{K}$  is the reference tetrahedron,
- $P_{\mathcal{K}} = R_k$ ,
- The degrees of freedom are of three types associated with edges  $\hat{e}$  of  $\hat{\mathcal{K}}$ , faces  $\hat{f}$  of  $\hat{\mathcal{K}}$  and  $\hat{\mathcal{K}}$  itself. We denote by  $\hat{\tau}$  a unit vector in the direction of  $\hat{e}$ . We define three different degrees of freedom as follows (see Fig. 5.7):
  - (1) the first set is associated with edges of the element:

$$M_{\hat{e}}(\hat{u}) = \left\{ \int_{\hat{e}} \hat{u} \cdot \hat{q} \hat{d}\hat{s} \text{ for all } \hat{q} \in P_{k-1}(\hat{e}) \text{ for each edge } \hat{e} \text{ of } \hat{\mathcal{K}} \right\},$$

- (2) the second set is associated with faces of the element:

$$M_{\hat{f}}(\hat{u}) = \left\{ \frac{1}{\text{area}(\hat{f})} \int_{\hat{f}} \hat{u} \cdot \hat{q} d\hat{A} \text{ for each face } \hat{f} \text{ of } \hat{\mathcal{K}}, \right. \\ \left. \hat{q} \in \left( P_{k-2}(\hat{f}) \right)^3 \text{ and } \hat{q} \cdot \hat{\nu} = 0 \right\},$$

- (3) the last set of degrees of freedom are associated with the volume:

$$M_{\hat{\mathcal{K}}}(\hat{u}) = \left\{ \int_{\hat{\mathcal{K}}} \hat{u} \cdot \hat{q} d\hat{V} \text{ for all } \hat{q} \in \left( P_{k-3}(\hat{\mathcal{K}}) \right)^3 \right\}.$$

Then  $\sum_{\mathcal{K}} = M_{\hat{e}}(\hat{u}) \cup M_{\hat{f}}(\hat{u}) \cup M_{\hat{\mathcal{K}}}(\hat{u})$ .

**Remark 5.31** The degrees of freedom  $M_{\hat{f}}$  need some comment since they appear to differ from those in [233]. If  $\hat{q} \in (P_{k-2}(\hat{f}))^3$  and  $\hat{q} \cdot \hat{\nu} = 0$  then  $\hat{q} = (\hat{\nu} \times \hat{q}) \times \hat{\nu}$ , so

$$\int_{\hat{f}} \hat{u} \cdot \hat{q} d\hat{A} = \int_{\hat{f}} \hat{u} \times \hat{\nu} \cdot \hat{q} \times \hat{\nu} d\hat{A}.$$

But  $\hat{q} \times \hat{\nu} = \hat{r} \in (P_{k-2}(\hat{f}))^2$ , where  $(P_{k-2}(\hat{f}))^2$  denotes the set of vector fields tangential to  $\hat{f}$  with each component in  $P_{k-2}(\hat{f})$ . So the new degrees of freedom and the original ones from [233] are equivalent. However, we find the new set easier to manipulate from the point of view of proving affine equivalence.

Because we are working in  $H(\text{curl}; \hat{\mathcal{K}})$ , the general results of Section 3.9 show that vectors in  $R_k$  must be transformed in a special way. If  $\hat{u} \in R_k$  on  $\hat{\mathcal{K}}$ , we define  $u$  on  $K$  by the transformation (3.76), which, in the special case of an affine map, becomes (5.33)

$$u \circ F_K = \left( B_K^T \right)^{-1} \hat{u}.$$

An important consequence of this formula is that the curl of  $u$  is related in an easy way to the curl of  $\hat{u}$  (see Corollary 3.58), i.e.

$$\nabla \times u = \frac{1}{\det(B_K)} B_K \widehat{\nabla} \times \hat{u}. \quad (5.34)$$

We shall also need to show how tangent vectors transform under the affine map. Let  $\hat{\tau}$  be a unit vector in the direction of an edge  $\hat{e}$  of the tetrahedron  $K$ . Then the vector  $\tau$  given by(5.35)

$$\tau = \frac{B_K \hat{\tau}}{\|B_K \hat{\tau}\|}$$

is a unit tangent vector to the edge  $e$  of  $K$  that is the image of  $\hat{e}$  under  $F_K$ .

Our next lemma shows that  $R_k$  is indeed invariant under the transformation (5.33). This is the first step in proving the affine invariance of the finite element space.

**Lemma 5.32**  $R_k$  is invariant under (5.33).

**Proof** If  $u \in R_k$ , then  $u = \hat{p}_1 + \hat{p}_2$ , where  $\hat{p}_1 \in (P_{k-1})^3$  and  $\hat{p}_2 \in S_k$ . Then

$$\begin{aligned} u(x) &= \left[ \left( B_K^\tau \right)^{-1} \hat{p}_1 + \left( B_K^\tau \right)^{-1} \hat{p}_2 \right] \left( F_K^{-1}(x) \right) \\ &= \left[ \left( B_K^\tau \right)^{-1} \hat{p}_1 \right] \left( F_K^{-1}(x) \right) + \left( B_K^\tau \right)^{-1} \hat{p}_2 \left( B_K^{-1}x - B_K^{-1}b_K \right). \end{aligned}$$

Since  $\hat{p}_2 \in (P_k)^3$ , we see that  $\hat{p}_2 \left( B_K^{-1}x - B_K^{-1}b_K \right) = \hat{p}_2 \left( B_K^{-1}x \right) + \hat{p}_3(x)$ , where  $\hat{p}_3 \in (P_{k-1})^3$ . Hence

$$u(x) = \left[ \left( B_K^\tau \right)^{-1} p_1 \left( F_K^{-1}(x) \right) + \hat{p}_3(x) \right] + \left( B_K^\tau \right)^{-1} \hat{p}_2 \left( B_K^{-1}x \right),$$

and  $\left( B_K^\tau \right)^{-1} \hat{p}_1 \circ F_K^{-1} + \hat{p}_3 \in (P_{k-1})^3$ . Furthermore,

$$\left( B_K^\tau \right)^{-1} \hat{p}_2 \left( B_K^{-1}x \right) \cdot x = \hat{p}_2 \left( B_K^{-1}x \right) \cdot \left( B_K^{-1}x \right) = 0,$$

and we conclude that  $u \in R_k$ .  $\square$

Now we can define the curl conforming element on a general tetrahedron  $K$ .

**Definition 5.33** The curl conforming finite element on an element  $K$  is defined by

- $K$  is a tetrahedron,
  - $P_k = R_k$ ,
  - The degrees of freedom are of three types associated with edges  $e$  of  $K$ , faces  $f$  of  $K$  and  $K$  itself. We denote by  $\tau$  a unit vector in the direction of  $e$ . There are three different degrees of freedom as follows:
- (1) the set is associated with edges of the element:(5.36)

$$M_e(u) = \begin{cases} \int_e u \cdot \tau q ds \text{ for all } q \in P_{k-1}(e) \\ \text{for each edge } e \text{ of } K \end{cases};$$

(2) the second set is associated with faces of the element.(5.37)

$$M_f(u) = \left\{ \frac{1}{\text{area}(f)} \int_f u \cdot q dA \text{ for each face } f \text{ of } K, \text{ and for all} \right. \\ \left. q = B_K \hat{q}, \hat{q} \in \left( P_{k-2}(\hat{f}) \right)^3, \hat{q} \cdot \hat{v} = 0 \right\};$$

(3) the last set is associated with the volume:(5.38)

$$M_K(u) = \left\{ \int_K u \cdot dV \text{ for all } q \text{ obtained by mapping } \hat{q} \in (P_{k-3})^3 \right. \\ \left. \text{by } q \circ F_K = (1 / \det(B_K)) B_K \hat{q} \right\}.$$

Then  $\sum_k = M_e(u) \cup M_f(u) \cup M_K(u)$ .

These degrees of freedom are affine invariant under the transformation from  $K$  to  $\hat{K}$ .

**Lemma 5.34** Suppose  $\det(B_K) > 0$  and the tangent vectors  $\tau$  on the edges of  $K$  are related to those on  $\hat{K}$  by (5.35). Then each of the sets of degrees of freedom (5.36)-(5.38) for  $u$  on  $K$  are identical to the degrees of freedom for  $\hat{u}$  on  $\hat{K}$  (provided(5.33)is used).

**Proof** Using the change of variables (5.33), and canceling the determinant terms gives

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

Next we consider facial degrees of freedom (5.37). Then using the change of variables (5.33) we can write

$$\frac{1}{\text{area}(\hat{f})} \int_{\hat{f}} \hat{u} \cdot \hat{q} d\hat{A} = \frac{1}{\text{area}(f)} \int_f (B_K^T \hat{u}) \cdot (B_K^{-1} \hat{q}) \frac{\text{area}(\hat{f})}{\text{area}(f)} dA \\ = \frac{1}{\text{area}(f)} \int_f u \cdot q dA.$$

To prove the result for edge degrees of freedom (5.36), we use (5.35) to show that, provided (5.33) is used,

$$\int_e u \cdot \tau q ds = \int_{\hat{e}} (B_K^{-T} \hat{u}) \cdot \hat{\tau} \hat{q} \frac{\text{length}(e)}{\text{length}(\hat{e})} d\hat{s} = \int_{\hat{e}} \hat{u} \cdot (B_K^{-1} \hat{\tau}) \hat{q} \frac{\text{length}(e)}{\text{length}(\hat{e})} d\hat{s}.$$

Now  $\tau = (b - a)/\text{length}(e)$ , where  $a$  and  $b$  are the end points of  $e$ . Hence

$$B_K^{-1} \tau = \frac{(a - b)}{\text{length}(e)} \frac{\text{length}(\hat{e})}{\text{length}(e)}.$$

Thus

$$\int_e u \cdot \tau q ds = \int_{\hat{e}} \hat{u} \cdot \hat{\tau} \hat{q} d\hat{s}.$$

□

We need to establish that the element is curl conforming. Using Theorem 5.3 we know that it suffices to prove that across every face separating two elements in the mesh  $\tau_i$ , the tangential trace  $u \times v$  of a finite element is continuous across the face. Since the degrees of freedom of type (5.37) agree across the common face, and for each edge of the face the degrees of freedom of type (5.36) also agree, we know that the degrees of freedom for the difference of the functions on each element all vanish. Thus it suffices to prove the following lemma (see the proof of Theorem 5.23 for more details of this type of argument in the context of divergence conforming elements). We shall prove the lemma for general  $k \geq 3$ . For  $k < 3$  the proof terminates at an earlier stage.

**Lemma 5.35** *If  $u \in R_k$  is such that the degrees of freedom of type (5.37) vanish on a given face  $f$  and such that the degrees of freedom of type (5.36) vanish for each edge of  $f$  then  $u \times v = 0$  on  $f$ .*

**Proof** It is convenient to prove this result on the reference element, which is possible since Lemma 5.34 guarantees that all degrees of freedom associated with  $f$  (including edges) vanish simultaneously with those associated with  $\hat{f}$ . We can also assume that  $\hat{f}$  is the face of  $\hat{K}$  in the plane  $\hat{x}_3 = 0$ . Then the tangential component of  $u$  is

$$\hat{u}_T = (\hat{u}_1(\hat{x}_1, \hat{x}_2, 0), \hat{u}_2(\hat{x}_1, \hat{x}_2, 0), 0)^T$$

and by Stokes Theorem 3.21, for every  $\hat{q} \in P_{k-1}(\hat{f})$ ,

$$\int_{\hat{f}} \vec{\nabla}_{\hat{f}} \times \hat{u}_T \hat{q} d\hat{A} = \int_{\hat{f}} u_T \cdot \vec{\nabla}_{\hat{f}} \times \hat{q} d\hat{A} + \int_{\partial \hat{f}} \hat{u}_T \cdot \hat{\tau} \hat{q} ds_{\hat{e}}.$$

Since  $\hat{q} \in P_{k-1}(\hat{f})$ , we have  $\vec{\nabla}_{\hat{f}} \times \hat{q} \in (P_{k-2}(\hat{f}))^2$ , so the fact that the degrees of freedom of type (5.37) and (5.36) vanish on  $\hat{f}$  and  $\partial \hat{f}$  implies that  $\vec{\nabla}_{\hat{f}} \times \hat{u}_T = 0$  on  $\hat{f}$ . Then since  $\hat{u}_T \in R_k(\hat{f})$  (using the analogue of  $R_k$  in two dimensions; see Remark 5.29), the analogue of Lemma 5.28 in two dimensions shows that  $\hat{u}_T = \vec{\nabla}_{\hat{f}} \hat{p}$  for some  $\hat{p} \in P_k$ , where  $\vec{\nabla}_{\hat{f}}$  is the surface gradient on  $\hat{f}$ . But for each  $\hat{q} \in P_{k-1}(\hat{e})$ , using the vanishing of degrees of type (5.36) on each edge  $\hat{e}$ ,

$$0 = \int_{\hat{e}} \hat{u}_T \cdot \hat{\tau} \hat{q} ds = \int_{\hat{e}} \frac{\partial \hat{p}}{\partial \hat{s}} \hat{q} ds.$$

Choosing  $q = \partial \hat{p} / \partial s$  shows that  $\partial \hat{p} / \partial s = 0$  on each edge  $\hat{e}$  of  $\hat{f}$  and we can choose  $\hat{p} = 0$  on  $\partial \hat{f}$ . Then using the geometry of  $\hat{f}$  we see that,  $\hat{p} = \hat{x}_1 \hat{x}_2 (1 - \hat{x}_1 - \hat{x}_2) r$

for some  $r \in P_{k-3}(\hat{f})$ , and so integration by parts shows that, for a fixed tangent vector  $\hat{\tau}$ ,

$$\int_{\hat{f}} \widehat{\nabla}_{\hat{f}} \hat{p} \cdot \hat{\tau} \hat{q} d\hat{A} = - \int_{\hat{f}} \hat{p} \hat{\tau} \cdot \widehat{\nabla}_{\hat{f}} \hat{q} d\hat{A}.$$

We now select  $\hat{q}$  to satisfy  $\hat{\tau} \cdot \widehat{\nabla}_{\hat{f}} \hat{q} = r$  and, since  $\hat{\tau}$  is constant, we can assume that  $\hat{q} \in P_{k-1}(\hat{f})$ . Thus, using the face degrees of freedom,

$$\int_{\hat{f}} \hat{x}_1 \hat{x}_2 (1 - \hat{x}_1 - \hat{x}_2) r^2 dA = 0$$

and we can conclude that  $r = 0$ , which completes the proof.  $\square$

Since the number of degrees of freedom and the dimension of  $R_k$  are the same, the next result proves unisolvence of the element.

**Lemma 5.36** Suppose  $u \in R_k$  is such that all its degrees of freedom (5.36)–(5.38) vanish. Then  $u = 0$ .

**Proof** We first map to the reference element, where all degrees of freedom also vanish. By the previous lemma,  $\hat{u} \times \hat{v} = 0$  on  $\partial\hat{K}$ . Then the integration by parts result (3.51) together with the vanishing of the moments (5.38) show that (5.39)

$$\int_{\hat{K}} \widehat{\nabla} \times \hat{u} \cdot \hat{q} d\hat{V} = \int_{\hat{K}} \hat{u} \cdot \widehat{\nabla} \times \hat{q} d\hat{V} = 0 \text{ for all } \hat{q} \in (P_{k-2})^3.$$

Using Corollary 3.21 (Stokes theorem in the plane) on each face  $\hat{f}$  of  $\hat{K}$  together with the vanishing of the moments (5.37) shows that

$$\int_{\hat{f}} \widehat{\nabla}_{\hat{f}} \times \hat{u}_T \hat{q} d\hat{A} = \int_{\hat{f}} u_T \cdot \vec{\nabla}_{\hat{f}} \times \hat{q} d\hat{A} = 0 \text{ for all } \hat{q} \in P_{k-1}(f).$$

Hence  $\nabla_{\hat{f}} \times \hat{u}_T = (\nabla \times \hat{u}) \cdot \hat{v} = 0$  on  $\hat{f}$  and hence on  $\partial\hat{K}$ .

Since  $\nabla \times \hat{u} \in (P_{k-1})^3$ , we know that

$$\widehat{\nabla} \times \hat{u} = (\hat{x}_1 \psi_1, \hat{x}_2 \psi_2, \hat{x}_3 \psi_3)^T,$$

where  $\psi = (\psi_1, \psi_2, \psi_3)^T \in (P_{k-2})^3$  and so picking  $q = \psi$  in (5.39) shows that  $\nabla \times \hat{u} = 0$  in  $\hat{K}$ . By Lemma 5.28,  $\hat{u} = \nabla \hat{p}$  for some  $\hat{p} \in P_k$ . But the fact that  $\hat{u} \times \hat{v} = 0$  on  $\partial\hat{K}$  implies that we can take  $\hat{p} = 0$  on  $\partial\hat{K}$  and so

$$\hat{p} = \hat{x}_1 \hat{x}_2 \hat{x}_3 \hat{r} \text{ for some } \hat{r} \in P_{k-3}.$$

The fact that the degrees of freedom (5.38) vanish implies that  $\hat{r} = 0$ , and the lemma is proved.  $\square$

To summarize the situation so far, we have proved the following theorem:

**Theorem 5.37** *The finite element defined in Definition 5.33 is  $H(\text{curl}; \Omega)$  conforming and unisolvent.*

As a result of this theorem, we can characterize the global finite element space on a mesh  $\tau_b$  by(5.40)

$$V_h = \{u \in H(\text{curl}; \Omega) | u|_K \in R_k \text{ for all } K \in \mathcal{T}_h\} .$$

Using the degrees of freedom (5.36)–(5.38), we can define an interpolant for  $V_b$ . Assuming that  $u$  has the necessary smoothness, the element-wise interpolant denoted by  $r_k u \in R_k$ , where  $K \in \tau_b$ , is characterized by the vanishing of the degrees of freedom on  $u - r_k u$ : (5.41)

$$M_e(u - r_k u) = M_f(u - r_k u) = M_K(u - r_k u) = \{0\} .$$

The global interpolant  $r_b u \in V_b$  is then defined element by element using

$$r_h u|_K = r_k u \text{ for all } K \in \mathcal{T}_h .$$

Unfortunately, because of the degrees of freedom (5.36) the interpolant is not defined for a general function in  $H(\text{curl}; \Omega)$ . To date, the best characterization of the functions for which the interpolant is defined is from [12]. We give a slightly simplified version.

**Lemma 5.38** *Suppose there are constants  $\delta > 0$  and  $p > 2$  such that  $u \in (H^{1/2+\delta}(K))^3$  and such that  $\nabla \times u \in (L^p(K))^3$  for each  $K$  in  $\tau_b$ . Then  $r_b u$  is well-defined and bounded.*

**Remark 5.39** *In fact, [12] gives an even weaker characterization requiring only that  $\nabla \times u \in (L^p(\Omega))^3$ ,  $u \in (L^p(K))^3$  and  $u \times v \in (L^p(\partial K))^3$  for some  $p > 2$ . This is implied by the result above, which is what we need for our analysis.*

**Proof of Lemma 5.38** We can assume that  $0 < \delta < 3/2$ . It is clear that (5.37) and (5.38) are well defined since  $u \in (L^2(K))^3$  and, by the Trace Theorem 3.9, we have  $u \in (H^0(\partial K))^3 \subset (L^p(\partial K))^3$  for any  $2 \leq p \leq 6/(3 - 2\delta)$  (here we have also used Theorem 3.7).

It remains to show that (5.36) is well defined. We select  $p$  such that  $2 < p \leq 6/(3 - 2\delta)$ . Let  $p'$  be such that  $1/p + 1/p' = 1$ . Given a polynomial  $q \in P_{k+1}(\ell)$  we extend it by zero to a function, still denoted by  $q$  on  $\partial f$ , where  $f$  is a face containing  $\ell$  on its boundary. Such a function is in  $W^{1-1/p, p'}(\partial f)$  since  $1-1/p' < 1/2$ .

Now using the right inverse of the trace map (see Theorem 3.9), we see that  $q$  is the trace of a function (again denoted  $q$ ) in  $W^{1, p'}(\ell)$ . Then (see Corollary 3.21, using a density argument and noting that since  $u \in H(\text{curl}; K)$  we know  $\nabla_j \times u = \nabla \times u \cdot v \in H^{-1/2}(\partial K)$ ), we conclude that

$$\int_\ell u \cdot \tau q \, ds = \int_f \nabla_f \times u \cdot q \, dA - \int_f u \cdot \vec{\nabla}_f \times q \, dA$$

Extending  $q$  by zero to all of  $\partial K$  and finally extending  $q \in W^{1,1/p, p}(\partial K)$  to a function in  $q \in W^{1, p'}(K)$ , and using (3.23) extended by density to functions in  $W^{1, p'}(K)$ , we have

$$\int_e u \cdot \tau q ds = \int_K \nabla \times u \cdot \nabla q dV - \int_f u \cdot \vec{\nabla}_f \times q dA.$$

Thus, using the boundedness of the extension operator in each step, we have

$$\begin{aligned} \left| \int_e u \cdot \tau q ds \right| &\leq C \left( \| \nabla \times u \|_{(Lp(K))^3} \| \nabla q \|_{(Lp'(K))^3} \right. \\ &\quad \left. + \| u \|_{(Lp(\partial K))^3} \| \vec{\nabla}_f \times q \|_{(Lp'(\partial K))^3} \right) \\ &\leq C \left( \| \nabla \times u \|_{(Lp(\partial K))^3} \| u \|_{(H^{1/2+\delta}(K))^3} \right) \| q \|_{W^{1-1/p', p'}(e)}. \end{aligned}$$

This proves the boundedness of the edge degrees of freedom for functions with the stated smoothness and completes the proof.  $\square$

We can now proceed to analyze the properties of the interpolant. Our first result establishes a link between the curl conforming elements of this section and the divergence conforming element of first type defined in Definition 5.14.

**Lemma 5.40** Suppose  $W_b$  is the divergence conforming finite element space in (5.28) and  $V_b$  is the curl conforming space given by (5.40). Then  $\nabla \times V_b \subset W_b$ , and if  $u$  is a function such that both the interpolants  $r_b u$  and  $w_b(\nabla \times u)$  are defined, then  $w_b(\nabla \times u) = \nabla \times r_b u$ .

**Proof** Let  $u_b \in V_b$ . Then  $\nabla \times u_b|_K \in (P_{k-1})^3$  for each  $K \in \tau_b$ . Furthermore,  $\nabla \cdot (\nabla \times u_b) = 0$ . Thus  $\nabla \times u_b \in H(\text{div}; \Omega)$  and so by characterization (5.28) we see that  $\nabla \times u_b \in W_b$ . It follows that  $\nabla \times r_b u \in W_b$ , and so it suffices to prove that the divergence element degrees of freedom for  $w_b(\nabla \times u)$  and  $\nabla \times r_b u$  agree element by element. But, on a given element  $K \in \tau_b$  with face  $f$ , for every  $q \in P_{k-1}(f)$ , first using the face degrees of freedom (5.22) for the divergence element, and then using the edge and face degrees of freedom for the curl element, we derive

$$\begin{aligned} &\int_f (\nabla \times r_b u - w_b(\nabla \times u)) \cdot v q dA \\ &= \int_f (\nabla \times r_b u - \nabla \times u) \cdot v q dA \\ &= \int_{\partial f} u \cdot \tau q ds + \int_f [u_T \cdot \vec{\nabla}_f \times q - (\nabla \times u) \cdot v] q dA \\ &= \int_f (\nabla \times u - \nabla \times u) \cdot v q dA = 0. \end{aligned}$$

In deriving the above equality, we have used the integration by parts result (3.28).

For the internal degrees of freedom (5.23), we see that for every  $q \in (P_{k+2})^3$ , using the internal divergence element degrees of freedom, followed by the surface and internal degrees of freedom (5.37) and (5.38) for the curl conforming element, that

$$\begin{aligned} & \int_K [\nabla \times r_h u - w_h \nabla \times u] \cdot q dV \\ &= \int_K r_h u \cdot \nabla \times q dV + \int_{\partial K} (v \times r_h u) \cdot q dA - \int_K \nabla \times u \cdot q dV \\ &= \int_K u \cdot \nabla \times q dV + \int_{\partial K} v \times u \cdot q dA - \int_K \nabla \times u \cdot q dV \\ &= \int_K \nabla \times u \cdot q dV - \int_K \nabla \times u \cdot q dV = 0. \end{aligned}$$

Here we have used the integral identity (3.27).  $\square$

It follows from the previous lemma that we can expect good approximation properties for the curl of the interpolant since we have already proved good error estimates for the divergence conforming interpolant  $w_h$ .

Our next result provides error estimates for the interpolant.

**Theorem 5.41** *Let  $\tau_b$  be a regular mesh on  $\Omega$ . Then*

(1) *If  $u \in (H(\Omega))^3$  and  $\nabla \times u \in (H(\Omega))^3$  for  $1/2 + \delta \leq s \leq k$  for  $\delta > 0$  then (5.42)*

$$\begin{aligned} \|u - r_h u\|_{(L^2(\Omega))^3} + \|\nabla \times (u - r_h u)\|_{(L^2(\Omega))^3} \\ \leq Ch^s \left( \|u\|_{(H^s(\Omega))^3} + \|\nabla \times u\|_{(H^s(\Omega))^3} \right). \end{aligned}$$

(2) *If  $u \in (H^{1/2+}(K))^3$ ,  $0 < \delta \leq 1/2$  and  $\nabla \times u|_K \in D_k$  where  $D_k$  is defined in (5.17), then*

$$\begin{aligned} \|u - r_h u\|_{(L^2(K))^3} &\leq C \left( h_K^{1/2+\delta} \|u\|_{(H^{1/2+\delta}(K))^3} \right. \\ &\quad \left. + h_K \|\nabla \times u\|_{(L^2(K))^3} \right). \end{aligned} \tag{5.43}$$

**Remark 5.42** *The first estimate holds tetrahedron by tetrahedron (i.e. with  $\Omega$  replaced by  $K$  and  $b$  replaced by  $b_K$ ). This is because it is proved element by element. Estimate (5.42) was proved in its current form by Alonso and Valli [9, 77]. Previous estimates along this line include the original paper of Nédélec [233], Dubois [132] and myself [216]. Estimate (5.43) is a generalization of the estimate (2.4) in [18] and is proved in [167] for Lipschitz polyhedral domains. This estimate is vital to our later proofs of convergence of the finite element method.*

*Note that a result of the proof of this theorem is the estimate*

$$\|\nabla \times (u - r_h u)\|_{(L^2(\Omega))^3} \leq Ch^s \|\nabla \times u\|_{(H^s(\Omega))^3},$$

*for  $1/2 + \delta \leq s \leq k$  and  $\delta > 0$ .*

Before we can prove this estimate, we summarize how Sobolev norms of curl conforming functions (i.e. those transformed by (5.33)) change as the functions are mapped.

**Lemma 5.43** Suppose the mesh  $\tau_b$  is regular and  $s \geq 0$ . Then there is a constant  $C$  independent of  $v$  such that if  $v$  is transformed by (5.33) to obtain  $v_{\text{circ}}$ ; then

$$\begin{aligned} |\hat{v}|_{(H^s(\hat{K}))^3} &\leq Ch_K^{s-1/2} |v|_{(H^s(K))^3}, \\ |\widehat{\nabla} \times \hat{v}|_{(H^s(\hat{K}))^3} &\leq Ch_K^{s+1/2} |\nabla \times v|_{(H^s(K))^3}. \end{aligned}$$

**Proof** For non integer  $s$ , the proof is given in [9]. We shall confine ourselves to proving the first result in the case of integer  $s$ . Using the fact that  $\hat{v} = B_K^T v \circ F_K$  we see that for any multi-index  $\alpha$

$$\frac{\partial^\alpha \hat{v}}{\partial \hat{x}^\alpha} = B_K^\top \frac{\partial^\alpha}{\partial x^\alpha} (v \circ F_K).$$

Thus the only difference between this and the classical mapping result in Lemma 5.9 is the presence of the matrix  $B_K$ . Using Lemma 5.9 we see that if  $|\alpha|_1 = s$  then

$$\begin{aligned} \left\| \frac{\partial^\alpha \hat{v}}{\partial \hat{x}^\alpha} \right\|_{(L^2(\hat{K}))^3} &\leq C |B_K| |B_K|^s |\det(B_K)|^{-1/2} \left\| \frac{\partial^\alpha v}{\partial x^\alpha} \right\|_{(L^2(K))^3} \\ &\leq Ch_K^{s-1/2} \left\| \frac{\partial^\alpha v}{\partial x^\alpha} \right\|_{(L^2(K))^3} \end{aligned}$$

and adding over all multi-indices  $|\alpha|_1 = s$  proves the result. For the curl estimates, we use transformation (5.34) in the same way.  $\square$

**Proof of Theorem 5.41** We shall only prove the result for integer  $s$ , to avoid technical complications. Thus we assume  $s$  is an integer with  $1 \leq s \leq k$  and the proof reduces to that in [233]. See [9] for more details particularly when  $1/2 + \delta \leq s < 1$ . We start by estimating the  $(L^2(\Omega))^3$  portion of (5.42) by decomposing the desired estimate into element-wise contributions:

$$\|u - r_h u\|_{(L^2(\Omega))^3}^2 = \sum_{K \in \tau_h} \|u - r_h u\|_{(L^2(K))^3}^2.$$

Next, for each element  $K$ , we transform to the reference element  $\hat{K}$  using (5.33):

$$\begin{aligned} \|u - r_h u\|_{(L^2(K))^3}^2 &\leq |\det(B_K)|^{1/2} \|B_K^{-1}\| \|\widehat{u - r_h u}\|_{(L^2(\hat{K}))^3} \\ &\leq Ch_K^{1/2} \|\widehat{u - r_h u}\|_{(L^2(\hat{K}))^3}. \end{aligned}$$

But, because of Lemmas 5.32 and 5.34, we have that  $\widehat{r_h u} = r_{\hat{K}} \widehat{u}$ . So, using the fact that  $(P_{k-1})^3 \subset R_k$ ,

$$\|\hat{u} - r_{\hat{K}} \hat{u}\|_{(L^2(\hat{K}))^3}^2 = \|(I - r_{\hat{K}})(\hat{u} + \phi)\|_{(L^2(\hat{K}))^3} \quad (5.44)$$

for every  $\phi \in (P_{k-1})^3$ . But the degrees of freedom of  $u$ , given by (5.36)–(5.38), may be estimated using Lemma 5.38 so that:

$$\|(I - r_{\hat{K}})(\hat{u} + \phi)\|_{(L^2(\hat{K}))^3} \leq C \left( \|\hat{u} + \phi\|_{(H^s(\hat{K}))^3} + \|\widehat{\nabla} \times (\hat{u} + \phi)\|_{(H^s(\hat{K}))^3} \right).$$

Now, using Theorem 5.5, we have (using the fact that  $s$  is an integer) (5.45)

$$\begin{aligned} & \inf_{\phi \in (P_{k-1})^3} \|(I - r_{\hat{K}})(\hat{u} + \phi)\|_{(L^2(\hat{K}))^3} \\ & \leq C \left( |\hat{u}|_{(H^s(\hat{K}))^3} + |\widehat{\nabla} \times \hat{u}|_{(H^s(\hat{K}))^3} \right). \end{aligned}$$

Mapping back to the reference element (using Lemma 5.43) shows that, for any integer

$$s > 1/2, |\hat{u}|_{(H^s(\hat{K}))^3} \leq Ch_K^{s-1/2} |u|_{(H^s(K))^3}$$

and  $|\widehat{\nabla} \times \hat{u}|_{(H^s(\hat{K}))^3} \leq Ch_K^{s+1/2} |\nabla \times u|_{(H^s(K))^3}$ . Using this estimate in (5.45) and then using (5.45) in (5.44) shows that (5.46)

$$\|u - r_K u\|_{(L^2(K))^3} \leq Ch_K^s \left( |u|_{(H^s(K))^3} + |\nabla \times u|_{(H^s(K))^3} \right).$$

In fact, for integer  $s > 1$  this estimate can be proved without the need for the curl term.

To prove the curl estimate, we can use Lemma 5.40 and Theorem 5.25:

$$\|\nabla \times (u - r_K u)\|_{(L^2(\Omega))^3} \leq \|(I - w_h) \nabla \times u\|_{(L^2(\Omega))^3} \leq Ch^s \|\nabla \times u\|_{(H^s(\Omega))^3}.$$

Combining this with (5.46) proves the  $H(\text{curl}; \Omega)$  estimate.

The proof of (5.43) follows the same pattern. Mapping to the reference element and using Lemma 5.38, we see that all the degrees of freedom of  $r_K \hat{u}$  can be estimated by  $\|\hat{u}\|_{(H^{1/2+\delta}(\hat{K}))^3} + \|\nabla \times \hat{u}\|_{(L(\hat{K}))^3}$ ,  $p > 2$ . But, since  $\nabla \times \hat{u} \in D_p$ , the equivalence of norms on a finite-dimensional vector space shows that  $r_K \hat{u}$  can be estimated by  $\|\hat{u}\|_{(H^{1/2+\delta}(\hat{K}))^3} + \|\nabla \times \hat{u}\|_{(L(\hat{K}))^3}$ . Thus we have

$$\|\hat{u} - r_{\hat{K}} \hat{u}\|_{(L^2(\hat{K}))^3} \leq C \inf_{\phi \in (P_0)^3} \left( \|\hat{u} + \phi\|_{(H^{1/2+\delta}(\hat{K}))^3} + \|\widehat{\nabla} \times \hat{u}\|_{(L^2(\hat{K}))^3} \right),$$

where  $\phi$  vanishes from the curl part of the estimate because it is a constant function. Thus, by Theorem 5.5, we have

$$\|\hat{u} - r_{\hat{K}} \hat{u}\|_{(L^2(\hat{K}))^3} \leq C \left( |\hat{u}|_{(H^{1/2+\delta}(\hat{K}))^3} + \|\widehat{\nabla} \times \hat{u}\|_{(L^2(\hat{K}))^3} \right).$$

Now mapping back to the reference element using Lemma 5.43 completes the estimate.  $\square$

We have seen how to construct subspaces  $V_b \subset H(\text{curl}; \Omega)$  but we also need to construct subspaces of  $X$ . Fortunately, this is easy, as the following lemma shows.

**Lemma 5.44** Suppose  $u \in X$  and  $r_b u$  is well defined. Then  $r_b u \in X$ .

**Remark 5.45** This implies that it is sufficient to set the degrees of freedom associated with edges and face on  $\Gamma$  to zero in order to construct finite element functions in  $X$ . Our proof uses the fact that  $\Gamma$  and  $\sum$  are disjoint and covered completely by triangulations. Thus we can define

$$X_h = \left\{ u_h \in H(\text{curl}; \Omega) \mid u_h|_K \in R_k \text{ for all } K \in \tau_h, u_h \times v = 0 \text{ on } \Gamma \right\}$$

and for any suitably smooth function  $u$  in  $X$  we know that  $r_b u \in X_b$  so the error estimates of Theorem 5.41 hold for interpolation in  $X$  as well.

**Proof of Lemma 5.44** Let  $f$  be a face of the mesh on  $\Gamma$ . The degrees of freedom (5.36) and (5.37) associated with  $f$  only depend on the tangential components of  $u$  on  $f$  which vanish. Hence all degrees of freedom associated with  $f$  vanish and so, by Lemma 5.35,  $r_b u \times v = 0$  on  $f$ .  $\square$

### 5.5.1 Linear edge element

Because of their importance in practice, let us consider the linear elements ( $k = 1$ ) in more detail. For  $k = 1$ , Nédélec [233] shows that

$$R_1 = \left\{ u(x) = a + b \times x, \text{ where } a, b \in \mathbb{C}^3 \right\}.$$

The six constants (the components of  $a$  and  $b$ ) are determined from the moments  $\int_e u \cdot \tau \, ds$  on the six edges,  $e$ , of an element  $K$ . It is for this reason that such elements (and by extension the entire family) are called *edge elements* (see Fig. 5.7).

Direct computation shows that the basis function with unit integral on the edge joining vertices  $i$  and  $j$  is given by (5.47)

$$\psi_{i,j} = \lambda_i \nabla \lambda_j - \lambda_j \nabla \lambda_i,$$

where  $\lambda_i$  is the barycentric coordinate function corresponding to node  $a_i$ . The function  $\psi_{i,j}$  also arises in the work of Whitney [295] and so the  $k = 1$  element is sometimes called the *Whitney element* (see [53, 54, 164] for a more extensive discussion of this aspect of edge elements). It is easy to see, using mid-point quadrature on the edge joining vertices  $i$  and  $j$  and the fact that this quadrature is exact for quadratic polynomials,  $\int_e \psi_{i,j} \cdot \tau \, ds = 1$ . We also remark that (5.48)

$$\nabla \times \psi_{i,j} = 2 \nabla \lambda_i \times \nabla \lambda_j.$$

The Whitney representation of the basis function is a convenient way of programming the lowest-order edge element.

Note that, on each element  $K \in \tau_b$ , a simple calculation shows that  $\nabla \cdot (\psi_{i,j}|_K) = 0$ . But this does not imply that  $\psi_{i,j}$  is globally divergence free. Since the normal component of  $\psi_{i,j}$  is not continuous across faces in the mesh there is a singular contribution to the divergence at the faces in the mesh.

### 5.5.2 Quadratic edge elements

Now we shall give details of how to compute a basis for quadratic edge elements  $k = 2$  on a tetrahedron. The problem is to find a basis that is easily constructed and which yields conformity of the global finite element space without elaborate modifications. This is relatively easy because of the degrees of freedom defined previously in this chapter, once we have agreed on a global orientation for the various geometric elements (edges and faces) in the mesh.

A straightforward computation using MAPLE (or similar algebraic software — or even by hand!) shows that if we define the vectors  $p_j, j = 1, 2, 3$ , as follows

$$p_1 = \begin{pmatrix} 0 \\ -\hat{x}_3 \\ \hat{x}_2 \end{pmatrix}, \quad p_2 = \begin{pmatrix} \hat{x}_3 \\ 0 \\ -\hat{x}_1 \end{pmatrix}, \quad p_3 = \begin{pmatrix} -\hat{x}_2 \\ \hat{x}_1 \\ 0 \end{pmatrix},$$

then  $R_2$  is spanned by the following twenty vectors:

$$\begin{aligned} \hat{v}_1 &= e_1, & \hat{v}_2 &= e_2, & \hat{v}_3 &= e_3, & \hat{v}_4 &= \hat{x}_1 e_1 \\ \hat{v}_5 &= \hat{x}_1 e_2, & \hat{v}_6 &= \hat{x}_1 e_3, & \hat{v}_7 &= \hat{x}_2 e_1, & \hat{v}_8 &= \hat{x}_2 e_2, \\ \hat{v}_9 &= \hat{x}_2 e_3, & \hat{v}_{10} &= \hat{x}_3 e_1, & \hat{v}_{11} &= \hat{x}_3 e_2, & \hat{v}_{12} &= \hat{x}_3 e_3, \\ \hat{v}_{13} &= \hat{x}_1 p_1, & \hat{v}_{14} &= \hat{x}_2 p_1, & \hat{v}_{15} &= \hat{x}_3 p_1, & \hat{v}_{16} &= \hat{x}_1 p_2, \\ \hat{v}_{17} &= \hat{x}_2 p_2, & \hat{v}_{18} &= \hat{x}_3 p_2, & \hat{v}_{19} &= \hat{x}_1 p_3, & \hat{v}_{20} &= \hat{x}_2 p_3, \end{aligned}$$

where  $e_j, j = 1, 2, 3$ , are the usual Cartesian unit vectors. Thus any  $\hat{u} \in R_2$  can be written as  $\hat{u} = \sum_{i=1}^{20} \hat{u}_i \hat{v}_i$ . We now wish to compute the basis functions corresponding to the affine invariant degrees of freedom in Definition 5.30. As usual our reference tetrahedron has vertices

$$\hat{a}_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \hat{a}_2 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \hat{a}_3 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \hat{a}_4 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

and we label the edges and faces as in Fig. 5.8 .

To each edge  $[\hat{a}_i, \hat{a}_j], i < j$  we assign a unit tangent vector

$$\hat{t}_{ij} = \frac{\hat{a}_j - \hat{a}_i}{\|\hat{a}_j - \hat{a}_i\|}.$$

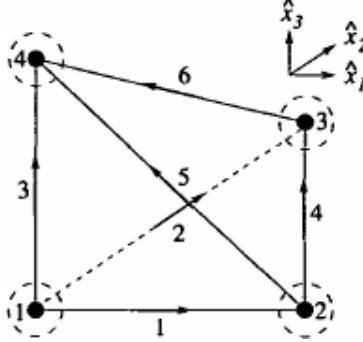
The edge degrees of freedom on the reference element are thus

$$\int_{\hat{a}_i}^{\hat{a}_j} \hat{u} \cdot \hat{t}_{ij} q ds$$

for  $\hat{q} \in P_1([\hat{a}_i, \hat{a}_j])$ . Since  $\hat{u} \cdot \hat{t}_{ij} \hat{q} \in P_2(\hat{e})$ , we can compute the above integral by Gaussian quadrature

$$\int_{\hat{a}_i}^{\hat{a}_j} \hat{u} \cdot \hat{t}_{ij} q ds = \frac{1}{2} \left( \hat{u} \left( \hat{\xi}_{ij}^{(1)} \right) \hat{q} \left( \hat{s}_{ij}^{(1)} \right) + \hat{u} \left( \hat{\xi}_{ij}^{(2)} \right) \hat{q} \left( \hat{s}_{ij}^{(2)} \right) \right) \cdot (\hat{a}_j - \hat{a}_i),$$

Fig. 5.8. Labeling of the edges on the reference tetrahedron with the direction of the edge (the direction of the associated tangent vector) marked. Vertex numbers are surrounded by a dashed circle. Faces are triangles with oriented via the right-hand rule: face 1 is the triangle [1,2,3] (i.e. the triangle with vertices numbered 1,2 and 3) with normal vector in the direction  $(\hat{a}_2 - \hat{a}_1) \times (\hat{a}_3 - \hat{a}_1)$ , face 2 is the triangle [1,2,4], face 3 is the triangle [1,3,4] and face 4 is the triangle [2,3,4].



where  $\hat{s}_{ij}^{(1)}$  and  $\hat{s}_{ij}^{(2)}$  are the two Gaussian quadrature points in arc length along  $[\hat{a}_i, \hat{a}_j]$  with  $\hat{s}_{ij}^{(1)}$  closest to  $\hat{a}_i$  and  $\hat{s}_{ij}^{(2)}$ ,  $i = 1, 2$ , are the corresponding three-dimensional coordinate of the Gauss point on  $[\hat{a}_i, \hat{a}_j]$ . Now we choose the linear polynomial  $\hat{q}$  so that  $\hat{q}(\hat{s}_{ij}^{(1)}) = 1$ ,  $\hat{q}(\hat{s}_{ij}^{(2)}) = 0$ , (and vice versa), so the corresponding degrees of freedom are just(5.49)

$$\hat{u}\left(\hat{\xi}_{ij}^{(1)}\right) \cdot (\hat{a}_j - \hat{a}_i) \quad \text{and} \quad \hat{u}\left(\hat{\xi}_{ij}^{(2)}\right) \cdot (\hat{a}_j - \hat{a}_i) .$$

For each face  $\hat{f}$  with vertices  $\{\hat{a}_i, \hat{a}_j, \hat{a}_k\}$ ,  $i < j < k$ , we choose (note these are *not* unit tangent vectors)  $\hat{t}_1^{(\hat{f})} = (\hat{a}_j - \hat{a}_i)$  and  $\hat{t}_2^{(\hat{f})} = (\hat{a}_k - \hat{a}_i)$ . Then the degrees of freedom associated with  $\hat{f}$  are

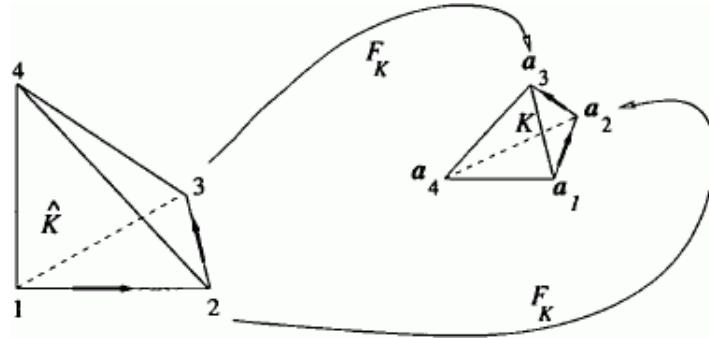
$$\frac{1}{\text{area}(\hat{f})} \int_{\hat{f}} \hat{u} \cdot \hat{t}_1^{(\hat{f})} dA \quad \text{and} \quad \frac{1}{\text{area}(\hat{f})} \int_{\hat{f}} \hat{u} \cdot \hat{t}_2^{(\hat{f})} dA .$$

For the purpose of computing the basis elements, it is convenient to compute these integrals by quadrature at the midpoints of the edges of the tetrahedron, so since  $\hat{u}$  is at most quadratic,(5.50)

$$\begin{aligned} \frac{1}{\text{area}(\hat{f})} \int_{\hat{f}} \hat{u} \cdot \hat{t}_1^{(\hat{f})} dA = & \quad \frac{1}{3} \left( \hat{u}\left(\frac{\hat{a}_i + \hat{a}_j}{2}\right) + \hat{u}\left(\frac{\hat{a}_i + \hat{a}_k}{2}\right) \right. \\ & \left. + \hat{u}\left(\frac{\hat{a}_j + \hat{a}_k}{2}\right) \right) \cdot \hat{t}_l^{(\hat{f})}, \quad l = 1, 2. \end{aligned}$$

Using these degrees of freedom, we can now compute the basis functions for  $R_2$ . For example, we could compute the basis function that has degree of freedom

Fig. 5.9. The choice of degrees of freedom and affine mapping guarantees that the direction of the tangent vectors agree across element boundaries.



1 at the first Gauss point on the edge  $[\hat{a}_1, \hat{a}_2]$  but zero at all other degrees of freedom by using the expansion  $\hat{u} = \sum_{j=1}^{20} \hat{u}_j \hat{v}_j$  in the following equations:

$$\begin{aligned}\hat{u}\left(\hat{\xi}_{1,2}^{(1)}\right) \cdot (\hat{a}_2 - \hat{a}_1) &= 1, \\ \hat{u}\left(\hat{\xi}_{1,2}^{(2)}\right) \cdot (\hat{a}_2 - \hat{a}_1) &= 0, \\ \hat{u}\left(\hat{\xi}_{1,m}^{(n)}\right) \cdot (\hat{a}_m - \hat{a}_1) &= 0, \quad l < m, \quad (l, m) \neq (1, 2), \quad 1 \leq l, m \leq 4, \\ \frac{1}{\text{area}(\hat{f}_k)} \int_{\hat{f}_k} \hat{u} \cdot \hat{t}_l^{(\hat{f}_k)} dA &= 0, \quad 1 \leq k \leq 4, \quad 1 \leq l, m \leq 4,\end{aligned}$$

where we use the discrete equations (5.50) to actually implement the area integral. Here  $\hat{f}_k$  is the  $k$ th face of the reference tetrahedron as defined in the caption to Fig. 5.8. Using this procedure, we can successively compute the coefficients of the 20 basis functions for  $K$  in terms of  $\hat{v}_1, \dots, \hat{v}_{20}$  once and for all. To obtain basis functions on any other element  $K$ , we need only transform to that element.

Our careful choice of tangent vector directions and degrees of freedom makes it possible to ensure that tangent directions are always followed consistently on the target elements—provided we follow the following simple rule. Suppose the vertices in the mesh are enumerated and the set of vertices is written  $\{a_i\}_{i=1}^{N_h}$ , where  $N_h$  is the total number of the vertices in the mesh. Then a tetrahedron  $K$  in the mesh will have vertices  $a_{i_1}, a_{i_2}, a_{i_3}, a_{i_4}$  and we assume that the vertices are ordered so that  $i_1 < i_2 < i_3 < i_4$ . We choose the affine map so as to map  $\hat{a}_l$  to  $a_{i_l}$ ,  $1 \leq l \leq 4$ . Thus, the component matrix and vector of the affine map are given by

$$b_K = a_{i_1} \quad \text{and} \quad B_K = (a_{i_2} - a_{i_1}, a_{i_3} - a_{i_1}, a_{i_4} - a_{i_1}).$$

In Fig. 5.9 we indicate how this forces the tangent directions to be consistent between the orientation of edges on the reference element  $K$  and the global orientation on  $K$ . To see this, suppose tetrahedra  $K_1$  and  $K_2$  meet at face  $f$

with vertices  $a_{j_1}, a_{j_2}, a_{j_3}$ ,  $j_1 < j_2 < j_3$ . Then, for example, the tangent vector will point from  $a_{j_1}$  to  $a_{j_2}$  along the appropriate edge, regardless of whether  $K_1$  or  $K_2$  is the target element. This trick is due to Paul Wesson.

Obviously, this procedure for computing basis functions can be extended (at the cost of some considerable labor) to higher-order edge elements. The key point is that our choice of vertex order for edges, faces and tetrahedra establishes an intrinsic orientation for each geometric quantity which unambiguously forces the degrees of freedom to agree.

## 5.6 $H^1(\Omega)$ conforming finite elements

We continue our description of the spaces of functions needed for Maxwell's equations by discussing the scalar space suitable for discretizing the potential  $p$  (or the space  $S$  given by (4.6)). This material is entirely classical, so we shall mainly give references to the appropriate literature. We start by defining the  $H^1$  conforming element:

**Definition 5.46** (*Scalar finite element space*) On a general tetrahedron  $K$ ,

- $K = \text{tetrahedron};$
- $P_K = P_k;$
- the degrees of freedom  $\sum_K$  fall into four classes (depending upon  $k$ , some may not be needed):
  - (1) vertex degrees of freedom: Let  $a_i$ ,  $1 \leq i \leq 4$  be the vertices of  $K$ , then(5.51)

$$M_v(p) = \{p(a_i), 1 \leq i \leq 4\},$$

(2) edge degrees of freedom:(5.52)

$$M_e(p) = \left\{ \frac{1}{\text{length}(e)} \int_e pq ds \text{ for all } q \in P_{k-2}(e), \right. \\ \left. \text{for all edges } e \right\},$$

(3) face degrees of freedom:(5.53)

$$M_f(p) = \left\{ \frac{1}{\text{area}(f)} \int_f pq dA \text{ for all } q \in P_{k-3}(f), \right. \\ \left. \text{for all faces } f \right\},$$

(4) volume degrees of freedom:(5.54)

$$M_K(p) = \left\{ \frac{1}{\text{volume}(K)} \int_K pq dV \text{ for all } q \in P_{k-4} \right\}.$$

Then  $\sum_k = M_e(p) \cup M_v(p) \cup M_f(p) \cup M_K(p)$ .

Of course, if  $k < 4$  then  $M_k(p)$  is not used, if  $k < 3$  then  $M_j(p)$  is not used, and if  $k < 2$  then  $M_e(p)$  is not used. We note that the total number of degrees of freedom is

$$\begin{aligned} 4 + 6\dim(P_{k-2}(e)) + 4\dim(P_{k-3}(f)) + \dim(P_{k-4}(K)) \\ = \frac{1}{6}(k+3)(k+2)(k+1), \end{aligned}$$

so that there are the same number of degrees of freedom as the dimensions of  $P_k$ .

**Lemma 5.47** *The element defined above is  $H^1(\Omega)$  conforming and unisolvant.*

**Proof** First we prove conformance in the general case  $k \geq 4$  (for  $k < 4$  the proof terminates early). Using Lemma 5.3 this reduces to showing that if all vertex, edge and face degrees of freedom of  $p \in P_k$  vanish for a particular face  $f$  of the tetrahedron, then  $p = 0$  on that face. Since the vertex degrees of freedom for  $p$  vanish, for each  $e \in \partial f$  we may write

$$\int_e \frac{\partial p}{\partial s} q ds = - \int_e p \frac{\partial q}{\partial s} ds = 0$$

for any  $q \in P_{k-1}(e)$ . Choosing  $q = \partial p / \partial s$  shows that  $\partial p / \partial s = 0$  along this edge and hence  $p = 0$  along each edge. Since  $p = 0$  on  $\partial f$ ,  $p = \lambda_1 \lambda_2 \lambda_3 r$ , where  $\lambda_1, \lambda_2, \lambda_3$  are the area barycentric coordinate functions for  $f$ . Then using the facts that  $r \in P_{k-3}(f)$  and that the face degrees of freedom vanish, we have

$$0 = \int_f p r dA = \int_f \lambda_1 \lambda_2 \lambda_3 r^2 dA.$$

Hence  $r = 0$ , and we have proved that  $p = 0$  on  $f$ .

Now we prove unisolvence. This reduces to showing that if all degrees of freedom of the polynomial  $p \in P_k$  vanish, then  $p = 0$ . But from the previous conformance calculation, we know that the vanishing degrees imply  $p = 0$  on  $\partial K$ . Thus,  $p = \lambda_1 \lambda_2 \lambda_3 \lambda_4 r$  for some  $r \in P_{k-4}$ , where  $\lambda_1, \dots, \lambda_4$  are the volume barycentric coordinate functions for  $K$ . Using the volume degrees of freedom (that vanish)

$$0 = \int_K p r dV = \int_K \lambda_1 \lambda_2 \lambda_3 \lambda_4 r^2 dV.$$

Hence  $r = 0$ , and we have proved that  $p = 0$ , as desired.  $\square$

For any  $p \in H^{3/2+\delta}(K)$ ,  $\delta > 0$ , we can now define an interpolation operator  $\pi_K$  by requiring that

$$M_v(p - \pi_K p) = M_e(p - \pi_K p) = M_f(p - \pi_K p) = M_K(p - \pi_K p, p) = \{0\}.$$

For piecewise linear functions, this just says that  $\pi_K p$  interpolates  $p$  at the vertices, but for  $k \geq 2$  the interpolation is via the moments or weighted integrals

along the edges, over face and volumes (besides the usual vertex interpolation). The restriction on the regularity of  $p$  allows us to use the Sobolev Embedding Theorem 3.5 to show that that  $H^{3/2+\delta}(K) \subset C(K)$  and hence to ensure vertex values are well defined.

The conformance lemma above also allows us to define the space(5.55)

$$U_h = \left\{ p_h \in H^1(\Omega) \mid p_h|_K \in P_k \text{ for all } K \in \mathcal{T}_h \right\},$$

which is just the standard space of piecewise  $k$ -degree, continuous piecewise polynomials. Via the local interpolation operators, we then define the global interpolation operator  $\pi_h : H^{3/2+\delta}(\Omega) \rightarrow U_h$ ,  $\delta > 0$ , by(5.56)

$$(\pi_h p)|_K = \pi_K p, \quad \text{for all } K \in \mathcal{T}_h.$$

The following theorem summarizes the accuracy properties of the interpolant.

**Theorem 5.48** *Let  $\tau_h$  be a regular family of meshes of  $\Omega$ . Then there exists a constant  $C$  independent of  $h$  and  $p$  such that*

$$\|p - \pi_h p\|_{H^1(\Omega)} \leq Ch^{s-1} \|p\|_{H^s(\Omega)}, \quad \frac{3}{2} + \partial \leq s \leq k + 1.$$

**Proof** For  $2 \leq s \leq k + 1$ , this theorem is entirely classical and may be found in [80]. For  $\frac{3}{2} + \partial \leq s \leq 2$  it may be proved, for example, by using the techniques from [9]. We shall not provide the proof here, but the usual technique of mapping to the reference element and using Theorem 5.5 suffices.  $\square$

Now we show that the scalar space  $U_h$  and curl conforming space  $V_h$  are connected in an intimate way.

**Theorem 5.49** *Suppose  $U_h$  is defined by(5.55) and  $V_h$  is defined by (5.40), then  $\nabla U_h \subset V_h$  and if  $p$  is sufficiently smooth that both  $\pi_h p$  and  $r_h \nabla p$  are well defined (e.g.  $p \in H^{3/2+\delta}(\Omega)$ ,  $\delta > 0$ ), we have*

$$\nabla \pi_h p = r_h \nabla p.$$

**Proof** This result is proved in [233]. First we see that if  $p_h \in U_h$ , then  $\nabla p_h \in H(\text{curl}; \Omega)$  and on each tetrahedron  $K$ ,  $(\nabla p_h)|_K \in (P_{k-1})^3 \subset R_k$ . Hence  $\nabla U_h \subset V_h$ .

To show that  $\nabla \pi_h p = r_h \nabla p$ , it then suffices to show that all the degrees of freedom for  $\nabla \pi_h p$  and  $r_h \nabla p$  agree tetrahedron by tetrahedron. We start with the edge degrees. For  $q \in P_{k-1}(e)$ ,

$$\int_e (\nabla \pi_h p - r_h \nabla p) \cdot \mathbf{n} q ds = - \int_e \left( \frac{\partial \pi_h p}{\partial s} - \frac{\partial p}{\partial s} \right) q ds,$$

where we have used the edge degrees (5.36) for  $r_h \nabla p$ . Now for  $q = 1$ , and if  $e = [a, b]$ ,

$$\int_e \left( \frac{\partial \pi_h p}{\partial s} - \frac{\partial p}{\partial s} \right) q ds = (\pi_h p(b) - p(b)) - (\pi_h p(a) - p(a)) = 0,$$

where we have used the vertex interpolation property of  $p$  from (5.52). For general  $q \in P_{k-1}(\ell)$ , integrating the right hand side of (5.57) by parts and using the vertex interpolation property and the degrees of freedom (5.52) of  $\pi_h$ , we obtain

$$\int_e (\nabla \pi_h p - r_h \nabla p) \cdot \tau q ds = - \int_e (\pi_h p - p) \frac{\partial q}{\partial s} ds = 0$$

since  $\partial q / \partial s \in P_{k-2}(\ell)$ . For the face degrees of freedom, using a constant tangent vector  $\tau$  and  $q \in P_{k-2}(f)$  we have

$$\int_f (\nabla \pi_h p - r_h \nabla p) \cdot \tau q dA = \int_f \frac{\partial}{\partial \tau} (\pi_h p - p) q dA,$$

where we have used the face degrees of freedom (5.37) for  $r_h$  to remove this operator. Using a general unit tangent vector  $\tau$ , we conclude that

$$\int_f (\nabla \pi_h p - r_h \nabla p) \cdot \tau q dA = 0$$

if and only if

$$\int_f \nabla_f (\pi_h p - p) \cdot \xi dA = 0 \quad \text{forall } \xi \in (P_{k-2}(f))^2.$$

Using the Divergence Theorem 3.19 in the plane containing  $f$ , we conclude that

$$\int_f \nabla_f (\pi_h p - p) \cdot \xi dA = - \int_f (\pi_h p - p) \nabla_f \cdot \xi dA + \int_{\partial f} (\pi_h p - p) v_f \cdot \xi ds,$$

where  $v_f$  is the outward normal to  $f$  in the plane of  $f$ . But  $\nabla_f \cdot \xi \in P_{k-3}(f)$  and  $v_f \cdot \xi \in P_{k-2}(\ell)$  for each edge, so the right-hand side vanishes using the face and edge degrees of freedom (5.53) and (5.52) for  $\pi_h$ . We have thus proved that the face degrees of freedom (5.37) for  $\nabla \pi_h p$  and  $r_h p$  agree.

Finally, for the volume degrees of freedom we use the degrees of freedom in (5.38) together with the integral identity (3.24) to show that if  $q \in (P_{k-3})^3$  then

$$\begin{aligned} & \int_K (\pi_h p - r_h \nabla p) \cdot q dV \\ &= \int_K \nabla (\pi_h p - p) \cdot q dV \\ &= - \int_K (\pi_h p - p) \nabla \cdot q dV + - \int_{\partial K} (\pi_h p - p) q \cdot v dA. \end{aligned}$$

But  $\nabla \cdot q \in P_{k-4}$  and  $q \cdot v \in P_{k-3}(f)$  for each face  $f$ , so the right-hand side vanishes, using the face and volume degrees of freedom for  $\pi_h$ . This completes the proof.  $\square$

### 5.6.1 The Clément interpolant

In some situations, it is useful to have an interpolant defined on discontinuous functions. In particular, in Section 13.4, we shall need to interpolate a function  $p \in H^1(\Omega)$ , where  $\Omega \subset \mathbb{R}^3$  is a bounded Lipschitz polyhedron. Functions in  $H^1(\Omega)$  are not necessarily continuous and so the standard interpolant  $\pi_h p$  is not well defined on this space (hence the restriction  $p \in H^s(\Omega)$ ,  $s > \frac{3}{2}$  in Theorem 5.48). In this section we shall define a generalized interpolant called the Clément interpolant [81] as generalized by Bernardi [40]. Other generalized interpolants are also available with many of the same properties (see, e.g. Chapter 4.8 of [60]) so our choice of this interpolant is, to some extent, arbitrary. These generalized interpolants have the desirable property of interpolating less smooth functions than the standard interpolant but have obvious defects which make them less desirable for general use. For example, the generalized interpolants all use average values of the function to be interpolated, and hence the value of the interpolant on one tetrahedron may depend on values of the function on other tetrahedra. This non-local behavior complicates the theory and implies that the generalized interpolant of a finite element function does not necessarily reproduce the finite element function exactly.

We start by following [143] and describe the Clément interpolant without regard for boundary conditions. Let  $\tau_b$  denote a regular tetrahedral mesh of  $\Omega$  using elements of maximum diameter  $b$ . Let  $U_b$  denote the space of continuous piecewise-linear finite element functions on  $\tau_b$  given by (5.55) with  $k = 1$  (higher-order generalizations are possible but we shall not require them here—see [40]). The standard set of global degrees of freedom  $\sum_b$  of  $U_b$  is given by function values at the vertices of the mesh. Let  $a_1, \dots, a_{N_b}$  denote the  $N_b$  vertices of  $\tau_b$ .

By the unisolvence property (Lemma 5.47), for each  $a_i$  there is a unique finite element function  $\theta_i \in U_b$  such that

$$\Theta_i(a_j) = \delta_{i,j}, \quad 1 \leq j \leq N_b,$$

and the functions  $\theta_i$ ,  $1 \leq i \leq N_b$ , are a standard Lagrange basis for  $U_b$ .

For each  $i$  we then define the macro-element  $\Delta_i$  by

$$\begin{aligned} \nabla_i &= \bigcup \{K \in \tau_b \mid \text{support}(\Theta_i) \cap K \neq \emptyset\} \\ &= \bigcup \left\{ K \in \tau_b \mid a_i \in \bar{K} \right\}. \end{aligned}$$

Since the mesh  $\tau_b$  is regular, the interior angles of the tetrahedra are bounded away from zero, independent of  $b$  and so there is a maximum number  $M$  of tetrahedra that meet at any vertex (the number  $M$  depends on the regularity constant for the mesh but is independent of  $b$ ). In addition, the number of macro-elements containing a given element  $K$  is also bounded independently of  $b$ . Using these facts, Bernardi [40] proves the following lemma, which makes critical use of the regularity of the mesh.

**Lemma 5.50** *If  $K$  and  $K'$  are contained in a macro-element  $\Delta_i$  we have  $h_K \leq Ch_{K'}$  where  $C$  is independent of  $b$  and  $i$ .*

With this lemma, Bernardi argues that it is possible to construct a finite number of reference domains  $\hat{\Delta}_l$ ,  $1 \leq l \leq L$  ( $L$  independent of  $h$ ), consisting of unions of at most  $M$  tetrahedra such that each macro-element  $\Delta_l$  is the image of some  $\hat{\Delta}_l$ . By this we mean that each tetrahedron in  $\Delta_l$  is mapped by an affine map to a distinct element in  $\hat{\Delta}_l$  and that the union of mapped elements is all of  $\Delta_l$ . The domains  $\hat{\Delta}_l$  play the role of the reference domain  $K$  in the theory of the previous sections of this chapter. In particular, because there are finitely many reference configurations  $\hat{\Delta}_l$  independent of  $h$ , the various interpolation and continuity constants appearing in the theory for these domains are bounded independently of  $h$ .

We can now define the Clément interpolation operator  $\Pi_{\text{Clem}} : H^1(\Omega) \rightarrow U_h$ . Let  $\Delta_l$  be a macro-element with  $\hat{\Delta}_l$  its reference configuration. Then let  $p \in H^1(\Omega)$  and define  $\hat{p}_j \in P_1$  to be the unique function such that

$$\int_{\hat{\Delta}_l} (\hat{p}_j - \hat{p}) \hat{\xi} d\hat{V} = 0 \quad \text{forall } \hat{\xi} \in P_1(\hat{\Delta}_l)$$

where, for each  $K \in \Delta_l$ , we obtain  $\hat{p}$  by the usual mapping. Thus  $\hat{p} = p \circ F_{K,K}$ , where  $F_{K,K}$  is the affine map between the appropriate  $K \in \hat{\Delta}_l$  and  $K \in \Delta_l$ . Then, if  $\hat{a}_j \in \hat{\Delta}_l$  maps to  $a_j \in \Delta_l$ , we have (5.58)

$$\Pi_{\text{Clem}} p = \sum_{j=1}^{N_h} \hat{p}_j(\hat{a}_j) \Theta_j.$$

This defines  $\Pi_{\text{Clem}}$  on  $H^1(\Omega)$  (i.e. when no boundary conditions are present), but because of the averaging procedure, even if  $p = 0$  on  $\partial\Omega$ , it is possible that  $\Pi_{\text{Clem}} p \neq 0$  on  $\partial\Omega$ .

We have already encountered in Chapter 4 the situation where  $\Omega$  is a connected domain with disconnected boundary consisting of two connected components  $\Gamma$  and  $\Sigma$ . Recall that

$$S = \left\{ p \in H^1(\Omega) \mid p = 0 \text{ on } \Gamma \text{ and } p = \text{constant on } \Sigma \right\}.$$

Note that  $\Gamma$  is exactly covered by a union of faces of the mesh (if  $\Gamma$  is just a portion of one component of  $\partial\Omega$ , it would be necessary that the mesh be chosen to exactly cover  $\Gamma$  — see [40]). In this case, for  $p \in S$ , we can define a conforming Clément interpolant, still denoted  $\Pi_{\text{Clem}}$ , in the same way as (5.58) but omitting from the sum terms corresponding to vertices on  $\Gamma$  and enforcing the interpolant to be constant on  $\Sigma$ . In this case  $\Pi_{\text{Clem}} p \in S_h = S \cap U_h$ . Theorem 5.1 of [40] then gives the following estimates for any  $p \in S$  and  $K \in \tau_h$ :

$$\|p - \Pi_{\text{Clem}} p\|_{L^2(K)} \leq Ch_K \|p\|_{H^1(D_K)} \quad \text{and}$$

$$\|p - \Pi_{\text{Clem}} p\|_{H^1(K)} \leq C \|p\|_{H^1(D_K)},$$

where  $D_K$  is the union of all macro-elements  $\Delta_l$  containing  $K$ .

Using the interpolation estimate that for any  $u \in H^1(K)$

$$\|u\|_{L^2(\partial K)}^2 \leq C \left( h_K^{-1} \|u\|_{L^2(K)}^2 + h_K \|\nabla u\|_{(L^2(K))^3}^2 \right)$$

with constant  $C$  independent of  $K$  (this result is proved in two dimensions in [226] and the same proof, with obvious modifications, proves it in three dimensions) allows us to obtain the following theorem.

**Theorem 5.51** *Let  $\tau_b$  be a regular mesh of  $\Omega$ . Then the modified Clément operator with boundary data (i.e. respecting the boundary conditions on  $\Gamma$  and  $\Sigma$ ) denoted by  $\Pi_{\text{Clem}} : S \rightarrow S_b$  satisfies*

$$\sum_{K \in \tau_h} \left( \frac{1}{h_K^2} \|p - \Pi_{\text{Clem}} p\|_{L^2(K)}^2 + \frac{1}{h_K} \|p - \Pi_{\text{Clem}} p\|_{L^2(\partial K)}^2 \right) \leq C \|p\|_{H^1(\Omega)}^2,$$

where  $C$  is independent of  $b$  and  $p$ .

## 5.7 An $L^2(\Omega)$ conforming space

We will now define a finite element space in  $L^2(\Omega)$  to complete the discrete de Rham complex. Let  $Z_b$  denote the space of piecewise ( $k-1$ )-degree discontinuous scalar elements on  $\tau_b$ , so

$$Z_h = \left\{ q_h \in L^2(\Omega) \mid q_h|_K \in P_{k-1} \text{ for all } K \in \tau_h \right\}.$$

It is easy to see that the  $L^2(\Omega)$  projection  $P_{0,b} : L^2(\Omega) \rightarrow Z_b$  is related to  $w_b$ . In fact, for all sufficiently smooth functions,

$$\nabla \cdot w_h u = P_{0,h} \nabla \cdot u$$

since, for all  $q \in P_{k-1}$ , using the definition of  $P_{0,b}$  followed by the integral identity (3.33), we obtain

$$\begin{aligned} & \int_K (\nabla \cdot w_h u - P_{0,h} \nabla \cdot u) q dV \\ &= \int_K (\nabla \cdot w_h u - \nabla \cdot u) q dV \\ &= - \int_K (w_h u - u) \cdot \nabla q dV + \int_{\partial K} (w_h u - u) \cdot v q dA. \end{aligned}$$

Here, both integrals on the right-hand side vanish, owing to the face and volume degrees of freedom defining  $w_b$ , so the commuting property of  $w_b$  and  $P_{0,b}$  is established. Thus, if we define, for  $\delta > 0$ ,

$$\begin{aligned} U &= H^{3/2+\delta}(\Omega), \\ V &= \left\{ v \in \left( H^{1/2+\delta}(\Omega) \right)^3 \mid \nabla \times v \in \left( H^{1/2+\delta}(\Omega) \right)^3 \right\}, \\ W &= \left\{ w \in \left( H^{1/2+\delta}(\Omega) \right)^3 \mid \nabla \cdot w \in L^2(\Omega) \right\}, \end{aligned}$$

then the above result and Theorems 5.40 and 5.49 show that the following discrete de Rham complex commutes:

(5.59)

$$\begin{array}{ccccccc}
H^1(\Omega) & \xrightarrow{\nabla} & H(\text{curl}; \Omega) & \xrightarrow{\nabla \times} & H(\text{div}, \Omega) & \xrightarrow{\nabla \cdot} & L^2(\Omega) \\
\cup & & \cup & & \cup & & \\
U & & V & & W & & \\
\pi_h \downarrow & & r_h \downarrow & & w_h \downarrow & & P_{0,h} \downarrow \\
U_h & \xrightarrow{\nabla} & V_h & \xrightarrow{\nabla \times} & W_h & \xrightarrow{\nabla \cdot} & Z_h .
\end{array}$$

Here  $U_b$  is given by (5.55),  $V_b$  by (5.40),  $W_b$  by (5.28) and  $Z_b$  as above.

We have established error estimates for  $\pi_b$ ,  $r_b$  and  $w_b$ . We could also easily establish such results for  $P_{0,b}$  (this is standard finite element theory) and hence establish the following result for  $w_b$ : (5.60)

$$\|\nabla \cdot (w_h u - u)\|_{(L^2(\Omega))^3} = \|P_{0,h} \nabla \cdot u - \nabla \cdot u\|_{(L^2(\Omega))^3} \leq Ch^s \|\nabla \cdot u\|_{H^s(\Omega)}$$

for  $0 \leq s \leq k-1$ , provided  $u$  is sufficiently smooth such that  $\nabla \cdot u \in H^s(\Omega)$ , and  $w_b u$  is well defined ( $u \in W$  is enough). This proves the estimate in the remark following Theorem 5.25.

We see that by establishing convergence in  $L^2(\Omega)$  for  $\pi_b$  in  $(L^2(\Omega))^3$  for  $r_b$  and  $w_b$ , and in  $L^2(\Omega)$  for  $P_{0,b}$ , we can establish higher norm convergence by using the commuting diagram. For example, using the estimates for  $r_b$ , we have

$$\begin{aligned}
\|\nabla(\pi_h p - p)\|_{(L^2(\Omega))^3} &\leq \|r_h \nabla p - \nabla p\|_{(L^2(\Omega))^3} \\
&\leq Ch^s \|\nabla p\|_{H^s(\Omega)} \quad \text{for } \frac{1}{2} + \partial \leq s \leq k-1.
\end{aligned}$$

## 5.8 Boundary spaces

A tetrahedral mesh  $\tau_b$  of  $\Omega$  induces a triangular mesh on  $\partial\Omega$  in the sense that the faces of the elements in  $\tau_b$  that lie on  $\partial\Omega$  cover the boundary and obey the usual finite element meshing constraints (see Section 5.3). Furthermore if the mesh on  $\Omega$  is regular, then so is the mesh on the boundary (interpreting regularity in terms of the inscribed and circumscribed circles in the same way as in Definition 5.11).

Let us denote the mesh induced by  $\tau_b$  on  $\partial\Omega$  as  $\tau_b(\partial\Omega)$ . For a suitably smooth function  $u$ , we shall need to estimate the error in (5.61)

$$\|\gamma_T(u - r_h u)\|_{L_t^2(\partial\Omega)} = \|(v \times (u - r_h u)|_{\partial\Omega}) \times v\|_{L_t^2(\partial\Omega)},$$

where  $\gamma_T$  is the trace operator defined in (3.46). This estimate can be performed in two ways. We could estimate the error in terms of boundary norms of  $\gamma_T(u)$  noting that  $\gamma_T(r_h u)$  lies in the appropriate two-dimensional analogue of the set  $R_k$  (this is just the rotated Raviart–Thomas space). However, we wish to relate the convergence rate to volume norms of the function  $u$ . We, therefore, prove the following (most likely suboptimal) result.

**Lemma 5.52** Suppose  $u \in (H^s(\Omega))^3$  and  $\nabla \times u \in (H^s(\Omega))^3$  for some  $1/2 < s \leq k$ . Then

$$\|Y_T(u - r_h u)\|_{L_t^2(\partial\Omega)} \leq Ch^{s-1/2} (\|u\|_{H^s(\Omega)} + \|\nabla \times u\|_{H^s(\partial\Omega)}).$$

**Proof** Using (5.61) we have

$$\|Y_T(u - r_h u)\|_{L_t^2(\partial\Omega)}^2 \leq \|u - r_h u\|_{(L^2(\partial\Omega))^3}^2 = \sum_{f \in T_h(\partial\Omega)} \|u - r_h u\|_{(L^2(f))^3}^2.$$

If we now map to the reference element using the usual volume change of variables (5.33) we can ensure that the face  $f$  maps to the face of the reference element in the  $(\hat{x}_1, \hat{x}_2)$ -plane. This is exactly the standard two-dimensional reference element. We obtain

$$\|u - r_h u\|_{(L^2(f))^3}^2 = \int_{\hat{f}} \left| B_K^{-T} (\hat{u} - r_{\hat{K}} \hat{u}) \right|^2 \frac{\text{area}(f)}{\text{area}(\hat{f})} d\hat{A}.$$

Thus we see that

$$\begin{aligned} \|u - r_h u\|_{(L^2(f))^3}^2 &\leq C \left| B_K^{-T} \right|^2 \text{area}(f) \|(\hat{u} - r_{\hat{K}} \hat{u})\|_{(L^2(\hat{f}))^3}^2 \\ &\leq C \|(\hat{u} - r_{\hat{K}} \hat{u})\|_{(L^2(\hat{f}))^3}^2. \end{aligned}$$

Now, using Theorem 3.9 on  $K$ , we obtain

$$\|(\hat{u} - r_{\hat{K}} \hat{u})\|_{(L^2(\hat{f}))^3}^2 \leq C \|(\hat{u} - r_{\hat{K}} \hat{u})\|_{(H^s(\hat{K}))^3}^2.$$

We may now proceed as in the proof of Theorem 5.41 (following (5.45)) by using the Deny–Lions theorem 5.5 and mapping back to the element  $K$  having  $f$  as a face using Lemma 5.43.  $\square$

Putting together this lemma and Theorem 5.41 provides the following estimate for the error in  $X$ . This result is not optimal, since we obtain at best an estimate of  $O(h^{1/2})$  for linear edge elements. At the expense of a higher norm on the right-hand side, we could improve the result, but this would not fit with our later theory.

**Lemma 5.53** Suppose  $u \in (H^s(\Omega))^3$  and  $\nabla \times u \in (H^s(\Omega))^3$  for some  $1/2 < s \leq k$  then

$$\|(u - r_h u)\|_X \leq Ch^{s-1/2} (\|u\|_{(H^s(\Omega))^3} + \|\nabla \times u\|_{(H^s(\Omega))^3}).$$

Sometimes, we shall also need to assume that on boundaries where the impedance boundary condition is applied (denoted  $\Sigma$ ), the mesh has an additional uniformity. For a given face  $f$  we define the diameter of  $f$ , denoted  $b_f$ , to be the diameter of the largest circle in the plane of  $f$  containing  $f$  in its interior. Now we say that a mesh  $\tau_b$ ,  $b > 0$ , is *quasi-uniform on  $\Sigma$*  if the following holds.

**Definition 5.54** A family of meshes  $\tau_b$ ,  $b > 0$ , is said to be *quasi-uniform on  $\Sigma$*  if there are constants  $\tau > 0$  and  $b_0 > 0$  independent of  $b$  such that

$$\tau h \leq h_f \text{ for all } f \in \mathcal{V}_h(\Sigma) \text{ and } h_0 \geq h > 0.$$

Effectively, this definition rules out an unbounded disparity in the diameter of the elements in the mesh on  $\Sigma$  as  $b \rightarrow 0$ . Usually, we would like the flexibility of allowing very small tetrahedra in some parts of the mesh (but maintaining a regular mesh of course), so we prefer to assume quasi-uniformity only when all else fails! However, in our problem  $\Sigma$  is an auxiliary boundary and we expect the solution to be smooth near  $\Sigma$ . Thus the restriction of quasi-uniformity on  $\Sigma$  is less significant than assuming the entire mesh  $\tau_b$  is quasi-uniform. Nevertheless, it would be desirable to do away with this assumption.

We shall use the inverse assumption to bound certain norms of discrete functions as follows.

**Lemma 5.55** Suppose  $p_b \in P_k$  for fixed  $k$  on a triangle  $T \in \tau_b(\Sigma)$ . Then there is a constant  $C$  independent of  $b$  such that

$$\|p_b\|_{L^\infty(T)} \leq Ch^{-1} \|p_b\|_{L^2(T)} \quad \text{and} \quad \|p_b\|_{H^1(T)} \leq Ch^{-1} \|p_b\|_{L^2(T)}.$$

**Remark 5.56** Results such as this are classical, and more general cases can be found in [80].

**Proof of Lemma 5.55** Both estimates are proved the same way. We only provide details for the first estimate. By mapping to the reference domain  $\tilde{T}$  using an affine map, we have

$$\|p_b\|_{L^\infty(T)} = \|\hat{p}\|_{L^\infty(\tilde{T})}.$$

On the reference domain, since  $P_k$  is a finite-dimensional vector space, the equivalence of norms on this space shows that there is a constant  $C$  such that

$$\|\hat{p}\|_{L^\infty(\tilde{T})} \leq C \|\hat{p}\|_{L^2(\tilde{T})}.$$

Now mapping back to the element  $T$ , we have

$$\|\hat{p}\|_{L^2(\tilde{T})} \leq \frac{1}{\sqrt{|\det(\tilde{B}_T)|}} \|p_b\|_{L^2(T)},$$

where  $\tilde{B}_T$  is the matrix in the affine map that maps  $\tilde{T}$  to  $T$ , where  $\tilde{T}$  is the two-dimensional reference element in the same plane as  $T$ . Using the two-dimensional analogue of Theorem 5.10, we have

$$|\det(\tilde{B}_T)| \geq Ch_T^2$$

and use of this estimate, together with the quasi-uniformity assumption, proves the desired result.  $\square$

The assumption of quasi-uniformity allows us to prove the following more technical inverse estimate from [58].

**Lemma 5.57** Suppose  $\tau_b(\Sigma)$ ,  $b > 0$  is a regular and quasi-uniform family of meshes on  $\Sigma$ . Let  $p_h$  be a piecewise  $k$ -degree polynomial on  $\tau_b(\Sigma)$ . Then for any  $\delta$  with  $0 \leq \delta < \frac{1}{2}$  there is a constant  $C$  such that

$$\|p_h\|_{H^\delta(\Sigma)} \leq Ch^{-\delta} \|p_h\|_{L^2(\Sigma)}.$$

**Remark 5.58** Note that this lemma shows that piecewise polynomials are in  $H^\delta(\Sigma)$  provided  $0 \leq \delta < \frac{1}{2}$ . Of course, by applying the estimate to each component of a piecewise polynomial vector function we see that quasi-uniformity implies an inverse inequality for vector piecewise polynomials. In particular,

$$\|\nu \times u_h\|_{(H^\delta(\Sigma))^3} \leq Ch^{-\delta} \|\nu \times u_h\|_{(L^2(\Sigma))^3},$$

when  $u_h \in V_b$ .

**Proof of Lemma 5.57** The proof is from [58] and uses the interpolation theory of Sobolev spaces which we have not discussed in this book. A good reference for an introduction to this theory is [60]. The difficulty with the proof is that the upper limit  $s = \frac{1}{2}$  is not included in the range of validity of the estimate. In addition the fractional Sobolev norm involves a double integral over the boundary  $\Sigma$ . This complicates the use of a mapping approach to the proof. Instead, we give the proof from [58] which proceeds directly via function space interpolation. The fractional Sobolev norm is equivalent to the following norm:(5.62)

$$\|p_h\|_{H^\delta(\Sigma)}^2 = \int_0^\infty K(p_h, t)^2 t^{-2\delta-1} dt,$$

where(5.63)

$$K(p_h, t)^2 = \inf_{v \in H^1(\Sigma)} \left( \|p_h - v\|_{L^2(\Sigma)}^2 + t^2 \|v\|_{H^1(\Sigma)}^2 \right).$$

Here  $v$  can depend on  $t$ .

For  $b \leq t \leq 1$  we choose  $v = 0$  and obtain(5.64)

$$\int_b^\infty K(p_h, t)^2 t^{-2\delta-1} dt \leq (2\delta)^{-1} h^{-2\delta} \|p_h\|_{L^2(\Sigma)}^2.$$

The choice for  $0 \leq t \leq b$  is much more complex. We want to approximate  $p_h$  on each triangle, but not up to the boundary of the triangle where the discontinuity of  $p_h$  occurs. Suppose the faces of  $\Sigma$  are  $f_1, f_2, \dots, f_M$ . We use a non-negative cutoff function  $\phi_l \in C(\Sigma)$  such that  $\phi_l|_{C_0^\infty(f_l)}$  and with the following properties:(5.65)

$$\phi_l(x) = \begin{cases} 1 & \text{if } x \in f_l, \text{ and } \text{dist}(x, \partial f_1) > t, \\ 0 & \text{if } x \notin f_l, \end{cases}$$

(5.66)

$$|\nabla_{f_l} \phi_l| \leq C t^{-1} \text{ for all } x,$$

where  $\text{dist}(x, \partial f)$  is the distance of  $x$  from the boundary  $\partial f$ . On  $f$ , such a function is just a standard cutoff function since each face is planar. The function  $v$  in (5.63) is taken to be

$$v = \sum_{l=1}^{M_h} \phi_l p_h.$$

Of course,  $v$  is a smooth function and using the arithmetic geometric mean inequality we have

$$\|v\|_{H^1(f)}^2 \leq C \left( \|p_h\|_{H^1(f)}^2 + \|p_h\|_{L^\infty(f)}^2 \|\nabla_{f_l} \phi_l\|_{(L^2(f))}^2 \right).$$

Note that  $\nabla_{f_l} \phi_l$  only non-zero in a strip of width  $t$  around the edge of the triangle (hence having an area  $O(bt)$ ). Using this fact and the estimate (5.66) we obtain

$$\|v\|_{H^1(f)}^2 \leq C \left( \|p_h\|_{H^1(f)}^2 + ht^{-1} \|p_h\|_{L^\infty(f)}^2 \right).$$

The same idea is used in the following estimate:

$$\|p_h - v\|_{L^2(f)}^2 = \|(1 - \phi_l)p_h\|_{L^2(f)}^2 \leq Ch t \|p_h\|_{L^\infty(f)}^2$$

Now using the above two estimates, and the inverse estimates in Lemma 5.55, we have

$$K(p_h, t)^2 \leq Ch t^{-1} \|p_h\|_{L^2(f)}^2.$$

Hence

$$\int_0^h K(p_h, t)^2 t^{-2\bar{\delta}-1} dt \leq Ch^{-2\bar{\delta}} \|p_h\|_{L^2(f)}^2.$$

Using this estimate and (5.64) in (5.62) proves the result.  $\square$

Finally, let us remark that if  $u_b \in V_b$  is an edge finite element function then  $\gamma_b(v_b) \in H(\text{Div}; \Sigma)$  where

$$H(\text{div}; \Sigma) = \left\{ v \in (L^2(\Sigma))^3 \mid \nabla_\Sigma \cdot v \in L^2(\Sigma), v \cdot n = 0 \text{ a.e. on } \Sigma \right\}.$$

Similarly,  $\gamma_T(v_b) \in H(\text{Curl}; \Sigma)$ , where

$$H(\text{Curl}; \Sigma) = \left\{ v \in (L^2(\Sigma))^3 \mid \nabla_\Sigma \times v \in L^2(\Sigma), v \cdot n = 0 \text{ a.e. on } \Sigma \right\}.$$

# 6 FINITE ELEMENTS ON HEXAHEDRA

## 6.1 Introduction

In this chapter we shall discuss the definition and properties of Nédélec's first family of finite elements on hexahedra [233]. The presentation follows his work closely, but, as in the previous chapter, we emphasize the connections between the various spaces. The outline of the chapter is roughly the same as the previous chapter, and, in particular, it is necessary to read Sections 5.2 (last part) and 5.3 before this chapter.

We assume a regular finite element mesh  $\tau_b$ ,  $b > 0$  (see Section 5.3), of hexahedra of maximum diameter  $b$  with the very strong assumption that all elements are parallelepipeds with edges parallel to the coordinate axes. This implies that each element  $K \in \tau_b$  can be obtained from the reference element  $\mathcal{K} = (0, 1)^3$  via a diagonal affine map  $F_K(\hat{x}) = B_K \hat{x} + b_K$ , where  $B_K$  is an invertible diagonal matrix. In this case the conclusions of Lemma 5.9 concerning the change of norms under  $F_K$  still holds and the estimate for the norm of  $B_K$ ,  $B_K^{-1}$ , and its determinant in Lemma 5.10 also still hold.

Because of the simplicity of the mapping, it is usual to work directly with basis functions on the mapped element and not work via the reference element and affine map as in the previous chapter, where we considered tetrahedral elements. Although we actually define the various elements on the reference domain  $\mathcal{K}$ , the same definitions can be used on a target element  $K$  in the mesh. Our analysis of hexahedral elements is a little less thorough than for tetrahedra, but having seen the analysis for tetrahedra, the reader can fill in the details!

Hexahedral elements are popular in engineering and are the basis of a number of successful codes. The elements we shall describe here are due to Nédélec [233] (in this case of divergence elements they are a straightforward generalization to three dimensions of the Raviart–Thomas mixed finite element [260]).

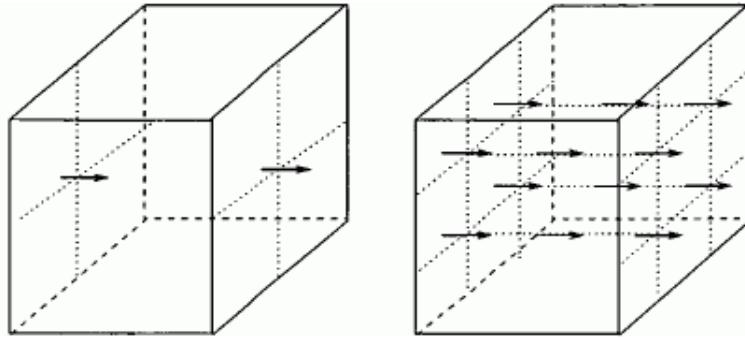
## 6.2 Divergence conforming elements on hexahedra

On the reference element  $\mathcal{K}$  we define the element as follows (see Fig. 6.1 for a graphical representation of these degrees of freedom):

**Definition 6.1** For given integer  $k \geq 1$  the divergence conforming element of Nédélec is defined as follows.

- (1) The reference element is  $\mathcal{K} = (0, 1)^3$ .
- (2) The polynomial space is  $P_{\mathcal{K}} = \mathcal{Q}_{k,k-1, k-1} \times \mathcal{Q}_{k-1, k,k-1} \times \mathcal{Q}_{k-1, k-1, k}$ .

Fig. 6.1. The degrees of freedom for the first two divergence conforming elements on hexahedra. For simplicity we only show the degrees of freedom for the  $x_1$ -component. *Left:*  $k = 1$ ; the average value of the normal component of the finite element vector field is given on each face (represented by the bold face normal vectors on each face). *Right:*  $k = 2$ ; there are four normal component degrees of freedom per face (denoted by thick arrows) and, in addition, four interior degrees of freedom represented by non-bold-face vectors.



(3) The degrees of freedom are given on faces  $\hat{f}$  with normal  $\hat{v}$  and in the interior of  $\hat{K}$  (they are well defined for  $\hat{u} \in (H^{1/2 + \delta}(\hat{K}))^3$ ,  $\delta > 0$ , by Lemma 5.15):

(a) for the faces(6.1)

$$M_{\hat{f}}(\hat{u}) = \left\{ \int_{\hat{f}} \hat{u} \cdot \hat{v} q dA \text{ for each } q \in Q_{k-1,k-1}(\hat{f}) \right. \\ \left. \text{and each face } \hat{f} \right\},$$

(b) for the volume

$$M_{\hat{K}}(\hat{u}) = \left\{ \int_{\hat{K}} \hat{u} \cdot \hat{q} dV \text{ for all } \right. \\ \left. \hat{q} \in Q_{k-2,k-1,k-1} \times Q_{k-1,k-2,k-1} \times Q_{k-1,k-1,k-2} \right\}.$$

Then  $\sum_k = M_{\hat{f}}(\hat{u}) \cup M_{\hat{K}}(\hat{u})$ .

Using the transformation (5.20), a basis function on  $\hat{K}$  can be mapped to a basis function on a general element  $K$ . Since the mapping is a diagonal affine map ( $B_K$  is diagonal), this simply scales each component of  $\hat{u}$  and so we do not define the element on a general hexahedron  $K$  (but see the tetrahedral degrees of freedom in Definition 5.18 for how  $\hat{q}$  on  $K$  must be mapped).

We start our analysis of this element by showing that the element is divergence conforming and unisolvant.

**Theorem 6.2** A vector function  $u \in \mathcal{Q}_{k,k-1,k-1} \times \mathcal{Q}_{k-1,k,k-1} \times \mathcal{Q}_{k-1,k-1,k}$  defined on the reference hexahedron  $K = (0, 1)^3$  is uniquely determined by the degrees of freedom (6.1) and (6.2). Moreover, the space  $W_h$  of finite elements on the mesh  $\tau_h$  defined by mapping the element in Definition 6.1 from the reference element using (5.20) is divergence conforming, so that  $W_h \subset H(\text{div}; \Omega)$ . In particular, (6.3)

$$W_h = \{u_h \in H(\text{div}; \Omega) | u_h|_K \in \mathcal{Q}_{k,k-1,k-1} \times \mathcal{Q}_{k-1,k,k-1} \times \mathcal{Q}_{k-1,k-1,k} \text{ for all } K \in \mathcal{T}_h\}.$$

**Proof** First we show conformity. Using Theorem 5.3 (as in the proof of Lemma 5.20), we need to prove that if all the degrees of freedom of type (6.1) of  $\hat{u} \in P_K$  vanish on a particular face, then  $\hat{u} \cdot \hat{\nu} = 0$  on that face. But, using the fact that the element has faces parallel to the coordinate planes, on  $\hat{f}$  the normal component of  $u$  is such that  $\hat{u} \cdot \hat{\nu} \in \mathcal{Q}_{k-1,k-1}$ , and so we may choose  $\hat{q} = \hat{u} \cdot \hat{\nu}$  in (6.1) and hence conclude the desired result.

To prove unisolvence, we first note that the dimension of  $P_K$  is  $3k^2(k+1)$ , and this is also the number of degrees of freedom in  $\sum_K$ . Hence it suffices to prove that if all degrees of freedom vanish for  $\hat{u} \in P_K$  then  $\hat{u} = 0$ . But by the conformity proof, we know that  $\hat{u} \cdot \hat{\nu} = 0$  on each face of  $K$  and hence (again using the fact that the faces are parallel to the coordinate axes) we have

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

for some  $\hat{r} \in \mathcal{Q}_{k-2,k-1} \times \mathcal{Q}_{k-1,k-2} \times \mathcal{Q}_{k-1,k-2}$ . Hence choosing  $\hat{q} = \hat{r}$  in (6.2) proves that  $\hat{r} = 0$  and we are done.  $\square$

Now that we have a well-defined finite element, we can define a global interpolation operator  $w_h$  using the element-wise interpolation operator (see (5.26) and (5.27)). We have the following error estimate using the same argument as for Theorem 5.25 (but substantially simpler — owing to the fact that  $B_K$  is diagonal).

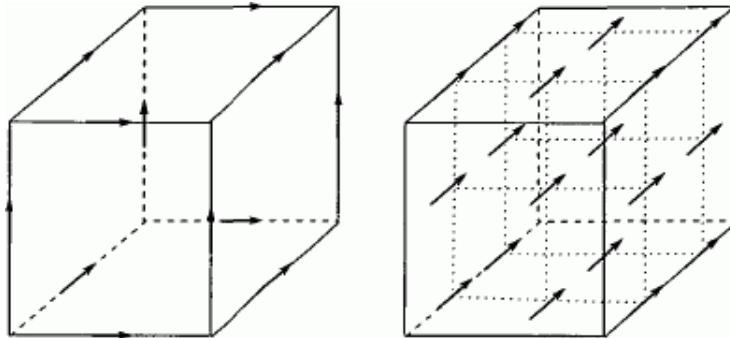
**Theorem 6.3** Suppose  $\tau_h$  is a regular family of hexahedral meshes on  $\Omega$  with edges parallel to the coordinate axes. Assume  $0 < \delta < \frac{1}{2}$ . Then if  $u \in (H(\Omega))^3$ ,  $1/2 + \delta \leq s \leq k$ , there is a constant  $C$  independent of  $h$  and  $u$  such that (6.4)

$$\|u - w_h u\|_{(L^2(\Omega))^3} \leq Ch^s \|u\|_{(H^s(\Omega))^3}, \quad \frac{1}{2} + \delta \leq s \leq k.$$

Let us close by remarking that the simplest element in this family, when  $k = 1$  (see Fig. 6.1) has

$$u|_K \in \mathcal{Q}_{1,0,0} \times \mathcal{Q}_{0,1,0} \times \mathcal{Q}_{0,0,1};$$

Fig. 6.2. Degrees of freedom for the first two curl conforming elements on hexahedra. *Left:*  $k = 1$ ; the average value of tangential component of the finite element vector field is given on each edge. *Right:*  $k = 2$ ; only the degrees of freedom for the second component  $u_2$  of the field are shown. There are two tangential component degrees of freedom per edge, two per face and two interior degrees of freedom.



therefore

$$u|_K = (a_1 + b_1 x_1, a_2 + b_2 x_2, a_3 + b_3 x_3)^\top.$$

The six degrees of freedom are the average flux on each face and these determine  $a$  and  $b$  in the above expression for  $u$ . Similar expansions are easy to write down for any  $k$ .

### 6.3 Curl conforming hexahedral elements

To continue our discussion of finite elements on hexahedra, we now present the edge elements due to Nédélec [233]. Again we shall restrict ourselves to right hexahedra with edges parallel to the coordinate axis. We assume that the domain  $\Omega$  is covered with a mesh of regular parallelepipeds of maximum diameter  $h$  to form the mesh  $\tau_h$ . On the reference element  $K$ , we define the element as follows (see Fig. 6.2 for a graphical representation of these degrees of freedom):

**Definition 6.4** For given integer  $k \geq 1$  the curl conforming element of Nédélec is defined as follows:

- (1) the reference element is  $K = (0, 1)^3$ ;
- (2) the polynomial space is  $P_k = Q_{k-1, k, k} \times Q_{k, k-1, k} \times Q_{k, k, k-1}$ ;
- (3) the degrees of freedom are given on edges  $\hat{e}$  with unit tangent  $\hat{\tau}$ , on faces  $\hat{f}$  with normal  $\hat{\nu}$  and in the interior of  $K$ . They are well defined for  $\hat{u} \in (H^{1/2+\delta}(K))^3$ ,  $\delta > 0$  such that  $\nabla \times \hat{u} \in (L^p(K))^3$  for some  $p > 2$  (this follows from Lemma 5.38) as follows:

- (a) for the edges:(6.5)

$$M_{\hat{e}}(\hat{u}) = \left\{ \int_{\hat{e}} \hat{u} \cdot \hat{\tau} q ds \quad \text{foreach } q \in P_{k-1}(\hat{e}) \right. \\ \left. \text{and each edge } \hat{e} \right\};$$

(b) for the faces:(6.6)

$$M_{\hat{f}}(\hat{u}) = \left\{ \int_{\hat{f}} \hat{u} \times \hat{v} \cdot \hat{q} dA \text{ for each } \hat{q} \in Q_{k-2,k-1}(\hat{f}) \times Q_{k-1,k-2}(\hat{f}) \text{ and each face } \hat{f} \right\},$$

where we note that  $\hat{u} \times \hat{v}$  is a vector in the plane of  $\hat{f}$  and hence can be interpreted as a two dimensional vector;

(c) for the volume:(6.7)

$$M_{\hat{K}}(\hat{u}) = \left\{ \int_{\hat{K}} \hat{u} \cdot \hat{q} dV \text{ for all } \hat{q} \in Q_{k-1,k-2,k-2} \times Q_{k-2,k-1,k-2} \times Q_{k-2,k-2,k-1} \right\}.$$

Then  $\sum_k = M_e(\hat{u}) \cup M_f(\hat{u}) \cup M_K(\hat{u})$ .

Using the transformation (5.33) the basis function on  $K$  can be mapped to the basis function on a general element  $K$ . Since the mapping is a diagonal affine map ( $B_K$  is diagonal), this simply scales each component of  $\hat{u}$  and so we do not define the element on a general hexahedron  $K$ .

We start our analysis of this element by showing that the element is curl conforming and unisolvant.

**Theorem 6.5** A vector function  $u \in Q_{k-1,k,k} \times Q_{k,k-1,k} \times Q_{k,k,k-1}$ ,  $k \geq 1$ , defined on the reference hexahedron  $\hat{K}$  is uniquely determined by the degrees of freedom (6.5)–(6.7). Moreover, the space  $V_h$  of finite elements on the mesh  $\tau_h$  defined by mapping the element in Definition 6.4 from the reference element using (5.33) is curl conforming, so that  $V_h \subset H(\text{curl}; \Omega)$ . In addition, (6.8)

$$V_h = \left\{ u_h \in H(\text{curl}; \Omega) \mid u_h|_K \in Q_{k-1,k,k} \times Q_{k,k-1,k} \times Q_{k,k,k-1} \text{ for all } K \in \tau_h \right\}.$$

**Proof** First we show conformity. Using Theorem 5.3 (as in the proof of Lemma 5.35), we need to prove that if all the degrees of freedom of type (6.5) and (6.1) of  $\hat{u} \in P_k$  vanish on a particular face, then  $\hat{u} \times \hat{v} = 0$  on that face. Suppose we consider the face  $\hat{x}_3 = 0$ . Then on this face the tangential components of  $\hat{u}$  are  $\hat{u}_1 \in Q_{k-1,k}$  and  $\hat{u}_2 \in Q_{k,k-1}$ . On each edge of this face,  $\hat{u} \cdot \hat{\tau} \in P_{k-1}$ , and

hence choosing  $\hat{q} = \hat{u} \cdot \hat{\tau}$  in the degrees of freedom in (6.5) shows that  $\hat{u} \cdot \hat{\tau} = 0$  on each edge of this face.

Now consider the face  $\hat{J}$  with  $\hat{x}_3 = 0$ . Because  $\hat{u} \cdot \hat{\tau} = 0$  on the edges of this face, we know that the tangential components of  $\hat{u}$  on this face have the factorization

$$\begin{aligned}\hat{u}_1 &= \hat{x}_2(1 - \hat{x}_2)\hat{v}_1 \text{ for some } \hat{v}_1 \in Q_{k-1,k-2}, \\ \hat{u}_2 &= \hat{x}_1(1 - \hat{x}_1)\hat{v}_2 \text{ for some } \hat{v}_2 \in Q_{k-2,k-1}.\end{aligned}$$

Choosing  $\hat{q}_1 = \hat{v}_2$  and  $\hat{q}_2 = -\hat{v}_1$  in the degrees of freedom (6.6) shows that  $\hat{v}_1 = \hat{v}_2 = 0$  on this face and hence  $\hat{u} \times \hat{v}$  vanishes on this face, and we have verified conformity.

To prove unisolvence, we first note that the dimension of  $P_k$  is  $3k(k+1)^2$ , and this is also the number of degrees of freedom in  $\sum_k$ . Hence, it suffices to prove that if all degrees of freedom vanish for  $\hat{u} \in P_k$  then  $\hat{u} = 0$ . But by the conformity proof, we know that  $\hat{u} \times \hat{v} = 0$  on each face of  $K$  and hence (using the fact that the faces are parallel to the coordinate axes) we have

$$\hat{u}_1 = \left( \hat{x}_2(1 - \hat{x}_2)\hat{x}_3(1 - \hat{x}_3)\hat{\varphi}_1, \hat{x}_1(1 - \hat{x}_1)\hat{x}_3(1 - \hat{x}_3)\hat{\varphi}_2, \hat{x}_1(1 - \hat{x}_1)\hat{x}_2(1 - \hat{x}_2)\hat{\varphi}_3 \right)^T$$

for some  $\hat{\varphi} \in Q_{k-1, k-2, k-2} \times Q_{k-2, k-1, k-2} \times Q_{k-2, k-2, k-1}$ . Hence, choosing  $\hat{q} = \hat{\varphi}$  in (6.7) proves that  $\hat{\varphi} = 0$  and we are done.  $\square$

Now that we have a well-defined finite element, we can define a global interpolant using the element-wise interpolation operator as in (5.41). As in the previous chapter, we denote this operator by  $\tau_b$ . We have the following error estimate using the same argument as for Theorem 5.41.

**Theorem 6.6** Suppose  $\tau_b$  is a regular family of hexahedral meshes on  $\Omega$  with edges parallel to the coordinate axis. Provided  $u \in (H^s(\Omega))^3$ , and  $\nabla \times u \in (H^s(\Omega))^3$  for  $\frac{1}{2} + \delta \leq s \leq k$ ,  $0 < \delta < \frac{1}{2}$ , there is a constant  $C$  independent of  $b$  and  $u$  such that (6.9)

$$\begin{aligned}\|u - r_h u\|_{(L^2(\Omega))^3} + \|\nabla \times (u - r_h u)\|_{(L^2(\Omega))^3} \\ \leq Ch^s \left( \|u\|_{(H^s(\Omega))^3} + \|\nabla \times u\|_{(H^s(\Omega))^3} \right)\end{aligned}$$

for  $\frac{1}{2} + \delta \leq s \leq k$ . In addition, if  $\nabla \times u \in Q_{k, k-1, k-1} \times Q_{k-1, k, k-1} \times Q_{k-1, k-1, k}$  on an element  $K$  then (6.10)

$$\|u - r_h u\|_{(L^2(K))^3} \leq C \left( h_K^{1/2+\delta} \|u\|_{(H^{1/2+\delta}(K))^3} + h_K \|\nabla \times u\|_{(L^2(K))^3} \right).$$

Our final result of this section links the space presented above with the divergence conforming space  $W_b$  from (6.3).

**Theorem 6.7** Let  $W_b$  be the space given by (6.3) and  $V_b$  be the space given by (6.8). Then  $\nabla \times V_b \subset W_b$ . Furthermore, using the degrees of freedom in this chapter, if  $u$  is smooth enough such that  $r_b u$  and  $w_b \nabla \times u$  are defined, then  $\nabla \times r_b u = w_b \nabla \times u$ .

**Proof** The first part is clear since  $\nabla \times V_b \subset H(\text{div}; \Omega)$ , and, if  $u_b \in V_b$ , a direct calculation shows that  $\nabla \times u_b|_K \in \mathcal{Q}_{k, k+1, k+1} \times \mathcal{Q}_{k-1, k, k+1} \times \mathcal{Q}_{k-1, k-1, k}$ .

The second part of the theorem is proved by verifying that the degrees of freedom given in (6.1) and (6.2) vanish for  $\nabla \times r_b u - w_b \nabla \times u$ . The result then follows from the unisolvence of the element proved in Theorem 6.2. We perform the analysis on the reference element. For degrees of the type given in (6.1) we let  $\hat{f}$  be a face of  $\hat{K}$  with normal  $\hat{v}$ , and let  $\hat{q} \in \mathcal{Q}_{k-1, k-1}(\hat{f})$ . Then using the degrees of freedom for  $w_{\hat{K}} \nabla \times \hat{u}$  from (6.1), (3.52), (3.15) and integration by parts, we have

$$\begin{aligned} & \int_{\hat{f}} (\widehat{\nabla} \times r_{\hat{K}} \hat{u} - \omega_{\hat{K}} \widehat{\nabla} \times \hat{u}) \cdot \hat{v} \hat{q} d\hat{A} \\ &= \int_{\hat{f}} (\widehat{\nabla} \times r_{\hat{K}} \hat{u} - \widehat{\nabla} \times \hat{u}) \cdot \hat{v} \hat{q} d\hat{A} \\ &= \int_{\hat{f}} \nabla_{\hat{f}} \cdot (\hat{v} \times (r_{\hat{K}} \hat{u} - \hat{u})) \hat{q} d\hat{A} \\ &= \int_{\hat{f}} \hat{v} \times (r_{\hat{K}} \hat{u} - \hat{u}) \cdot \nabla_{\hat{f}} \hat{q} d\hat{A} - \int_{\partial \hat{f}} \hat{v}_{\hat{f}} \cdot (\hat{v} \times (r_{\hat{K}} \hat{u} - \hat{u})) \hat{q} d\hat{s}, \end{aligned}$$

where  $\hat{v}_{\hat{f}}$  is the unit outward normal to  $\hat{f}$  in the plane of  $\hat{f}$ . The first term on the right-hand side vanishes by using the degrees of freedom (6.6) for  $r_{\hat{K}} u$  since  $\nabla_{\hat{f}} \hat{q} \in \mathcal{Q}_{k-2, k-1}(\hat{f}) \times \mathcal{Q}_{k-1, k-2}(\hat{f})$ . The second term vanishes using the degrees in (6.5) since  $\hat{v} \times (\hat{v} \hat{q}) = \pm \hat{n} \hat{q}$  and  $\hat{q} \in P_{k-1}(\hat{e})$  for each edge  $\hat{e}$  of  $\hat{f}$ .

For the volume degrees of freedom in (6.2), if

$$\hat{q} \in \mathcal{Q}_{k-2, k-1, k-1} \times \mathcal{Q}_{k-1, k-2, k-1} \times \mathcal{Q}_{k-1, k-1, k-2}$$

we have, using the degrees of freedom (6.2) for  $\hat{w}$  and integration by parts,

$$\begin{aligned} & \int_{\hat{K}} (\widehat{\nabla} \times r_{\hat{K}} \hat{u} - \omega_{\hat{K}} \widehat{\nabla} \times \hat{u}) \cdot \hat{q} d\hat{V} \\ &= \int_{\hat{K}} (\widehat{\nabla} \times r_{\hat{K}} \hat{u} - \widehat{\nabla} \times \hat{u}) \cdot \hat{q} d\hat{V} \\ &= \int_{\hat{K}} (r_{\hat{K}} \hat{u} - \hat{u}) \cdot \widehat{\nabla} \times \hat{q} d\hat{V} + \int_{\partial \hat{K}} (\hat{v} \times (r_{\hat{K}} \hat{u} - \hat{u})) \cdot \hat{q} d\hat{A}. \end{aligned}$$

Use of the degrees of freedom (6.7) and (6.6) shows that the right-hand side vanishes, and this completes the proof.  $\square$

Let us close by remarking that the simplest element in this family, when  $k = 1$  (see Fig. 6.2) has  $u|_K \in \mathcal{Q}_{0,1,1} \times \mathcal{Q}_{1,0,1} \times \mathcal{Q}_{1,0,0}$  so that

$$u|_K = \begin{pmatrix} a_1 + b_1 x_2 + c_1 x_3 + d_1 x_2 x_3 \\ a_2 + b_2 x_1 + c_2 x_3 + d_2 x_1 x_3 \\ a_3 + b_3 x_1 + c_3 x_2 + d_3 x_1 x_2 \end{pmatrix}.$$

The twelve degrees of freedom are given by the average of the tangential component on each edge and these determine the coefficients in the above expression for  $\boldsymbol{\pi}$ . Similar expansions are easy to write down for any  $k$ .

## 6.4 $H^1(\Omega)$ conforming elements on hexahedra

Now we continue our study of finite elements on parallelepipeds with edges parallel to the coordinate axes by describing a standard family of scalar finite elements in  $H^1(\Omega)$ . In keeping with our previous discussions of curl and divergence conforming elements, we shall not provide a great deal of detail about these elements.

**Definition 6.8** Let  $k \geq 1$ . On the reference element the gradient conforming element is defined as follows.

- (1)  $K = (0, 1)^3$ .
- (2)  $P_K = \mathcal{Q}_{k, k, k}$ .
- (3) Let  $\hat{e}$  be a general edge of  $\hat{K}$  and  $\hat{f}$  a general face. Let  $\hat{p} \in H^{3/2 + \delta}(\hat{K})$  for some  $\delta > 0$ . There are four families of degrees of freedom as follows:

(a) vertex degrees:

$$M_v(\hat{p}) = \left\{ \hat{p}(\hat{\alpha}) \text{ for the eight vertices } \hat{\alpha} \text{ of } \hat{K} \right\}; \quad (6.11)$$

(b) edge degrees:

$$M_e(\hat{p}) = \left\{ \int_{\hat{e}} \hat{p} \hat{q} d\hat{s} \text{ for all edges } \hat{e} \text{ of } \hat{K} \text{ and all } \hat{q} \in P_{k-2}(\hat{e}) \right\}; \quad (6.12)$$

(c) face degrees:

$$M_f(\hat{p}) = \left\{ \int_{\hat{f}} \hat{p} \hat{q} d\hat{A} \text{ for all faces } \hat{f} \text{ of } \hat{K} \text{ and all } \hat{q} \in \mathcal{Q}_{k-2, k-2}(\hat{f}) \right\}; \quad (6.13)$$

(d) volume degrees:

$$M_K(\hat{p}) = \left\{ \int_K \hat{p} \hat{q} d\hat{V} \text{ for all } \hat{q} \in \mathcal{Q}_{k-2, k-2, k-2} \right\}. \quad (6.14)$$

Then  $\sum_k M_k(\hat{p}) = M_v(\hat{p}) \cup M_e(\hat{p}) \cup M_f(\hat{p}) \cup M_K(\hat{p})$ .

Our first lemma proves that the element is  $H^1(\Omega)$  conforming using Lemma 5.3 in the usual way.

**Lemma 6.9** If all the degrees of freedom of a function  $\hat{p} \in \mathcal{Q}_{k, k, k}$  associated with a face  $\hat{f}$  of  $\hat{K}$  vanish (including vertices, and edges of the face) then  $\hat{p} = 0$  on  $\hat{f}$ .

**Proof** We use the fact that the vertex degrees of freedom vanish on each edge  $\hat{e}$  of  $\hat{f}$ . For example, on the edge  $\hat{x}_1 = \hat{x}_2 = 0$  we have  $\hat{p} = \hat{x}_3(1 - \hat{x}_3)r$  for some  $r \in P_{k-2}(\hat{e})$ . Choosing  $\hat{q} = r$  in the degrees of freedom (6.12) for this edge shows that  $r = 0$ .

Now using the fact that  $\hat{p} = 0$  on each edge of  $\hat{f}$ , which we assume to be the face  $\hat{x}_3 = 0$ , we have

$$\hat{p} = \hat{x}_1(1 - \hat{x}_1)\hat{x}_2(1 - \hat{x}_2)r$$

for some  $r \in Q_{k-2, k-2}(\hat{f})$ . Choosing  $\hat{q} = r$  in the degrees of freedom (6.13) shows that  $r = 0$  and hence  $\hat{p} = 0$  on  $\hat{f}$ , as required.  $\square$

Next we prove unisolvence of the element. The number of degrees of freedom and the dimension of  $Q_{k,k,k}$  are both  $(k+1)^3$  and thus it suffices to show the following result:

**Lemma 6.10** *If  $p \in Q_{k,k,k}$  and all the degrees of freedom of  $\hat{p}$  vanish, then  $\hat{p} = 0$ .*

**Proof** From the previous lemma we know that  $\hat{p} = 0$  on  $\partial K$ . Hence

$$\hat{p} = \hat{x}_1(1 - \hat{x}_1)\hat{x}_2(1 - \hat{x}_2)\hat{x}_3(1 - \hat{x}_3)r,$$

where  $r \in Q_{k-2, k-2, k-2}$ . Choosing  $\hat{q} = r$  in (6.14) proves  $r = 0$ , as required.  $\square$

The finite element on a general element  $K$  can be obtained by mapping using the diagonal affine map  $F_K$  via  $p \circ F_K = \hat{p}$  in the usual way.

Using the degrees of freedom (5.51)–(5.54) transformed on  $K$  we can define an interpolant

$$\pi_K: H^{3/2+\delta}(K) \rightarrow Q_{k,k,k}$$

by requiring the degrees of freedom of  $\pi_K p - p$  vanish. The global interpolant  $\pi p$  is then defined element-wise by  $\pi p|_K = \pi_K p$  for all elements in the mesh.

Using the same proof as for Theorem 5.48, we have the following result:

**Theorem 6.11** *The estimate of Theorem 5.48 holds for the element of this section.*

We can summarize the space as follows: (6.15)

$$U_h = \{p_h \in H_1(\Omega) | p_h|_K \in Q_{k,k,k} \text{ for all } K \in \mathcal{T}_h\}.$$

We have the following relation with  $V_b$  defined in (6.8).

**Theorem 6.12** *If  $U_b$  is defined by (6.15) and  $V_b$  by (6.8) then  $\nabla U_b \subset V_b$ . In addition, if  $p$  is such that  $r_b p$  and  $\pi_b p$  are defined then  $\nabla \pi_b p = r_b \nabla p$ .*

**Proof** Clearly, if  $p_b \in U_b$  then  $\nabla p_b \in H(\text{curl}; \Omega)$ ; we see directly that  $\nabla p_b \in Q_{k-1, k, k} \times Q_{k, k-1, k} \times Q_{k, k, k-1}$ , so  $\nabla p_b \in V_b$ .

To prove the commuting property, we map to the reference element and show that all degrees of freedom of type (6.5)–(6.7) vanish for  $\widehat{\nabla} \pi_{\hat{K}} \hat{p} - r_{\hat{K}} \widehat{\nabla} \hat{p}$ . Then, via Lemma 6.10, we conclude that  $\widehat{\nabla} \pi_{\hat{K}} \hat{p} - r_{\hat{K}} \widehat{\nabla} \hat{p} = 0$ .

For the edge degrees of freedom (6.5), if  $\hat{\tau}$  is tangent to  $\hat{e} = [\hat{a}, \hat{B}]$  and if  $\hat{q} \in P_{k+1}(\hat{e})$  then using (6.5) and integration by parts we have

$$\begin{aligned} & \int_{\hat{e}} (\widehat{\nabla} \pi_{\hat{K}} \hat{p} - r_{\hat{K}} \widehat{\nabla} \hat{p}) \cdot \hat{q} d\hat{s} \\ &= \int_{\hat{e}} (\widehat{\nabla} \pi_{\hat{K}} \hat{p} - \widehat{\nabla} \hat{p}) \cdot \hat{q} d\hat{s} \\ &= \int_{\hat{e}} \frac{\partial}{\partial \hat{s}} (\pi_{\hat{K}} \hat{p} - \hat{p}) \hat{q} d\hat{s} \\ &= (\pi_{\hat{K}} \hat{p} - \hat{p})(\hat{b}) - (\pi_{\hat{K}} \hat{p} - \hat{p})(\hat{a}) - \int_{\hat{e}} \frac{\partial \hat{q}}{\partial \hat{s}} (\pi_{\hat{K}} \hat{p} - \hat{p}) d\hat{s}. \end{aligned}$$

Since  $\partial \hat{q} / \partial \hat{s} \in P_{k+2}(\hat{e})$  using (6.11) and (6.12) we conclude that the right-hand side above vanishes.

The face degrees (6.6) and volume degrees (6.7) are treated in the same way. We do not give the details.  $\square$

Finally, by mapping to the reference cube and using the Deny–Lions Theorem 5.5 we can verify the following theorem.

**Theorem 6.13** *Assume that the mesh  $\tau_h$ ,  $h > 0$ , is regular and the elements in the mesh are parallelepipeds with edges parallel to the coordinate axes. Then the error estimates of Theorem 5.48 hold for the hexahedral elements discussed in this section.*

## 6.5 An $L^2(\Omega)$ conforming space and a boundary space

As for the elements on tetrahedra considered in the previous chapter we can complete the de Rham diagram by defining

$$Z_h = \left\{ p_h \in L^2(\Omega) \mid p_h|_K \in Q_{k-1,k-1,k-1} \text{ for all } K \in \mathcal{T}_h \right\},$$

and define  $P_{0,b}$  to be the  $L^2(\Omega)$  orthogonal projection into  $Z_h$ . It is then clear, using the same argument as in Section 5.7, that  $\nabla \cdot w_b u = P_{0,b} \nabla \cdot u$  for any  $u$  for which both sides are well defined.

Using this result and Theorems 6.7 and 6.12 proves that the discrete de Rham diagram in (5.59) also holds for the finite elements on hexahedra in this chapter. Also, the same error estimates hold, so the accuracy properties of both spaces are asymptotically the same.

In the same way as in Section 5.8, the volume mesh induces a regular mesh, denoted by  $\tau_h(\partial\Omega)$ , on the boundary of  $\Omega$ . In this case the edges of the mesh are parallel to the coordinate axes. Hence the same mapping argument shows that Lemma 5.53 holds in this case.

If the boundary mesh is quasi-uniform on some component  $\Sigma$  of the boundary, the same arguments show that Lemmas 5.55 and 5.57 also hold.

# 7 FINITE ELEMENTS METHODS FOR THE CAVITY PROBLEM

## 7.1 Introduction

In Chapter 4 we showed that a standard Galerkin formulation for the Maxwell problem in a cavity has a unique solution. In performing this analysis, we encountered various function spaces (in particular, the spaces  $X$  and  $S$  defined in (4.3) and (4.6)). Then in Chapters 5 and 6 we saw how these spaces can be discretized using finite elements. In this chapter we shall see how these two themes can be combined to produce a finite element approximation of Maxwell's equations posed on a bounded domain. As we saw in Chapter 4, there are two possibilities: for a given wavenumber  $\varkappa > 0$  there may be exactly one solution to the cavity problem or there may be non-trivial solutions to the homogeneous problem. We shall analyze both cases.

The obvious choice of using vector continuous piecewise-linear elements is dangerous since, if the domain has re-entrant corners, it is possible to compute finite element solutions that converge to a field that is not a solution of Maxwell's equations [105]. For this model problem, modifications to the bilinear form to restore convergence are given in [113, 114], but further modifications are needed to handle, for example, discontinuous coefficients. We prefer to use the edge finite elements of Nédélec [233]. These avoid the problem of spurious solutions at the cost of increased complexity. Furthermore, these elements can be applied to problems involving discontinuous coefficients (modeling different media) without modification. Indeed shortly after the publication of Nédélec's paper, these elements started to be used in engineering codes [55].

In presenting the analysis there seem to be two possible paths. Either one can use a special theory of mixed methods developed by Daniele Boffi and his co-workers [48, 47, 46, 50] to handle Maxwell's equations, or one can use discrete analogs of compactness arguments to derive the theory. The latter approach, which we shall follow, fits better with our approach in Chapter 4 of using compactness to analyze the continuous system. In fact, Boffi [46] has shown that the approach we shall follow here, using the discrete compactness concept of Kikuchi [183–185], is equivalent to his approach.

Within the general approach of discrete analogs of compactness, we shall present two convergence proofs. The first proof from [222] uses duality theory along the lines of [267] and is rather specialized, in that we need to simplify the cavity problem in order to apply the proof (it is likely that this proof can be extended to more general cases, but that would require a better understanding of Maxwell's equations). The advantage of the approach is that the argument

is relatively simple, and the results are rather precise. The proof is based on early work in [217], convergence is proved using duality via the ideas of Schatz [267] concerning the compact perturbation of coercive bilinear forms. Due to limitations on the understanding of edge elements and the regularity theory for Maxwell's equations at that time,  $\Omega$  was assumed to be convex, and the mesh was quasi-uniform. Here we lift these restrictions.

The second approach uses the theory of collectively compact operators to produce a general convergence theory applicable for a general class of coefficients and the general domain considered in Section 4. This approach was implicit in the work of Kikuchi [185] and first explicitly suggested in [38, 178] for the analysis of waveguide problems. In [119], Demkowicz and I applied the theory of collectively compact operators to prove convergence on general Lipschitz polyhedra. We assumed quasi-uniformity of the mesh to provide a certain inverse inequality which, as we shall see, is not necessary. Moreover, using the results of [71], our proof extends to include rather general spatially dependent coefficients in the equations (e.g. piecewise-constant coefficients). The drawback of this approach is its complexity.

A third compactness-based proof due to Hiptmair [164] will not be considered in this section. Instead, we shall use a similar approach when we prove the convergence of edge element methods for the full scattering problem in Chapter 10 .

A fourth and different approach is due to Boffi and Gastaldi [50]. They use the general convergence theory of Rappaz [123], together with their estimates of Maxwell eigenvalue convergence, to prove convergence on general regular meshes. This is the approach also suggested in [122], where it is shown that the convergence of the discrete eigenvalues is equivalent to the convergence of the source problem considered here (the proof can then be completed using the results in [49]).

Let us now discuss the finite element method for the discretization of the general problem (4.4). We suppose that the interior problem has a unique solution because either

- (1)  $\Sigma \neq \emptyset$ , and  $\lambda$  strictly positive, or
- (2)  $\Im(\varepsilon)$  is strictly positive definite on a ball contained in  $\Omega$ , or
- (3)  $\Im(\varepsilon) = 0$  and  $\lambda = 0$  but  $\varkappa$  is not a resonance or Maxwell eigenvalue for  $\Omega$ .

In any of these cases, we know that the variational problem (4.4) has a unique solution depending continuously on the data, and thus is suitable for finite element discretization (see Chapter 4 ).

From the previous two chapters, we know that edge finite elements on tetrahedral or rectilinear hexahedral meshes satisfy the discrete de Rham diagram (5.59) and have the same interpolation error estimates. Thus, from the point of view of analysis, the choice of elements is immaterial. However, for the sake of clarity we shall assume the use of a tetrahedral mesh. In particular, we suppose  $\Omega$  has been covered by a regular mesh of tetrahedra, and that the elements presented

in Chapter 5.5 are used. Thus we define the finite element subspace of  $X$  consisting of degree- $k$  edge elements by(7.1)

$$\begin{aligned} X_h = \{ & u_h \in H(\text{curl}; \Omega) | u_h|_K \in R_K \text{ for all } K \in \mathcal{T}_h \\ & \text{and } u_h \times v = 0 \text{ on } \Gamma \}, \end{aligned}$$

where  $k > 0$  is an integer. We assume that the mesh is consistent with the coefficients  $\mu_r$  and  $\varepsilon_r$ , by which we mean that any surfaces where  $\mu_r$  or  $\varepsilon_r$  are discontinuous are also unions of faces of the mesh. On each tetrahedron  $K$  the coefficient  $\mu_r$  is constant and the coefficient  $\varepsilon_r$  is an  $H^1(K)$  function of position, so it is continuous (more precisely, it satisfies the conditions in Section 4.2).

Given  $F \in (L^2(\Omega))^3$  and  $g \in L^2_t(\Sigma)$ , we seek to approximate the solution  $E \in X$  of (4.4) by finding  $E_h \in X_h$  such that(7.2)

$$\begin{aligned} (\mu_r^{-1} \nabla \times E_h, \nabla \times \varphi_h) - K^2 (\epsilon_r E_h, \varphi_h) - ik \langle \lambda E_{h,T}, \varphi_h, T \rangle \\ = (F, \varphi_h) + \langle g, \varphi_{h,T} \rangle \text{ for all } \varphi_h \in X_h. \end{aligned}$$

This variational problem should be compared with (4.4). Now we want to prove that (7.2) has a unique solution which approximates the solution  $E$  of (4.4) in a quasi-optimal way.

## 7.2 Error analysis via duality

In this section we give a simple proof of convergence of edge finite element approximations to the cavity problem for Maxwell's equations. This analysis is from [222] and is motivated by the work of Gopalakrishnan and Pasciak [147], who use similar estimates in their analysis of Schwarz methods for Maxwell's equations. It is based on the use of solutions to the dual variational problem.

Unfortunately, in order to use the duality theory, we need to simplify (7.2). In particular, for this section, we assume  $\varepsilon_r = \mu_r = 1$  in  $\Omega$ , and, in addition, the boundary of  $\Omega$  is assumed to have just one component  $\Gamma$  (thus, the boundary condition is perfectly conducting). With these simplifications, we wish to approximate the electric field  $E$  that satisfies the Maxwell equations(7.3a)

$$\begin{aligned} \nabla \times (\nabla \times E) - \kappa^2 E &= F \text{ in } \Omega, \\ v \times E &= 0 \text{ on } \Gamma = \partial\Omega. \end{aligned} \tag{7.3b}$$

As usual,  $F$  is a given function related to the imposed current sources and the parameter  $\kappa$  is the wavenumber assumed to be real and positive. Equation (7.3b) specifies a standard perfectly conducting boundary condition on the boundary of  $\Omega$ .

In this case, the space  $X$  given in (4.3) simplifies to  $X = H_0(\text{curl}; \Omega)$  and  $S$  given in (4.6) simplifies to  $S = H_0^1(\Omega)$ .

With the above simplifications, problem (4.4) becomes the problem of finding  $E \in H_0(\text{curl}; \Omega)$  such that(7.4)

$$(\nabla \times E, \nabla \times \varphi) - \kappa^2(E, \varphi) = (F, \varphi) \quad \text{for all } \varphi \in H_0(\text{curl}; \Omega).$$

Because  $\kappa$  is real, we can assume that any solution of this problem is real, so all spaces and functions in this section are real. From Chapter 4 we know that this problem has a unique solution unless  $\kappa$  is an interior Maxwell eigenvalue for  $\Omega$ . We assume this is not the case in this section.

The problem of approximating  $E$  by finite elements then reduces to using the finite element space  $X_h$  defined in (7.1) and computing  $E_h \in X_h$  such that the following simplified version of (7.2) is satisfied:(7.5)

$$(\nabla \times E_h, \nabla \times \varphi_h) - \kappa^2(E_h, \varphi_h) = (F, \varphi_h) \quad \text{for all } \varphi_h \in X_h.$$

The remainder of this section is devoted to proving the following theorem.

**Theorem 7.1** *Let  $\Omega$  be a simply connected Lipschitz polyhedron with connected boundary  $\Gamma$ . Let  $\tau_h$  be a regular mesh and suppose  $X_h$  is given by (7.1). In addition, suppose  $\kappa$  is not a Maxwell eigenvalue for  $\Omega$ . Then if  $E$  satisfies(7.4)and  $E_h \in X_h$  satisfies (7.5), there is a constant  $C$  independent of  $h$ ,  $E$  and  $E_h$  and a constant  $b_0 > 0$  independent of  $E$  and  $E_h$  such that, for all  $0 < h < b_0$ ,(7.6)*

$$\|E - E_h\|_{H(\text{curl}; \Omega)} \leq \frac{1}{1 - Ch^{1/2+\delta}} \inf_{v_h \in X_h} \|E - v_h\|_{H(\text{curl}; \Omega)}.$$

Here  $\delta > 0$  is the exponent in Lemma 7.6.

**Remark 7.2** *Choosing  $h$  small enough that (for example)  $Ch^{1/2+\delta} < \frac{1}{2}$  proves quasi-optimal convergence of the edge element approximation. Furthermore, the constant  $1 / (1 - Ch^{1/2+\delta})$  can be made arbitrarily close to unity. Note that we do not use  $\kappa$ -dependent norms, but of course the constant  $C$  depends on  $\kappa$  via the a priori estimate for the dual problem. Later, in Section 13.3, we shall see how this  $\kappa$  dependence can be included in the estimate.*

If  $u \in H^s(\text{curl}; \Omega)$  for some  $s$  with  $\frac{1}{2} < s \leq K$ , then Theorems 5.41 and 7.1 show that for all sufficiently small  $h$  there is a constant  $C$  such that  $\|E - E_h\|_{H(\text{curl}; \Omega)} \leq Ch$ . In general, the polyhedral boundary  $\Gamma$  causes singularities in the solution that prevent high global regularity [44, 106]. Nevertheless, as we have seen, we can expect sufficient regularity to guarantee a convergence rate of better than  $O(h^{1/2})$ . The mesh only needs to be regular so that it can be refined strongly near boundary singularities in  $E$ . Nicaise has shown in detail how to this near an edge [238].

Assuming Theorem 7.1 is proved, we then have the following corollary.

**Corollary 7.3** *For any  $F \in (L^2(\Omega))^3$ , there is an  $b_0 > 0$  such that, for all  $h < b_0$ , eqn(7.5)has a unique solution.*

**Proof** It suffices to prove uniqueness. Let  $F = 0$ . Then, since  $\nu$  is not a Maxwell eigenvalue,  $E = 0$  in (7.4) and  $E_b = 0$  is one solution of the discrete problem. By the error estimate (7.6), for any solution  $E_b$  of the discrete problem, we have the estimate  $\|E_b\|_{H(\text{curl};\Omega)} \leq C \inf_{v_h \in X_h} \|v_h\|_{H(\text{curl};\Omega)} = 0$ . Hence  $E_b = 0$ , and uniqueness is proved.  $\square$

### 7.2.1 The discrete Helmholtz decomposition

For the simplified problem considered in this section (see (7.5)), we have already commented that  $S = H_0^1(\Omega)$  and so we take (7.7)

$$S_h = \left\{ p_h \in H_0^1(\Omega) \mid p_h|_K \in P_K \text{ for all } K \in \mathcal{T}_h \right\}.$$

It follows from (5.59) that  $\nabla S_b \subset X_b$ . Thus  $\nabla S_b$  provides a large subspace of test functions in  $X_b$ . Using this space, we say a function  $u \in (L^2(\Omega))^3$  is *discrete divergence-free* if

$$(u, \nabla \xi_h) = 0 \text{ for all } \xi_h \in S_h.$$

We then have the following discrete Helmholtz decomposition analogous to (4.7)(7.8)

$$X_h = X_{0,h} \oplus \nabla S_h,$$

where  $X_{0,b}$  is the space of discrete divergence-free finite elements. In other words, (7.9)

$$X_{0,h} = \{ u_h \in X_h \mid (u_h, \nabla \xi_h) = 0 \text{ for all } \xi_h \in S_h \},$$

Now let  $Y_b$  denote the following space of degree- $k$  divergence conforming finite elements (see Section 5.4):

$$Y_h = \{ z_h \in H_0(\text{div};\Omega) \mid z_h|_K \in D_K \} \subset W_h.$$

First, we note that via (5.59) and taking into account the boundary conditions we have  $\nabla \times X_b \subset Y_b$ . Thus, as in [18], we can regard the curl as a bounded operator from  $X_b$  into  $Y_b$ . We denote the null-space of the curl operator in  $X_b$  by  $N_b(\text{curl})$ . Let  $u_b \in N_b(\text{curl})$ . Since the domain  $\Omega$  is simply connected and the boundary  $\Gamma$  is connected, the fact that  $\nabla \times u_b = 0$  in  $\Omega$  implies  $u_b = \nabla p$  for some  $p \in H_0^1(\Omega)$ . In addition since  $u_b \in X_b$ , we know that  $p \in S_b$ . Hence in  $X_b$  the null-space of the curl is given by  $N_b(\text{curl}) = \nabla S_b$ .

The discrete divergence-free space  $X_{0,b}$  is thus given by  $X_{0,b} = N_b(\text{curl})^\perp$ , where  $N_b(\text{curl})^\perp$  is the orthogonal complement of  $N_b(\text{curl}) \subset X_b$  in the  $(L^2(\Omega))^3$  inner product. Now, following [18], let  $\nabla_b \times$  denote the discrete adjoint operator for the curl by which we mean that for each  $z_b \in X_b$ , the function  $\nabla_b \times z_b \in X_b$  is the unique function such that

$$(\nabla_b \times z_b, \psi_h) = (z_b, \nabla \times \psi_h) \text{ for all } \psi_h \in X_h.$$

By a standard theorem from functional analysis, Theorem 2.15, we know that

$$N_b(\text{curl})^\perp = \nabla_b \times (\nabla \times X_b),$$

so that we have the following result.

**Lemma 7.4** For each  $v_b \in X_{0,b}$  there is a function  $\zeta_b \in \nabla \times X_b \subset Y_b$  such that  $v_b = \nabla b \times \zeta_b$ , in the sense that

$$(v_h, \varphi_h) = (z_h, \nabla \times \varphi_h) \quad \text{for all } \varphi_h \in X_h.$$

This lemma was first given in [18] where it is pointed out that an alternative way to write the discrete Helmholtz decomposition is as follows. Any function  $v_b \in X_b$  may be written as

$$v_h = \nabla_h \times z_h + \nabla p_h$$

for some  $\zeta_b \in \nabla \times X_b \subset Y_b$  and  $p_b \in S_b$ .

Next we need to define the  $H_0(\text{curl}; \Omega)$  orthogonal projection. This is denoted  $P_b: H_0(\text{curl}; \Omega) \rightarrow X_b$ , and is such that if  $u \in H_0(\text{curl}; \Omega)$  then  $P_b u \in X_b$  satisfies(7.10)

$$(\nabla \times (u - P_b u), \nabla \times \varphi_h) + ((u - P_b u), \varphi_h) = 0 \quad \text{for all } \varphi_h \in X_h.$$

Cea's Lemma 2.37 shows that this projection satisfies the optimal error estimate

$$\|u - P_b u\|_{H(\text{curl}; \Omega)} = \inf_{v_h \in X_h} \|u - v_h\|_{H(\text{curl}; \Omega)}.$$

If  $u \in H^s(\text{curl}; \Omega)$ ,  $s > \frac{1}{2}$ , Theorem 5.41 can then be used to provide order estimates for the right-hand side of the above equality.

Using the test function  $\varphi_b = \nabla \xi_b$ , for some  $\xi_b \in S_b$ , in (7.10) shows that  $u - P_b u$  is discrete divergence-free since(7.11)

$$((u - P_b u), \nabla \xi_b) = 0 \quad \text{for all } \xi_b \in S_b.$$

## 7.2.2 Preliminary error analysis

This section is devoted to proving two lemmas that will be used in the proof of the our main theorem (Theorem 7.1).

Under the assumptions of this section, the general sesquilinear form  $a(\cdot, \cdot)$  defined in (4.5) reduces to

$$a(u, \varphi) = (\nabla \times u, \nabla \times \varphi) - \kappa^2(u, \varphi).$$

At this stage, we do not know that  $E_b$  exists, but if it does exist we define  $e_b = E - E_b$ . Then, by subtracting (7.5) from (7.4), we obtain the Galerkin error equation,(7.12)

$$a(E_h, \psi_h) = 0 \quad \text{for all } \psi_h \in X_h.$$

In particular, choosing  $\psi_b = \nabla \xi_b$  for some  $\xi_b \in S_b$  shows that  $e_b$  is discrete divergence-free.

In [217] the problem of estimating  $\|E - E_b\|_{H(\text{curl}; \Omega)}$  was approached via a classical Gårding inequality. Our first lemma is a weaker form of the Gårding inequality as used in [147].

**Lemma 7.5** There is a constant  $C$  independent of  $h$ ,  $E$  and  $E_h$  such that (7.13)

$$\|e_h\|_{H(\text{curl}; \Omega)} \leq \|E - P_h E\|_{H(\text{curl}; \Omega)} + C \sup_{v_h \in X_h} \frac{|(e_h, v_h)|}{\|v_h\|_{H(\text{curl}; \Omega)}}.$$

**Proof** Using a very slight modification of the proof of Lemma 4.4 of [147] we see that by the definition of the curl norm and the definition of  $a(\cdot, \cdot)$  we have

$$\begin{aligned} \|e_h\|_{H(\text{curl}; \Omega)}^2 &= a(e_h, e_h) + (1 + \kappa^2)(e_h, e_h) \\ &= a(e_h, E - P_h E) + a(e_h, P_h E - E_h) + (1 + \kappa^2)(e_h, e_h). \end{aligned}$$

Now using the Galerkin condition (7.12), the definition of the curl norm, and the definition of  $\|\cdot\|_{H(\text{curl}; \Omega)}$  we have

$$\begin{aligned} \|e_h\|_{H(\text{curl}; \Omega)}^2 &= a(e_h, E - P_h E) + (1 + \kappa^2)(e_h, e_h) \\ &= (\nabla \times e_h, \nabla \times (E - P_h E)) + (e_h, (E - P_h E)) \\ &\quad + (1 + \kappa^2) \{ (e_h, e_h) - (e_h, (E - P_h E)) \} \\ &= (\nabla \times e_h, \nabla \times (E - P_h E)) + (e_h, (E - P_h E)) \\ &\quad + (1 + \kappa^2)(e_h, P_h E - E_h). \end{aligned}$$

Hence, using the Cauchy–Schwarz inequality and the boundedness of the projection  $P_h: H(\text{curl}; \Omega) \rightarrow X_h$

$$\begin{aligned} \|E_h\|_{H(\text{curl}; \Omega)}^2 &\leq \|E - P_h E\|_{H(\text{curl}; \Omega)} \|E_h\|_{H(\text{curl}; \Omega)} \\ &\quad + (1 + \kappa^2) \sup_{v_h \in X_h} \frac{|(E_h, v_h)|}{\|v_h\|_{H(\text{curl}; \Omega)}} \|P_h E - E_h\|_{H(\text{curl}; \Omega)} \\ &= \|E - P_h E\|_{H(\text{curl}; \Omega)} \|E_h\|_{H(\text{curl}; \Omega)} \\ &\quad + (1 + \kappa^2) \sup_{v_h \in X_h} \frac{\|E_h, v_h\|}{\|v_h\|_{H(\text{curl}; \Omega)}} \|P_h E_h\|_{H(\text{curl}; \Omega)} \\ &\leq \|E - P_h E\|_{H(\text{curl}; \Omega)} \|E_h\|_{H(\text{curl}; \Omega)} + \\ &\quad (1 + \kappa^2) \sup_{v_h \in X_h} \frac{\|E_h, v_h\|}{\|v_h\|_{H(\text{curl}; \Omega)}} \|E_h\|_{H(\text{curl}; \Omega)}. \end{aligned}$$

This proves the desired estimate with  $C = 1 + \kappa^2$ .  $\square$

Our error estimate will be completed if we can estimate the supremum on the right-hand side of (7.13). This is done in Lemma 7.7. Before we prove this lemma, we need to investigate discrete divergence-free functions in more detail. For such functions we can construct a nearby exactly divergence-free function. This construction was used, for example, by Girault [142] and myself [217] with an *ad hoc* analysis. The solution operator for (7.14) which maps a discrete divergence free vector to its divergence free component is called the *Hodge operator* by Hiptmair [164]. Building on the work of Hiptmair, the clearest analysis is from Arnold *et al.* [18].

For a given discrete divergence-free function  $v_b \in X_{0,b}$ , let us define  $v^b \in H_0(\text{curl}; \Omega)$  by(7.14a)

$$\begin{aligned} \nabla \times v^b &= \nabla \times v_h \quad \text{in } \Omega, \\ \nabla \cdot v^b &= 0 \quad \text{in } \Omega. \end{aligned} \tag{7.14b}$$

Note that  $v^b$  is the divergence-free component of  $v_b$  in the Helmholtz decomposition.

In [18] it is suggested to view the solution  $v^b$  of (7.14) as part of the solution of the mixed problem of finding  $v^b \in H_0(\text{curl}; \Omega)$  and  $\zeta^b \in \nabla \times H_0(\text{curl}; \Omega)$  such that(7.15a)

$$\begin{aligned} (v^b, \varphi) + (\nabla \times \varphi, z^b) &= 0 \quad \text{forall } \varphi \in H_0(\text{curl}; \Omega), \\ (\nabla \times v^b, \xi) &= (\nabla \times v_h, \xi) \quad \text{forall } \xi \in \nabla \times H_0(\text{curl}; \Omega). \end{aligned} \tag{7.15b}$$

Both the coercivity condition and Babuška–Brezzi condition for mixed methods are obviously satisfied and so  $(v^b, \zeta^b)$  exists. Thus we have the following lemma:

**Lemma 7.6** *Let  $v_b \in X_{0,b}$ . Suppose  $v^b \in H_0(\text{curl}; \Omega)$  satisfies (7.14). Then there are constants  $C$  and  $\delta > 0$  independent of  $b$  and  $v_b$  and  $v^b$ , such that*

$$\|v^b - v_h\|_{(L^2(\Omega))^3} \leq Ch^{1/2+\delta} \|\nabla \times v_h\|_{(L^2(\Omega))^3}.$$

**Proof** The proof follows [18]. From the characterization of  $v^b$  we see that  $v^b \in X_N$  and hence, by Theorem 3.50, there is an exponent  $\delta > 0$  such that  $v^b \in (H^{1/2+\delta}(\Omega))^3$ , and since  $\nabla \times v^b = \nabla \times v_b$ , we see that  $\nabla \times v^b \in W_b \subset (L^p(\Omega))^3$  for  $p > 2$ . Hence, using Lemma 5.38, the edge finite element interpolant  $r_b v^b$  is well defined. But then, using the commuting diagram property of edge elements (5.59), if  $w_b$  is the divergence conforming element interpolation operator,(7.16)

$$\nabla \times r_b v^b = \omega_b \nabla \times v^b = \omega_b \nabla \times v_h = \nabla \times v_h.$$

Since  $v_b$  is discrete divergence-free, by Lemma 7.4 there is a function  $\zeta_b \in \nabla \times X_b$  such that(7.17a)

$$\begin{aligned} (v_h, \varphi_h) + (\nabla \times \varphi_h, z_h) &= 0 \quad \text{for all } \varphi_h \in X_h, \\ (\nabla \times v_h, \xi_h) &= (\nabla \times v_h, \xi_h) \quad \text{for all } \xi_h \in \nabla \times X_h. \end{aligned} \tag{7.17b}$$

$$(\nabla \times v_h, \xi_h) = (\nabla \times v_h, \xi_h) \quad \text{for all } \xi_h \in \nabla \times X_h.$$

Of course the second equation above is trivially satisfied! Thus  $(v_b, \zeta_b)$  is nothing else than the mixed finite element approximation to  $(v^b, \zeta^b)$  defined by (7.15). Now, selecting  $\varphi = r_b v^b - v_b$  in (7.15a) and  $\varphi_b = r_b v^b - v_b$  in (7.17a) and using the fact that  $\nabla \times \varphi_b = 0$  (see(7.16)) we have  $(v^b - v_b, r_b v^b - v_b) = 0$ . Thus

$$(v^b - v_h, v^b - v_h) = (v^b - v_h, v^b - r_b v^b) + (v^b - v_h, r_b v^b - v_h).$$

Hence  $\|v^b - v_b\|_{(L^2(\Omega))^3} \leq \|v^b - r_b v^b\|_{(L^2(\Omega))^3}$  and, using Lemma 5.38, we have

$$\|v^h - v_h\|_{(L^2(\Omega))^3} \leq C \left( h^{1/2+\delta} \|v^h\|_{(H^{1/2+\delta}(\Omega))^3} + h \|\nabla \times v_h\|_{(L^2(\Omega))^3} \right).$$

The *a priori* estimate  $\|v^h\|_{(H^{1/2+\delta}(\Omega))^3} \leq C \|\nabla \times v_h\|_{(L^2(\Omega))^3}$  completes the proof.  $\square$

### 7.2.3 Duality estimate

Now we can estimate the troublesome term in (7.13).

**Lemma 7.7** *For all  $h$  small enough, there exist constants  $C$  and  $\delta$  with  $0 < \delta \leq 1/2$  such that*

$$\sup_{v_h \in X_h} \frac{|\langle E_h, v_h \rangle|}{\|v_h\|_{H(\text{curl}; \Omega)}} \leq Ch^{\delta+1/2} \|E_h\|_{H(\text{curl}; \Omega)}.$$

**Proof** This lemma is proved by a duality argument similar to the one in the proof of Lemma 4.3 of [147] and the duality argument in [217]. Using the continuous Helmholtz decomposition, there is a divergence-free function  $e_0^h \in H_0(\text{curl}; \Omega)$  and a scalar  $p^h \in H_0^1(\Omega)$  such that  $e_h = e_0^h + \nabla p^h$ . Here  $p^h \in H_0^1(\Omega)$  satisfies

$$(\nabla p^h, \nabla \xi) = (e_h, \nabla \xi) \quad \text{for all } \xi \in H_0^1(\Omega).$$

Thus, by choosing  $\xi = p^h$ , we see that  $\|\nabla p^h\|_{(L^2(\Omega))^3} \leq \|e_h\|_{(L^2(\Omega))^3}$ .

Using the discrete Helmholtz decomposition, we also can write  $v_h = v_{0,h} + \nabla \xi_h$  for some  $v_{0,h} \in X_{0,h}$  and  $\xi_h \in S_h$ . Since we have already shown that  $e_h$  is discrete divergence-free, we have (7.18)

$$(e_h, v_h) = (e_h, v_{0,h}) = (e_0^h, v_{0,h}) + (\nabla p^h, v_{0,h}).$$

The first term on the right-hand side is estimated by (7.19)

$$|(e_0^h, v_{0,h})| \leq \|e_0^h\|_{(L^2(\Omega))^3} \|v_{0,h}\|_{(L^2(\Omega))^3} \leq C \|e_0^h\|_{(L^2(\Omega))^3} \|v_h\|_{(L^2(\Omega))^3},$$

where we have made use of the fact that  $\|\nabla \xi_h\|_{(L^2(\Omega))^3} \leq \|v_h\|_{(L^2(\Omega))^3}$ . Thus we can estimate this term by estimating  $\|e_0^h\|_{(L^2(\Omega))^3}$ , which we do next.

We define the adjoint variable  $\zeta \in H_0(\text{curl}; \Omega)$  (unrelated to  $\zeta_b$  and  $\zeta^b$  in the previous section!) such that (7.20)

$$a(\varphi, z) = (e_0^h, \varphi) \quad \text{for all } \varphi \in H_0(\text{curl}; \Omega).$$

Clearly,  $\zeta$  is the weak solution in  $H_0(\text{curl}; \Omega)$  of

$$\nabla \times \nabla \times z - \kappa^2 z = E_0^h \quad \text{in } \Omega,$$

and the assumption that  $\kappa$  is not an interior Maxwell eigenvalue implies that  $\zeta$  is well defined and there is a constant  $C$  such that  $\|\zeta\|_{H(\text{curl}; \Omega)} \leq C \|E_0^h\|_{(L^2(\Omega))^3}$  (see Corollary 4.19).

Since  $E_0^h$  is divergence-free, it follows that  $\zeta$  is also divergence-free (to see this, take  $\varphi = \nabla \xi$  for  $\xi \in H_0^1(\Omega)$  in eqn (7.20)). Thus we have

$$\nabla \times z \in \left(L^2(\Omega)\right)^3, \quad \nabla \cdot z = 0 \text{ in } \Omega \text{ and } v \times z = 0 \text{ on } \Gamma.$$

Hence  $\zeta \in X_N$  and, by Theorem 3.50 and the remark after Corollary 3.51, we have  $\zeta \in (H^{1/2+\delta}(\Omega))^3$  for some  $\delta$  with  $0 < \delta \leq 1/2$  together with the norm bound  $\|\zeta\|_{(H^{1/2+\delta}(\Omega))^3} \leq C \|E_0^h\|_{(L^2(\Omega))^3}$ . In addition, we see that  $\nabla \times \zeta \in (L^2(\Omega))^3$  is the weak solution of

$$\begin{aligned} \nabla \times (\nabla \times z) &= -\kappa^2 z + e_0^h \quad \text{in } \left(L^2(\Omega)\right)^3, \\ \nabla \cdot (\nabla \times z) &= 0 \quad \text{in } \Omega, \\ v \cdot (\nabla \times z) &= 0 \quad \text{on } \Gamma. \end{aligned}$$

Thus  $\nabla \times \zeta \in X_T$  and again, by Theorem 3.50, we know that

$$\nabla \times z \in \left(H^{1/2+\delta}(\Omega)\right)^3$$

with the norm bound  $\|\nabla \times z\|_{(H^{1/2+\delta}(\Omega))^3} \leq C \|E_0^h\|_{(L^2(\Omega))^3}$ . We conclude that  $\zeta \in H^{1/2+\delta}(\text{curl}; \Omega)$ . Hence, by Lemma 5.38, the interpolant  $r_h \zeta$  is well defined, and we can use Theorem 5.41 to obtain the error estimate

$$\|z - P_h z\|_{H(\text{curl}; \Omega)} \leq \|z - r_h z\|_{H(\text{curl}; \Omega)} \leq Ch^{1/2+\delta} \|E_0^h\|_{(L^2(\Omega))^3}.$$

Now using (7.20) and the fact that  $\zeta$  is divergence-free, we have

$$\|e_0^h\|_{(L^2(\Omega))^3}^2 = a(e_0^h, z) = a(e_0^h + \nabla p^h, z) = a(e_h, z).$$

Then, by the Galerkin condition (7.12), and the above estimate for  $\zeta - P_h \zeta$

$$\begin{aligned} \|e_0^h\|_{(L^2(\Omega))^3}^2 &= a(e_h, z - P_h z) \\ &\leq C \|e_h\|_{H(\text{curl}; \Omega)} \|z - P_h z\|_{H(\text{curl}; \Omega)} \\ &\leq Ch^{1/2+\delta} \|E_0^h\|_{(L^2(\Omega))^3} \|E_h\|_{H(\text{curl}; \Omega)}. \end{aligned}$$

We have thus proved that (7.21)

$$\|e_0^h\|_{(L^2(\Omega))^3}^2 \leq Ch^{1/2+\delta} \|e_h\|_{H(\text{curl}; \Omega)}.$$

Now we estimate the term  $(\nabla p^h, v_{0,h})$  in (7.18). Since, by construction,  $v_{0,h}$  is discrete divergence-free, Lemma 7.6 implies that there is a divergence-free function  $v_0^h \in H(\text{curl}; \Omega)$  with

$$\begin{aligned} \|v_0^h - v_{0,h}\|_{(L^2(\Omega))^3} &\leq Ch^{1/2+\delta} \|\nabla \times v_{0,h}\|_{(L^2(\Omega))^3} \\ &= Ch^{1/2+\delta} \|\nabla \times v_h\|_{(L^2(\Omega))^3}. \end{aligned}$$

Now using the fact that  $v_0^h$  is divergence-free, and using the error estimate above, we have(7.22)

$$\begin{aligned} (\nabla p^h, v_{0,h}) &= (\nabla p^h, v_{0,h} - v_0^h) \\ &\leq Ch^{1/2+\delta} \|\nabla p^h\|_{(L^2(\Omega))^3} \|\nabla \times v_h\|_{(L^2(\Omega))^3}. \end{aligned}$$

Using (7.21) in (7.19) and using the resulting estimate together with (7.22) in (7.18) proves the desired result.  $\square$

We now prove our main theorem.

**Proof of Theorem 7.1** Lemma 7.7 shows that

$$\sup_{v_h \in X_h} \frac{|(e_h, v_h)|}{\|v_h\|_{H(\text{curl}; \Omega)}} \leq Ch^{1/2+\delta} \|e_h\|_{H(\text{curl}; \Omega)}.$$

Putting this together with (7.13) shows that

$$\|e_h\|_{H(\text{curl}; \Omega)} \leq \|E - P_h E\|_{H(\text{curl}; \Omega)} + Ch^{1/2+\delta} \|e_h\|_{H(\text{curl}; \Omega)}.$$

Choosing  $h$  small enough that  $1 - Ch^{1/2+\delta} > 0$  proves the result.  $\square$

Our proof rests critically on regularity results for the dual problem and on the estimate in Lemma 7.4 for the approximation of a discrete divergence-free function by a divergence-free function. For smooth coefficients these results still hold. For general coefficients  $\varepsilon$  and  $\mu$ , both results might be difficult to obtain. However, it is possible that using arguments like those in [71], the explicit estimates used here could be replaced by uniform convergence estimates based on the compactness arguments of the type used by Schatz and Wang [268]. Indeed, this is essentially the approach taken in the next section.

## 7.3 Error analysis via collective compactness

In this section we shall apply the theory of convergence of collectively compact operators to prove convergence of the solution of the general problem (7.2) to the solution of (4.4). This allows us to return to the full generality of the problem discussed in Section 7.1. However, we shall need to restrict the class of finite element meshes. In particular, we need to assume that the mesh  $\tau_b$  is quasiuniform on  $\Sigma$ .

We start by developing a discrete Helmholtz decomposition to reduce the problem to an operator equation suitable for analysis. Then we prove pointwise convergence of the appropriate operators and finally apply the theory of collectively compact operators to prove convergence.

From the properties of scalar finite element spaces in Section 5.6 and the commuting diagram (5.59), we know that if(7.23)

$$S_h = \{p \in u_h \mid p = 0 \text{ on } \Gamma, p \text{ is constant on } \Sigma\}$$

then  $S_b \subset S$  and  $\nabla S_b \subset X_b$ . In fact,  $\nabla S_b \subset X_b$ . To check this we need only check that the boundary conditions in the definition of  $X_b$  are satisfied. But these hold because any  $p_b \in S_b$  is constant on  $\Gamma$  and  $\Sigma$ , and hence has vanishing surface

gradient there. Thus we can write the following *discrete Helmholtz decomposition* as

$$X_h = X_{0,h} \oplus \nabla S_h,$$

where

$$X_{0,h} = \{u_h \in X_h \mid (\in_r u_h, \nabla \xi_h) = 0 \text{ for all } \xi_h \in S_h\}.$$

Obviously, if  $u_b \in X_{0,b}$  it is not necessarily the case that  $\nabla \cdot (\epsilon_r u_b) = 0$ , so  $X_{0,b} \not\subseteq X_0$ . This causes difficulties for analysis. However, the discrete functions in  $X_{0,b}$  are orthogonal to the gradient of all scalar finite element functions in  $S_b$ . Now we write

$$E_h = E_{0,h} + \nabla p_h \text{ for some } E_{0,h} \in X_{0,h} \text{ and } p_h \in S_h.$$

Substituting into (7.2), selecting  $\varphi_b = \nabla \xi_b$  for some  $\xi_b \in S_b$  and using the definition of  $X_{0,b}$ , we find the discrete analogue of (4.11):(7.24)

$$-\kappa^2(\in_r \nabla p_h, \nabla \xi_b) = (F, \nabla \xi_b) \text{ for all } \xi_b \in S_b.$$

Because  $\epsilon_r$  has a positive real part, we can apply the Lax–Milgram Lemma 2.21 to guarantee that (7.24) has a unique solution. Then Cea's Lemma 2.37 implies a quasi-optimal error estimate and we have proved the next lemma.

**Lemma 7.8** *Let  $p \in S$  satisfy (4.11). There is a unique solution  $p_b \in S_b$  of (7.24) and*

$$\|p - p_b\|_{H^1(\Omega)} \leq C \inf_{\xi_b \in S_b} \|p - \xi_b\|_{H^1(\Omega)}.$$

**Remark 7.9** *If  $p$  is a smooth solution of (4.10) then we can give explicit bounds on the error. For example if  $p \in H^s(\Omega)$ ,  $\frac{3}{2} + \delta \leq s \leq K + 1$ , we have, by Theorem 5.48, that*

$$\|p - p_b\|_{H^1(\Omega)} \leq Ch^{s-1} \|p\|_{H^s(\Omega)}.$$

Now that  $p_b$  is in hand, we may focus on the problem of finding  $E_{0,b} \in X_{0,b}$  such that (7.25)

$$\begin{aligned} (\mu_r^{-1} \nabla \times E_{0,h}, \nabla \times \varphi_h) - \kappa^2(\in_r E_{0,h}, \varphi_h) - i\kappa(\lambda E_{0,h,T}, \varphi_{h,T}) \\ = (F, \varphi_h) + (g, \varphi_{h,T}) + \kappa^2(\in_r \nabla p_h, \varphi_h) \end{aligned}$$

for all  $\varphi_b \in X_{0,b}$ . Paralleling the analysis of (4.13) in Chapter 4, we recall the sesquilinear form  $a_+$  given by (4.14) and define the discrete operator  $K_b : (L^2(\Omega))^3 \rightarrow (L^2(\Omega))^3$  and vector  $F_b \in (L^2(\Omega))^3$  by requiring, for any  $f \in (L^2(\Omega))^3$ , that  $Kf \in X_{0,b} \subset (L^2(\Omega))^3$  satisfies (7.26)

$$a_+(K_b F, \varphi_h) = -2\kappa^2(\in_r F, \varphi_h) \text{ for all } \varphi_h \in X_{0,h},$$

and the function  $F_b \in X_{0,b}$  satisfies (7.27)

$$a_+(\mathcal{F}_h, \varphi_h) = (F, \varphi_h) + (g, \varphi_{h,T}) + \kappa^2(\in_r \nabla p_h, \varphi_h) \text{ for all } \varphi_h \in X_{0,h}.$$

where  $a_+(\cdot, \cdot)$  is given by (4.14). As in the proof of Theorem 4.11, using the Lax–Milgram Lemma 2.21 now applied with  $X_{0,b}$  in place of  $X_0$ , we know that

$K_b$  and  $F_b$  are well defined. We can write the discrete problem (7.25) as the problem of finding  $E_{0,b} \in (L^2(\Omega))^3$  such that(7.28)

$$E_{0,h} + K_h E_{0,h} = \mathcal{F}_h .$$

Of course, a solution  $E_{0,b}$  of this problem will satisfy

$$E_{0,h} = \mathcal{F}_h - K_h E_{0,h} \in X_{0,h},$$

so we are effectively computing the same solution as before and have not changed the problem by posing the problem in  $(L^2(\Omega))^3$ . The way ahead is now clear. From Section 4.5 we know that  $E_0 \in (L^2(\Omega))^3$  satisfies (4.18) and by the above argument  $E_{0,b} \in (L^2(\Omega))^3$  satisfies (7.28). We shall apply the theory of collectively compact operators (see [16, 193]) and Section 2.3.3 to prove that  $\|E_0 - E_{0,b}\|_{(L^2(\Omega))^3} \rightarrow 0$ , as  $b \rightarrow 0$ .

### 7.3.1 Pointwise convergence

The first step in applying the theory of collectively compact operators given in Section 2.3.3 is to verify the pointwise convergence of  $Kf$  to  $Kf$  in  $(L^2(\Omega))^3$ . As a preliminary before doing this, we need to verify that the finite element spaces are dense in the appropriate way, which we do next.

**Lemma 7.10** *The space  $X_b$  is dense in  $X$ , in the sense that for any  $u \in X$*

$$\lim_{h \rightarrow 0} \inf_{\chi_h \in X_h} \|u - \chi_h\|_X = 0.$$

Similarly  $S_b$  is dense in  $S$ .

**Proof** Since the space  $\chi$  defined in Theorem 4.1 is dense in  $X$ , we can approximate any  $u \in X$  to arbitrary accuracy by a smooth function  $\chi \in \chi$ . This function can be approximated to arbitrary accuracy by its interpolant in  $X_b$  if  $b$  is taken to be sufficiently small (see Theorem 5.53).

Similarly, writing  $p \in S$  as  $p = p_0 + p_1$ , where  $p_0 \in H_0^1(\Omega)$  and  $p_1$  is a smooth function, taking the constant value of  $p$  on  $\Sigma$  and vanishing on  $\Gamma$  and then using the density of  $C_0^\infty(\Omega)$  in  $H_0^1(\Omega)$ , we can verify the density of  $S \cap C^\infty(\Omega)$  in  $S$ . Then the error estimates in Theorem 5.48 provide the desired result.  $\square$

**Theorem 7.11** *Given a function  $f \in (L^2(\Omega))^3$ , we have  $\|(K - K_b)f\|_X \rightarrow 0$  as  $b \rightarrow 0$ . Furthermore, if  $f \in X_0$ ,*

$$\|(K - K_b)F\|_X \leq C \inf_{\chi_h \in X_h} \|KF - \chi_h\|_X .$$

**Proof** We use the theory of mixed finite element methods given in Sections 2.2.3 and 2.3.2 to rewrite the variational problems for  $K$  and  $K_b$ . For the operator  $K$

this is done as follows. Given  $f \in (L^2(\Omega))^3$ , we seek  $Kf \in X$  and  $q \in S$  such that(7.29)

$$\begin{aligned} a + (KF, \varphi) + (\epsilon_r \varphi, \nabla q) &= -(\kappa^2 + 1)(\epsilon_r f, \varphi) \quad \text{for all } \varphi \in X, \\ (\epsilon_r KF, \nabla \xi) &= 0 \quad \text{for all } \xi \in S. \end{aligned}$$

To see that this is wellposed, we apply the Babuška–Brezzi theory of mixed methods given in Section 2.2. We note that Lemma 4.10 shows that  $a_+(\cdot, \cdot)$  is coercive on  $X$ . Furthermore,  $\nabla S \subset X$ , thus the inf-sup condition is easily verified, since taking  $\varphi = \nabla q$  we have (using the positive definiteness of the real part of  $\epsilon_r$ )

$$|(\epsilon_r \varphi \cdot \nabla q)| = |((\epsilon_r \nabla q) \cdot \nabla q)| \geq C \|\nabla q\|_{(L^2(\Omega))^3}^2 \geq C \|q\|_{H^1(\Omega)}^2,$$

where we have used the Poincaré inequality (Theorem 3.13) to bound the norm of  $q$  in  $H^1(\Omega)$  in terms of the semi-norm. Thus, (7.29) has a unique solution agreeing with the previous definition of  $Kf$  in (4.15), as we can see if we select  $\varphi \in X_0$  in (7.29).

Similarly, the finite element operator  $K_b$  can be defined as the solution of the mixed finite element problem of finding  $K_b f \in X_b$  and  $q_b \in S_b$  such that

$$\begin{aligned} a + (K_b f, \varphi_b) + (\epsilon_r \varphi_b, \nabla q_b) &= -(\kappa^2 + 1)(\epsilon_r f, \varphi_b) \quad \text{for all } \varphi_b \in X_b, \\ (\epsilon_r K_b f, \nabla \xi_b) &= 0 \quad \text{for all } \xi_b \in S_b. \end{aligned}$$

Again  $a_+(\cdot, \cdot)$  is coercive and, since  $\nabla S_b \subset X_b$ , the same argument as above verifies the Babuška–Brezzi condition for this discrete mixed problem. Choosing  $\varphi_b \in X_{0,b}$  shows that the problem reduces to the previous definition of  $K_b$  given in (7.26).

Hence, using Theorem 2.45, we have to estimate(7.30)

$$\begin{aligned} &\|(K - K_b)F\|_X + \|\nabla(q - q_b)\|_{(L^2(\Omega))^3} \\ &\leq C \left\{ \inf_{\mathcal{X}_h \in X_h} \|KF - \mathcal{X}_h\|_{\mathcal{X}} + \inf_{\xi_h \in S_h} \|\nabla(q - \xi_h)\|_{(L^2(\Omega))^3} \right\}. \end{aligned}$$

But, by Lemma 7.10,  $S_b$  is dense in  $S$ , and  $X_b$  is dense in  $X$  as  $b \rightarrow 0$ . Hence the right-hand side converges to zero as  $b$  decreases, and pointwise convergence is proved.

Now if  $f \in X_0$ , choosing  $\varphi = \nabla q$  in (7.29) we obtain (using also that  $Kf \in X_0$ ) that  $(\epsilon_r \varphi, \nabla q) = 0$  and hence  $q = 0$ . The estimate then follows from (7.30).  $\square$

Since we need to perform a similar analysis to estimate  $F - F_b$ , we do this next.

**Lemma 7.12** *Let  $F$  be defined by(4.16)and  $F_b$  by (7.27). Then*

$$\|\mathfrak{F} - \mathfrak{F}_b\|_X \leq C \left\{ \inf_{\mathcal{X}_h \in X_h} \|\mathfrak{F} - \mathcal{X}_h\|_{\mathcal{X}} + \inf_{\xi_h \in S_h} \|\nabla(p - \xi_h)\|_{(L^2(\Omega))^3} \right\},$$

where  $p \in S$  satisfies (4.11).

**Proof** We write the definition of  $F$  as the mixed problem of finding  $F \in X$ , and  $r \in S$  such that

$$\begin{aligned} a + (\mathcal{F}, \varphi) + (\epsilon_r \varphi, \nabla r) &= (F, \varphi) + \langle g, \varphi_T \rangle + \kappa^2 (\epsilon_r \nabla p, \varphi) \quad \text{for all } \varphi \in X, \\ (\epsilon_r \mathcal{F}, \nabla \xi) &= 0 \quad \text{for all } \xi \in S. \end{aligned}$$

Selecting  $\varphi = \nabla r$ , and using the definition of  $p$ , we see that the right-hand side vanishes, so that  $(\epsilon_{r\nabla}, \nabla r) = 0$  and so  $r = 0$ .

Next we define a new function  $F_b \in X_b$  and  $r_b \in S_b$  to satisfy

$$\begin{aligned} a + (\mathcal{F}_h, \varphi_h) + (\epsilon_r \varphi_h, \nabla r_h) &= (F, \varphi_h) + \langle g, \varphi_{h,T} \rangle + \kappa^2 (\epsilon_r \nabla p, \varphi_h) \\ &\quad \text{for all } \varphi_h \in X_h, \\ (\epsilon_r \mathcal{F}_h, \nabla \xi_h) &= 0 \quad \text{for all } \xi_h \in S_h. \end{aligned}$$

Exactly the same argument, choosing  $\varphi_b = \nabla r_b$ , shows that  $r_b = 0$ . Performing the mixed method analysis as in the proof of the previous theorem shows that  $\|F - F_b\|_X \leq C \inf_{\chi_b} \|F - \chi_b\|_X$ . But  $F_b \in X_{0,b}$  satisfies

$$a + (\widetilde{\mathcal{F}}_h, \varphi_h) = (\mathcal{F}, \varphi_h) + \langle g, \varphi_{h,T} \rangle + \kappa^2 (\epsilon_r \nabla p, \varphi_h) \quad \text{for all } \varphi_h \in X_{0,h},$$

so subtracting (7.27) from this equation gives  $a_+(F_b - F, \varphi_b) = \kappa^2 (\epsilon_{r\nabla} p - p_b, \varphi_b)$  and selecting  $\varphi_b = F_b - F$  shows that

$$\begin{aligned} C \|\widetilde{\mathcal{F}}_h - \mathcal{F}_h\|_X^2 &\leq |a(\widetilde{\mathcal{F}}_h - \mathcal{F}_h, \widetilde{\mathcal{F}}_h - \mathcal{F}_h)| = |\kappa^2 (\epsilon_r \nabla(p - p_b), \widetilde{\mathcal{F}}_h - \mathcal{F}_h)| \\ &\leq C \|\nabla(p - p_b)\|_{(L^2(\Omega))^3} \|\widetilde{\mathcal{F}}_h - \mathcal{F}_h\|_X. \end{aligned}$$

Then the triangle inequality implies that  $\|F - F_b\|_X \leq \|F - F_b\|_X + \|F_b - F\|_X$ , and the estimate follows from the estimate of  $\|\nabla(p - p_b)\|_{(L^2(\Omega))^3}$  in Lemma 7.8.  $\square$

### 7.3.2 Collective compactness

Let  $\Lambda$  denote a countable set of mesh sizes whose only accumulation point is zero. So  $\omega = \{h_n\}_{n=1}^\infty$  and  $b_n \rightarrow 0$  as  $n \rightarrow 0$ . This set is a sequence of decreasing mesh sizes.

We need to show that  $\{K_b\}_{b \in \Lambda}$  is a collectively compact set of operators. This will follow once we have proved that the finite element space has a discrete compactness property which was first described by Kikuchi for the lowest-order Nédélec element [185]. Using more recent regularity theory, Leszek Demkowicz and myself were able to extend the discrete compactness property to all orders of Nédélec elements on a reasonably large class of domains, but only if  $\epsilon_r = 1$  [119]. To complete the picture Caorsi *et al.* [71] have shown that, once the discrete compactness property is proved for  $\epsilon_r = 1$ , it holds for a general class of coefficients including the ones in this book. Thus we shall now state and prove the discrete compactness property and show that this implies that  $\{K_b\}_{b \in \Lambda}$  is collectively compact.

**Definition 7.13** We say  $X_{0,b}$ ,  $b \in \Lambda$ , has the *discrete compactness property* if for every sequence  $\{u_b\}_{b \in \Lambda}$  such that

- $u_b \in X_{0,b}$  for each  $b \in \Lambda$ ;
- there is a constant  $C$  independent of  $u_b$  such that  $\|u_b\|_X \leq C$  independent of  $b \in \Lambda$ ,

there exists a subsequence, still denoted  $\{u_b\}$ , and a function  $u \in X_0$  such that

$$u_b \rightarrow u \text{ strongly in } (L^2(\Omega))^3 \text{ as } b \rightarrow 0 \text{ in } \Lambda.$$

Supposing for the moment that  $X_{0,b}$  has the discrete compactness property, we can easily prove the collective compactness of  $\{K_b\}_{b \in \Lambda}$ .

**Theorem 7.14** If  $X_{0,b}$ ,  $b \in \Lambda$ , has the discrete compactness property then  $\{K_b\}_{b \in \Lambda}$  is collectively compact as a set of maps from  $(L^2(\Omega))^3$  to  $(L^2(\Omega))^3$ .

**Proof** Let  $U$  be a bounded set in  $(L^2(\Omega))^3$ . We need to show that  $K(U)$  is relatively compact in  $(L^2(\Omega))^3$ . Let  $\{\omega_n\}_{n=1}^\infty \subset K(u)$  be a sequence. Then for each  $n$  there is an  $b_n \in \Lambda$  and  $u_n \in U$  such that  $w_n = K_{b_n}(u_n)$ . Hence  $\omega_n \in X_{0,b_n}$  and  $\|w_n\|_X \leq C \|u_n\|_{(L^2(\Omega))^3} \leq C$ . Without loss of generality, we can assume  $b_n \rightarrow 0$  as  $n \rightarrow \infty$  (otherwise, we are in a finite-dimensional space and the convergence of a subsequence in  $\{\omega_n\}_{n=1}^\infty$  is guaranteed). But  $\{\omega_n\}_{n=1}^\infty$  is exactly as in the definition of discrete compactness and so the existence of a convergent subsequence is assured.  $\square$

Now it remains to prove the discrete compactness property. Unfortunately, we can only do this using regularity theory for Maxwell's equations. It would be highly desirable to obtain a proof without this constraint. The proof progresses in two steps. First we prove discrete compactness when  $\varepsilon_r = \mu_r = 1$  and then, using an argument of Caorsi *et al.* [71] we extend the result to general  $\varepsilon_r$ .

We start with a regularity result from [167].

**Lemma 7.15** Suppose  $\Omega$  is a bounded, simply connected Lipschitz polyhedron with boundary  $\partial\Omega$  consisting of two connected components  $\Sigma$  and  $\Gamma$ . Let  $\tau_b$  be a regular mesh that is also quasi-uniform on  $\Sigma$ . Let  $u_b \in X_{0,b}$  and suppose  $u \in X_0$  satisfies

$$\begin{aligned} \nabla \times u &= \nabla \times u_b && \text{in } \Omega, \\ v \times u &= v \times u_b && \text{on } \partial\Omega. \end{aligned}$$

Then there is a  $\delta > 0$  with  $\delta \leq 1/2$  such that  $u \in (H^{1/2+s}(\Omega))^3$ , for  $0 \leq s < \delta$  and

$$\|u\|_{(H^{1/2+s}(\Omega))^3} \leq C \left( \|\nabla \times u\|_{(L^2(\Omega))^3} + \|v \times u\|_{(H^s(\Sigma))^3} \right).$$

**Remark 7.16** In Theorem 3.47 this result is proved for  $s = 0$ . The result is proved for  $v \times u_b = 0$  in [12]. The result here is possible since  $v \times u_b$  is smoother than just square integrable. The proof combines features of [102] and [12], and is from [167].

**Proof of Lemma 7.15** In this proof we shall use the spaces  $H(\partial\Omega)$ ,  $l > 1$  defined in (3.12). Let  $O$  denote a smooth bounded and simply connected domain with connected boundary containing  $\Omega$  in its interior. First we construct a vector potential  $w \in (H^l(O))^3$  such that  $\nabla \times w = \nabla \times u_b$  in  $\Omega$  and  $\nabla \cdot w = 0$  in  $\Omega$ . This is done as follows.

Let  $\zeta$  be defined on  $O$  by

$$\zeta = \begin{cases} 0 & \text{in } D, \\ \nabla \times u & \text{in } \Omega, \\ \nabla \xi & \text{in } O \setminus (\overline{\Omega \cup D}). \end{cases}$$

Here  $\xi \in H^1(O \setminus (\overline{\Omega \cup D})) / \mathbb{R}$  solves the boundary value problem

$$\begin{aligned} \Delta \xi &= 0 \quad \text{in } O \setminus (\overline{\Omega \cup D}), \\ \frac{\partial \xi}{\partial v} &= v \cdot \nabla \times u_h \quad \text{on } \Sigma, \\ \frac{\partial \xi}{\partial v} &= 0 \quad \text{on } \partial O. \end{aligned}$$

Note that  $v \cdot \nabla \times u_h \in H^{1/2}(\Sigma)$  since  $\nabla \times u_h \in H(\operatorname{div}; \Omega)$ , and the fact that  $\nabla \cdot (\nabla \times u_h) = 0$  implies the necessary compatibility condition (of course,  $v \cdot \nabla \times u_h = 0$  on  $\Gamma$  since  $v \times u_h = 0$  there). Thus  $\zeta$  has continuous normal component across  $\Sigma$  and  $\Gamma$  and  $\nabla \cdot \zeta = 0$  in  $O$ . Hence, by Theorem 3.38, there is a function  $w \in (H^l(O))^3$  with the desired properties.

Since  $\Omega$  is simply connected and  $\nabla \times (u - w) = 0$  in  $\Omega$ , there is, by Theorem 3.37, a scalar potential  $p \in H^l(\Omega)$  such that  $u - w = \nabla p$ . But since  $\nabla \cdot (u - w) = \nabla \cdot u - \nabla \cdot w = 0$  in  $\Omega$  we have  $\nabla p = 0$  in  $\Omega$ . Now we follow [12] to show  $p \in H^{3/2+s}(\Omega)$  where  $s$  is the index in the statement of the theorem. Inside  $D$  we have  $\nabla \times w = 0$ , so since  $w \in (H^l(D))^3$ , there is a scalar potential  $\eta \in H^s(D)$  with  $w = \nabla \eta$ . On  $\Gamma$

$$(v \times (u - w)) \times v = (v \times \nabla p) \times v = \nabla_{\Gamma} p,$$

where  $\nabla_{\Gamma}$  denotes the surface gradient. But  $(v \times w) \times v = \nabla_{\Gamma} \eta$ . Thus  $\nabla_{\Gamma} \eta = \nabla_{\Gamma} p$  and so, possibly adjusting  $\eta$  by a constant,  $\eta = p$  and we have  $p \in H^{3/2}(\Gamma)$ .

In the case of  $\Sigma$ , clearly  $p|_{\Sigma} \in H^{1/2}(\Sigma)$  and  $w|_{\Sigma} \in (H^{1/2}(\Sigma))^3$ . Furthermore,  $(v \times (u - w)) \times v = -\nabla_{\Sigma} p$  (again  $\nabla_{\Sigma}$  is the surface gradient of  $p$ ). Hence on  $\Sigma$  we have  $u_T - w_T = \nabla_{\Sigma} p$ . In addition,  $u_T = u_{bT}$ . Now for each face  $F$  of  $\Sigma$ , the normal vector  $v$  is constant and  $u_b$  is piecewise polynomial, so via Lemma 5.57,

$$u_{h,T}|_F \in (H^s(F))^3, 0 \leq s < \frac{1}{2}.$$

Thus  $\nabla_{\Sigma} p \in (H(F))^3$  and so  $p \in H^{1+s}(F)$ ,  $0 \leq s \leq \frac{1}{2}$ .

Since  $p \in H^{1+s}(F)$ , it is continuous on  $F$  and so must be continuous on  $\Sigma$ , otherwise we would not have  $p|_{\Sigma} \in H^{1/2}(\Sigma)$ . But by the characterization of traces for polyhedral domains in Theorem 3.10, we have  $p \in H^{1+s}(\Sigma)$ . Thus  $p$

is the trace of a function  $\tilde{p} \in H^{3/2+s}(\Omega)$ . Now considering the function  $p - \tilde{p}$  we see that

$$\begin{aligned}\Delta(p - \tilde{p}) &= -\Delta\tilde{p} \quad \text{on } \Omega, \\ p - \tilde{p} &= 0 \quad \text{on } \partial\Omega.\end{aligned}$$

By Theorem 3.18, there is a  $\delta > 0$  such that  $p - \tilde{p} \in H^{3/2+s}(\Omega)$  for  $0 \leq s < \delta$  and we are done.  $\square$

Now we can prove the discrete compactness property for  $X_{0,b}$  when  $\varepsilon_r = \mu_r = 1$ . Note that in this theorem the boundary inverse property applies to  $\sum$  only.

**Theorem 7.17** Suppose  $\varepsilon_r = \mu_r = 1$  and  $\{\tau_b\}_{b \in \mathcal{A}}$  is regular and possesses the boundary inverse property on  $\sum$  (see Section 5.8). Then  $X_{0,b}$  possesses the discrete compactness property.

**Proof of Theorem 7.17** Let  $\omega_n \in X_{0,h_n}$ ,  $n = 1, 2, \dots$  and suppose  $h_n \rightarrow 0$  and  $\|\omega_n\|_X \leq C < \infty$ , for all  $n$ . Let  $p^n \in S$  satisfy  $(\nabla p^n, \nabla \xi) = (w_n, \nabla \xi)$  for all  $\xi \in S$ . Then let  $w^n = w_n - \nabla p^n$ . Clearly,  $w^n$  satisfies

$$\begin{aligned}\nabla \times \omega^n &= \nabla \times \omega_n \quad \text{and } \nabla \cdot \omega^n = 0 \quad \text{in } \Omega, \\ \nu \times \omega^n &= \nu \times \omega_n \quad \text{on } \partial\Omega.\end{aligned}$$

Hence  $w^n \in X_0$  and  $\|w^n\|_X \leq C$ . So, by the continuous compactness result in Theorem 4.7, there is a subsequence, still denoted by  $\{\omega^n\}_{n=1}^\infty$  and a function  $w \in X_0$  such that  $w^n \rightarrow w$  as  $n \rightarrow \infty$  strongly in  $(L^2(\Omega))^3$ . Since, by our previous lemma,  $w^n \in (H^{1/2+s}(\Omega))^3$ ,  $s > 0$ , and  $\nabla \times w^n = \nabla \times w_n \in W_b$ , we know by Theorem 5.41 that the interpolant  $r_{h_n} \omega_n$  is well defined. Using the fact that  $r_{h_n} \omega_n = \omega_n$ , the interpolant of  $\nabla p^n$  is well defined and so, using the commuting diagram (5.59), we have  $r_{h_n} \omega^n = \omega_n - \nabla \pi_{h_n} p^n$ . Hence, using the fact that  $w \in X_0$  and  $\omega_n \in X_{0,h_n}$ ,

$$\begin{aligned}&\int_\Omega (\omega - \omega_n) \cdot (\overline{\omega - \omega_n}) dV \\ &= \int_\Omega (\omega - \omega_n) \cdot (\overline{\omega - r_{h_n} \omega^n}) dV + \int_\Omega (\omega - \omega_n) \cdot (r_{h_n} \omega^n - \omega_n) dV \\ &= \int_\Omega (\omega - \omega_n) \cdot (\overline{\omega - r_{h_n} \omega^n}) dV + \int_\Omega (\omega - \omega_n) \cdot (-\overline{\nabla \pi_{h_n} p^n}) dV \\ &\leq \|\omega - \omega_n\|_{(L^2(\Omega))^3} \leq \|\omega - r_{h_n} \omega^n\|_{(L^2(\Omega))^3}.\end{aligned}$$

Thus  $\|\omega - \omega_n\|_{(L^2(\Omega))^3} \leq \|\omega - r_{h_n} \omega^n\|_{(L^2(\Omega))^3}$ . But

$$\|\omega - r_{h_n} \omega^n\|_{(L^2(\Omega))^3} \leq \|\omega - \omega^n\|_{(L^2(\Omega))^3} + \|\omega^n - r_{h_n} \omega^n\|_{(L^2(\Omega))^3}.$$

Now we expand the second term and use the error estimate (5.43) to obtain

$$\begin{aligned}\|\omega^n - r_{h_n} \omega^n\|_{(L^2(\omega))^3}^2 &= \sum_{K \in \mathcal{T}_h} \|\omega^n - r_{h_n} \omega^n\|_{(L^2(K))^3}^2 \\ &\leq \sum_{K \in \mathcal{T}_h} C \left( h_K^{1/2+s} \|\omega^n\|_{(H^{1/2+s}(K))^3} + h_K \|\nabla \times \omega^n\|_{(L^2(K))^3} \right)^2.\end{aligned}$$

Using the triangle inequality

$$\begin{aligned} & \left\| \omega^n - rh_n \omega^n \right\|_{(L^2(\Omega))^3} \\ & \leq C \left( h_n^{1/2+s} \|\omega^n\|_{(H^{1/2+s}(\Omega))^3} + h_n \|\nabla \times \omega^n\|_{(L^2(\Omega))^3} \right). \end{aligned}$$

But, from the previous lemma,

$$\|\omega^n\|_{(H^{1/2+s}(\Omega))^3} \leq C \left( \|\nabla \times \omega_n\|_{(L^2(\Omega))^3} + \|v \times \omega_n\|_{(H^s(\Sigma))^3} \right).$$

Using the boundary quasi-uniformity assumption, the inverse estimate for fractional order spaces from Lemma 5.57 and the remark following that lemma, we have

$$\|v \times \omega_n\|_{(H^s(\Sigma))^3} \leq Ch_n^{-s} \|\nabla \times \omega_n\|_{(L^2(\Sigma))^3}.$$

Thus

$$\|\omega - \omega_n\|_{(L^2(\Omega))^3} \leq \|\omega - \omega^n\|_{(L^2(\Omega))^3} + Ch_n^{1/2} \|\omega_n\|_X.$$

The first term on the right-hand side converges to zero by construction and the second because  $h_n \rightarrow 0$  and  $\|\omega_n\|_X \leq C$  as  $n \rightarrow \infty$ . Hence we have proved that  $\omega_n \rightarrow \omega$  in  $(L^2(\Omega))^3$  as  $n \rightarrow \infty$  and we are done.  $\square$

We now prove the discrete compactness property for general  $\epsilon_r$ .

**Theorem 7.18** *Let  $\Omega$  be a bounded simply connected Lipschitz domain with boundary  $\partial\Omega$  consisting of two connected components  $\Sigma$  and  $\Gamma$ . Suppose  $\epsilon_r$  satisfies the assumptions in Section 4.2, and that the mesh is regular and quasi-uniform on  $\Sigma$ . Then  $\{X_{0,h}\}_{h \in A}$  has the discrete compactness property.*

**Remark 7.19** *The proof is due to Caorsi et al. [71], who also investigate the connection between various properties of discrete spaces to clarify the relationship between the various convergence theories.*

**Proof of Theorem 7.18** We again use the notation of Theorem 4.7, where the dependence on  $\epsilon_r$  is explicitly recognized by a superscript. In particular, let

$$X_{0,h}^{(\epsilon_r)} = \{u_h \in X_h \mid (\epsilon_r u_h, \nabla \xi_h) = 0 \text{ for all } \xi_h \in S_h\}$$

for  $\epsilon = 1$  or  $\epsilon = \epsilon_r$ . We have the Helmholtz decomposition with respect to the standard  $(L^2(\Omega))^3$  inner product, (7.31)

$$X_h = X_{0,h}^{(1)} \oplus \nabla S_h,$$

and using the bilinear form  $(u, v)_{L^2_{\epsilon_r}(\Omega)} = (\epsilon_r u, v)$  from Section 4.7, we have the Helmholtz decomposition (7.32)

$$X_h = X_{0,h}^{(\epsilon_r)} \oplus \nabla S_h,$$

where  $u_h \in X_{0,h}^{(\epsilon_r)}$  if and only if  $u_b \in X_b$  and  $(\epsilon_r u_b, \nabla \xi_b) = 0$  for all  $\xi_b \in S_b$ .

Now suppose we have a sequence  $\{\omega_n\}_{n=1}^{\infty}$  such that

- $\omega_n \in X_{0,h_n}^{(\epsilon_r)}$  for each  $n$ ,
- $\|w_n\|_x \leq C$  for all  $n$ ,

and  $b_n \rightarrow 0$  as  $n \rightarrow \infty$ . Using the decomposition (7.31) we have

$$\omega_n = \omega_n^{(1,0)} + \nabla p_n^{(1)}, \quad \text{where } \omega_n^{(1,0)} \in X_{0,h_n}^{(1)} \text{ and } p_n^{(1)} \in S_{h_n},$$

and  $\|\omega_n^{(1,0)}\|_X \leq C\|\omega_n\|_X$ , so there exists a convergent subsequence (using Theorem 7.17) such that  $\{\omega_n^{(1,0)}\}_{n=1}^{\infty}$  converges in  $(L^2(\Omega))^3$  to  $\omega^{(1,0)} \in X_0^{(1)}$ .

Now let

$$\omega^{(1,0)} = \omega^{(\epsilon_r, 0)} + \nabla p^{(\epsilon_r)}, \quad \text{where } \omega^{(\epsilon_r, 0)} \in X_0^{(\epsilon_r)} \text{ and } p^{(\epsilon_r)} \in S.$$

We shall now show that  $w_n \rightarrow w^{(\epsilon_r, 0)}$  in  $(L^2(\Omega))^3$  as  $n \rightarrow \infty$ . Using the orthogonality of  $w^{(\epsilon_r, 0)}$  and  $w_n$  to gradients of functions in  $S_{h_n}$  we have

$$\begin{aligned} & \left( \in \epsilon_r (\omega^{(\epsilon_r, 0)} - \omega_n), (\omega^{(\epsilon_r, 0)} - \omega_n) \right) \\ &= \left( \in \epsilon_r (\omega^{(\epsilon_r, 0)} - \omega_n), (\omega^{(\epsilon_r, 0)} - \omega_n^{(1,0)} - \nabla p_n^{(1)}) \right) \\ &= \left( \in \epsilon_r (\omega^{(\epsilon_r, 0)} - \omega_n), (\omega^{(\epsilon_r, 0)} - \omega_n^{(1,0)} + \nabla \xi_n) \right) \end{aligned}$$

for all  $\xi_n \in S_{h_n}$ . Hence

$$\begin{aligned} & \|\omega^{(\epsilon_r, 0)} - \omega_n\|_{(L^2(\Omega))^3} \\ &\leq C \|\omega^{(\epsilon_r, 0)} - \omega_n^{(1,0)} + \nabla \xi_n\|_{(L^2(\Omega))^3} \\ &\leq C \left( \|\omega^{(\epsilon_r, 0)} - \omega_n^{(1,0)} + \nabla p^{(\epsilon_r)}\|_{(L^2(\Omega))^3} + \|\nabla(\xi_n - p^{(\epsilon_r)})\|_{(L^2(\Omega))^3} \right) \\ &= C \left( \|\omega^{(1,0)} - \omega_n^{(1,0)}\|_{(L^2(\Omega))^3} + \|\nabla(\xi_n - p^{(\epsilon_r)})\|_{(L^2(\Omega))^3} \right). \end{aligned}$$

The first term on the right hand side converges to zero by the convergence of  $\{\omega_n^{(1,0)}\}_{n=1}^{\infty}$  to  $\omega^{(1,0)}$  and the second by the density of  $S_{h_n}$  in  $S$  as  $n \rightarrow \infty$ .  $\square$

Now that we have verified the discrete compactness property, we can prove the following result that guarantees that the low-frequency limit of Maxwell's equations is well approximated by edge elements. This is the discrete version of the Friedrichs inequality in Corollary 4.8.

**Lemma 7.20** Under the assumptions of Theorem 7.18, there exists a positive constant  $C$  independent of  $h \in \Lambda$  such that if  $u_b \in X_{0,b}$ , for  $b \in \Lambda$  small enough then

$$\|u_h\|_{(L^2(\Omega))^3} \leq C \left( \|\nabla \times u_h\|_{(L^2(\Omega))^3} + \|V \times u_h\|_{(L^2(\Sigma))^3} \right).$$

**Remark 7.21** It would be better to have a proof of this directly from the definition of the space. This is possible in some cases [233].

**Proof of Lemma 7.20** Note that since  $\nabla \times u_b = 0$  in  $\Omega$  and  $V \times u_b = 0$  on  $\partial\Omega$  implies  $u_b = \nabla p_b$  for some  $p_b \in S_b$ . Hence  $u_b = 0$ , we know that this estimate holds for any  $b > 0$ , but with  $C = C(b)$ .

The proof that  $C$  is independent of  $b$  is by contradiction. Suppose the result does not hold. Then there is a sequence of mesh sizes  $h_n \in \Lambda$ ,  $n = 1, 2, \dots$  and functions  $u_{h_n} \in X_{0,h_n}$  such that  $\|u_{h_n}\|_{(L^2(\Omega))^3} = 1$  and  $\|\nabla \times u_{h_n}\|_{(L^2(\Omega))^3} + \|\nabla \times u_{h_n}\|_{(L^2(\Sigma))^3} \leq 1/n$ . But by discrete compactness, a subsequence, still denoted by  $(u_{h_n})_{n=0}^\infty$ , converges in  $(L^2(\Omega))^3$  to a function  $u \in X_0$ . Clearly,  $\|u\|_{(L^2(\Omega))^3} = 1$ , but

$$\|\nabla \times u\|_{(L^2(\Omega))^3} + \|\nabla \times u\|_{(L^2(\Sigma))^3} = 0.$$

Since  $\nabla \times u = 0$  in  $\Omega$   $\nabla \times u = 0$  on  $\partial\Omega$  and  $u \in X_0$ , we know that  $u = 0$ . But this is a contradiction and the proof is complete.  $\square$

Before continuing with our analysis of the finite element problem (7.2), we state and prove a result that will be useful later when we analyze Schwarz iterative methods. Note that in this result we take  $\varepsilon_r$  to be a symmetric matrix.

**Corollary 7.22** Suppose the discrete Friedrichs inequality in Lemma 7.20 holds when  $\varepsilon_r = 1$  with constant  $C_1$ . Now suppose that  $\varepsilon_r$  is a real, symmetric, positive-definite and continuous matrix-valued function of position on  $\Omega$ . Let

$$X_{0,h}^{(\varepsilon_r)} = \{v_h \in X_h \mid (\varepsilon_r v_h, \nabla \xi_h) = 0 \text{ for all } \xi_h \in S_h\}$$

then, for all  $v_h \in X_{0,h}^{(\varepsilon_r)}$ , we have (7.33)

$$(\varepsilon_r v_h, v_h) \leq C_1^2 \max_{x \in \bar{\Omega}} \rho(x) \left( \|\nabla \times v_h\|_{(L^2(\Omega))^3} + \|\nabla \times v_h\|_{(L^2(\Sigma))^3} \right)^2,$$

where  $\rho(\varepsilon_r)$  is the spectral radius of  $\varepsilon_r$ .

**Remark 7.23** This also holds for complex valued and discontinuous  $\varepsilon_r$ , but since we shall not need this case we do not prove it here.

**Proof of Lemma 7.22** Let  $v_h \in X_{0,h}^{(\varepsilon_r)}$ . Then using the standard discrete Helmholtz decomposition with  $\varepsilon_r = 1$

$$v_h = \bar{v}_h + \nabla p_h \quad \text{for some unique } \bar{v}_h \in X_{0,h} \text{ and } p_h \in S_h.$$

Using the fact that  $\varepsilon_r v_h$  is discrete divergence-free and using the Cauchy-Schwarz inequality, we have (7.34)

$$\begin{aligned} (\varepsilon_r v_h, v_h) &= (\varepsilon_r v_h, v_h - \nabla p_h) = (\varepsilon_r v_h, \bar{v}_h) \\ &\leq \|\varepsilon_r^{1/2} v_h\|_{(L^2(\Omega))^3} \|\varepsilon_r^{1/2} \bar{v}_h\|_{(L^2(\Omega))^3}. \end{aligned}$$

But

$$\|\varepsilon_r^{1/2} \bar{v}_h\|_{(L^2(\Omega))^3} = (\varepsilon_r \bar{v}_h, \bar{v}_h) \leq \int_{\Omega} \rho(x) |\bar{v}_h|^2 dV \leq \max_{x \in \bar{\Omega}} \rho(x) \|\bar{v}_h\|_{(L^2(\Omega))^3}^2.$$

Thus, using this estimate in (7.34) and then using the estimate from Lemma 7.20 on  $\bar{v}_h$ , we obtain

$$\begin{aligned} (\varepsilon_r v_h, v_h) &\leq \|\varepsilon_r^{1/2} v_h\|_{(L^2(\Omega))^3} \\ &\leq C_1 \max_{x \in \bar{\Omega}} \rho(x)^{1/2} \left( \|\nabla \times \bar{v}_h\|_{(L^2(\Omega))^3} + \|\nabla \times \bar{v}_h\|_{(L^2(\Sigma))^3} \right). \end{aligned}$$

Since  $\nabla \times \tilde{v}_b = \nabla \times v_b$  in  $\Omega$  and  $v \times \tilde{v}_b = v \times \tilde{v}_b$  on  $\sum$ , the estimate is proved.  $\square$

Now we continue with the analysis of (7.2). Having verified the discrete compactness, and hence the collective compactness of  $\{K_b\}_{b \in \Lambda}$  we have, by Theorem (2.51), that the discrete finite element equations have a unique solution and convergence occurs in  $(L^2(\Omega))^3$ .

**Theorem 7.24** *Let  $\tau_b$  be a regular mesh which is, in addition, quasi-uniform on  $\sum$ . Then under the standard assumptions on the domain and data from Section 4.2 and for  $b \in \Lambda$  sufficiently small,  $(I + K_b)^{-1}$  exists and is uniformly bounded as a map from  $(L^2(\Omega))^3$  to  $(L^2(\Omega))^3$ , thus (7.28) has a unique solution  $E_{0,b} \in X_{0,b}$ . Furthermore, the following error estimate holds:*

$$\|E_{0,h} - E_0\|_{(L^2(\Omega))^3} \leq C \left( \|\mathcal{F} - \mathcal{F}_h\|_{(L^2(\Omega))^3} + \|(K - K_h)E_0\|_{(L^2(\Omega))^3} \right).$$

Now, given the special properties of the finite element problem, we can actually prove the desired theorem on convergence in  $X$ .

**Theorem 7.25** *Under the assumptions for Theorem 7.24, and provided  $b \in \Lambda$  is small enough, the finite element discretization of Maxwell's equations given by (7.2) has a unique solution  $E_b \in X_b$ . Furthermore,*

$$\begin{aligned} \|E_b - E\|_X &\leq C \left( \inf_{\varphi_h \in X_h} \|\mathcal{F} - \varphi_h\|_X + \inf_{\eta_h \in X_h} \|KE_0 - \eta_h\|_X \right. \\ &\quad \left. + \inf_{\xi_h \in S_h} \|p - \xi_h\|_{H^1(\Omega)} \right), \end{aligned}$$

where  $E_0 \in X$  satisfies (4.13) and  $p$  satisfies (4.11).

**Proof** Recall that we wrote  $E_b = E_{0,b} + \nabla p_b$ . We know by the previous theorem that  $E_{0,b} \in X_{0,b}$  exists. In addition, Lemma 7.8 shows that  $p_b$  converges quasi-optimally. But, using the equations for  $E_{0,b}$  and  $E_0$  we have

$$\begin{aligned} \|E_b - E\|_X &\leq \|E_{0,h} - E_0\|_X + \|p - p_h\|_{H^1(\Omega)} \\ &\leq \|K_h E_{0,h} - KE_0\|_X + \|\mathcal{F}_h - \mathcal{F}\|_X + \|p - p_h\|_{H^1(\Omega)} \\ &\leq \|K_h(E_{0,h} - E_0)\|_X + \|K_h - K\|_X + \|\mathcal{F}_h - \mathcal{F}\|_X \\ &\quad + \|p - p_h\|_{H^1(\Omega)}. \end{aligned}$$

The uniform continuity of  $K$  implies

$$\|K_h(E_{0,h} - E_0)\|_X \leq C \|E_{0,h} - E_0\|_{(L^2(\Omega))^3}$$

and this can now be estimated by the previous theorem. The remaining terms on the right-hand side can be estimated using the pointwise estimates of Theorem

7.11 and Lemmas 7.12 and 7.8. Note that  $E_0 \in X_0$ , so we can use the estimate of Theorems 7.11 and 7.12.  $\square$

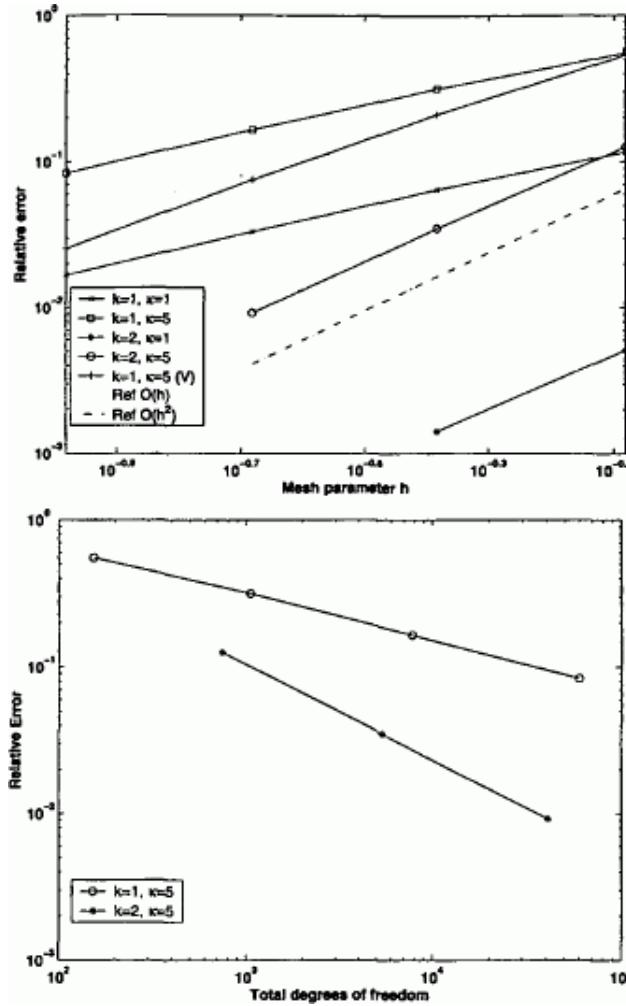
### 7.3.3 Numerical results for the cavity problem

Now we present some very simple numerical results that suggest that the error analysis we have given reflects the convergence rate seen in practice. We solve (7.2) with  $F = 0$  and  $g = \nabla \times E^i - ikE_T^i$ , where  $E^i = p\exp(ikd \cdot x)$ , and  $d = D/|D|$ , where  $D = (11, 1, 5)^\top$  and  $p = (0, -5, 1)^\top/10$ . The domain  $\Omega = [0, 1]^3$  and  $\Gamma = \emptyset$ . The boundary  $\Sigma$  is the surface of the unit cube. Thus the exact solution is  $E^i$  and is an analytic function of position. Hence we should see the optimal approximation rate in the error (for  $k = 1$  edge elements this is  $O(h)$ , and for  $k = 2$  we should see  $O(h^2)$ ). Starting with a coarse mesh of 96 tetrahedra, we subdivide the mesh by bisecting each edge in the mesh to provide successively finer meshes. On each mesh, we solve (7.2) with  $k = 1$  or  $k = 2$ . For small numbers of degrees of freedom we use GMRES and incomplete LU (ILU) preconditioning [146]. Since this is a memory intensive algorithm, for the finer grids we use a symmetric successive over relaxation preconditioned biconjugate gradient scheme. In general, the low wavenumber computation when  $\kappa = 1$  requires more fill-in in the ILU preconditioner than when  $\kappa = 5$ . It is often observed that for small  $\kappa$  the discrete problem becomes less well conditioned (most likely due to a diminishing weight on the discrete divergence condition). We shall comment more on this conditioning problem in the next section.

Having computed the solution, the error is computed by fifth order numerical quadrature on each element. The result for wavenumber  $\kappa = 1$  and  $\kappa = 5$ , and for degrees  $k = 1$  and  $k = 2$  is shown in Figure 7.1. In the left panel of Fig. 7.1, we see that, when  $k = 1$ , the relative  $(L^2(\Omega))^3$  norm error is decreasing consistent with a convergence rate of approximately  $O(h)$ . Similarly for  $k = 2$  we observe  $O(h^2)$ , regardless of  $\kappa$ . However, when  $\kappa = 5$  the error is now much larger than when  $\kappa = 1$ . We shall examine in detail the  $\kappa$  dependence of the error in Section 13.3. One further graph in Fig. 7.1 (left panel) shows the result of computing the error at the vertices in the triangulation when  $\kappa = 1$  and  $k = 1$ . We simply average the value of the finite element solution on each element meeting at each vertex. This graph suggests quadratic convergence ( $O(h^2)$ ) even though  $k = 1$ , and may be evidence of “super-convergence”. Often, particularly for highly refined meshes, there are positions in the mesh at which the solution converges at a faster rate than is expected globally. This is known to occur for edge elements on hexahedral grids [219, 211], and for two-dimensional edge elements [59]. However for edge elements on tetrahedra the problem is open.

The right-hand panel of Fig. 7.1 shows the global relative  $(L^2(\Omega))^3$  error as a function of the total number of degrees of freedom (or unknowns) in the problem for  $k = 1$  and 2 when  $\kappa = 1$ . Obviously the second order ( $k = 2$ ) method attains any given accuracy with fewer unknowns than the first order ( $k = 1$ ) method. This indicates the superiority of higher order schemes when the solution is smooth, as is the case for this model problem.

Fig. 7.1. Log-log graphs of the relative  $(L^2(\Omega))^3$  error. *Top:* Error against mesh parameter  $h$ . We show results for edge elements with  $k = 1$  and  $k = 2$  and for wavenumber  $\kappa = 1$  and  $\kappa = 5$ . We also show a graph of the error computed by averaging the edge element solution to the vertices (marked  $\nabla$ ). *Bottom:* Error against the total number of degrees of freedom. These results are consistent with the error analysis in this chapter, and show that quadratic elements can approximate the solution to higher accuracy than linear elements for the same number of degrees of freedom.



## 7.4 The ellipticized Maxwell system

It is reasonable to think of using standard continuous piecewise degree- $k$  finite elements to approximate Maxwell's equations. Success in this would allow the use of standard software and graphical interfaces for the electromagnetic problem. Indeed such methods have been used successfully in engineering computations [56].

Consider, for example, the simplified problem considered in Section 7.2, and

suppose again, for simplicity, that the right-hand side  $F \in H(\text{div};\Omega)$  is such that  $\nabla \cdot F = 0$ , so that  $E \in X$  (see (4.3) for the definition of  $X$ ) satisfies(7.35)

$$\nabla \times \nabla \times E - \kappa^2 E = F \quad \text{in } \Omega, \tag{7.36}$$

$$\begin{aligned} \nabla \cdot E &= 0 \quad \text{in } \Omega, \\ \nabla \times E &= 0 \quad \text{on } \Gamma. \end{aligned} \tag{7.37}$$

We assume that  $\kappa$  is not a Maxwell eigenvalue for  $\Omega$  so this problem has a unique solution in  $H_0(\text{curl};\Omega)$ . Multiplying (7.35) by a test function  $\varphi \in X$  and integrating by parts results in the usual problem of seeking  $E \in X$  such that(7.38)

$$(\nabla \times E, \nabla \times \varphi) - \kappa^2(E, \varphi) = (F, \varphi) \quad \text{for all } \varphi \in X.$$

As usual, for this simplified problem, we can assume that the fields are real.

One problem is that the sesquilinear form  $(\nabla \times E, \nabla \times \varphi)$  in (7.38) is not coercive. It is thus often suggested to “ellipticize” the variational problem by adding a “penalty term” that helps to control the divergence of the field. In particular, we introduce a parameter  $\gamma > 0$  such that  $E \in X_N$  (see (3.65)) satisfies(7.39)

$$(\nabla \times E, \nabla \times \varphi) + \gamma(\nabla \cdot E, \nabla \cdot \varphi) - \kappa^2(E, \varphi) = (F, \varphi) \quad \text{for all } \varphi \in X_N.$$

Here we have now restricted to  $X_N$  since we need  $\nabla \cdot E \in L^2(\Omega)$ .

Our first observation is that if  $\gamma$  is large enough then the solution of (7.35)–(7.37) and of the solution of (7.39) are identical. Of course, we have already shown that (7.35)–(7.37) have a unique solution in  $X_{N,0}$  (see Corollary 4.19) and this is also a solution of (7.39). We need only show that this is the only solution.

To show uniqueness of the solution of (7.39), we start by using the Helmholtz decomposition to write the solution  $E \in X_N$  as  $E = E_0 + \nabla p$  for some  $p \in G = \{q \in H_0^1(\Omega) \mid \Delta p \in L^2(\Omega)\}$  and  $E_0 \in (\nabla G)^\perp$ . Then choosing  $\varphi = \nabla \xi$ , for some  $\xi \in G$ , we have

$$\gamma(\nabla \cdot (E_0 + \nabla p), \nabla \xi) - \kappa^2(E_0 + \nabla p, \nabla \xi) = (F, \nabla \xi).$$

Using the fact that  $\nabla \cdot E_0 = \nabla \cdot F = 0$  we see that  $\gamma(\Delta p, \Delta \xi) = \kappa^2(\nabla p, \nabla \xi) = 0$  for all  $\xi \in G$ . Integrating by parts we obtain  $(\Delta p + (\kappa^2/\gamma)p, \Delta \xi) = 0$ . Hence, we conclude that  $p$  satisfies

$$\begin{aligned} \Delta p + \left(\frac{\kappa^2}{\gamma}\right)p &= 0 \quad \text{in } \Omega, \\ p &= 0 \quad \text{on } \Gamma. \end{aligned}$$

Then, if  $\gamma$  is chosen so that  $\kappa^2/\gamma$  is less than the first Dirichlet eigenvalue of  $\Omega$ , the only solution of the above system is  $p = 0$ . It would also suffice to assume that  $\kappa^2/\gamma$  is not a Dirichlet eigenvalue for  $\Omega$ , but this would be difficult to ensure in practice. We have proved that  $E = E_0$ , which satisfies (7.38) and we see that

the addition of the term  $\gamma(\nabla \cdot E, \nabla \cdot \varphi)$  does not change the solution of (7.38), but only provides a stabilization of the variational formulation. In particular via Theorem 3.50 and Corollary 3.51 and using the simple geometry of  $\Omega$ , we have that there is a constant  $C > 0$  such that, for all  $u \in X_N$ ,

$$\|u\|_{(L^2(\Omega))^3} \leq C \left\{ \|\nabla \times u\|_{(L^2(\Omega))^3} + \|\nabla \cdot u\|_{(L^2(\Omega))} \right\}.$$

Hence the differential operator in (7.39) is coercive. However, we need to be mindful of the warnings in Section 3.8 regarding the use of finite elements in  $X_N$ .

Now suppose we wish to discretize (7.39) using finite elements. Since  $X_N \subset H_0(\text{curl}; \Omega)$  and  $X_N \subset H(\text{div}; \Omega)$ , Theorem 5.3 implies that any piecewise polynomial subspace must be continuous across faces between elements. Thus the resulting piecewise polynomial subspace must be a subset of  $(H^1(\Omega))^3$ . For example, we could take

$$X_{N,h} = (u_h)^3 \cap X_N,$$

where  $U_b$  is given by (5.55). We could then try to compute the function  $E_b \in X_{N,b}$  that satisfies

$$(\nabla \times E_h, \nabla \times \varphi_h) + \gamma(\nabla \cdot E_h, \nabla \cdot \varphi_h) - \kappa^2(E_h, \varphi_h) = (F, \varphi_h)$$

for all  $\varphi \in X_{N,b}$ . Unfortunately, if  $\Omega$  has re-entrant corners, we know that  $(H^1(\Omega))^3 \cap X_N$  is a closed proper subspace of  $X_N$  (see Lemma 3.56). Thus, if we compute a sequence of fields  $E_b \in X_{N,b}$  for  $b > 0$ , we will find that

$$E_h \rightarrow \tilde{E} \text{ in } (H^1(\Omega))^3 \cap X_N \text{ as } h \rightarrow 0.$$

However,  $\tilde{E} \neq E$ ! [105]. Thus we have the terrible situation that we compute a convergent solution (we can check this “in practice” by examining the solution on successively finer meshes) but the numerical solution converges to the wrong answer.

This bad state of affairs can be corrected either by adding suitable singular functions to the finite element space  $X_{N,b}$  (at least in two dimensions, see [51]) or by modifying the term  $(\nabla \cdot E_b, \nabla \cdot \varphi_b)$  using weight functions that vanish sufficiently fast approaching re-entrant corners or edges on  $\Gamma$  [114]. This implies that we are no longer working in  $(H^1(\Omega))^3 \cap X_N$ , and corrects the difficulty. Use of the method in [114] seems to be the best “fix” if continuous finite elements are to be used.

#### 7.4.1 Discrete ellipticized variational problem

In this section we continue with our assumption that  $\Sigma = \emptyset$ , but now allow a general  $F \in (L^2(\Omega))^3$  and general  $\epsilon_r$  and  $\mu_r$ . We also assume that the wavenumber  $\kappa$  is not a Maxwell eigenvalue for  $\Omega$ . Given that the addition of the term  $\gamma(\nabla \cdot \epsilon_r E, \nabla \cdot \epsilon_r \varphi)$  in (7.39) restores the coercivity of the differential operator for Maxwell’s equations (at least if  $\gamma$  is large enough and  $\epsilon_r = 1$ ), it is natural

to ask if such a stabilization is possible with edge elements. The straightforward answer is ‘no’, because functions in  $X_b$  do not have a well-defined divergence.

However, we can define a discrete divergence via the discrete Helmholtz decomposition. For any  $u \in (L^2(\Omega))^3$  we define  $\nabla_b \cdot u \in S_b$  by(7.40)

$$(\nabla_h \cdot u \xi_h) = -(u, \nabla \xi_h) \quad \text{for all } \xi_h \in S_h .$$

The Lax–Milgram lemma guarantees that  $\nabla_b \cdot u$  is well defined. Now using the test function  $\varphi_b = \nabla \xi_b$  in (7.2) we see that  $\kappa^2(\epsilon_r E_b, \nabla \xi_b) = (F, \nabla \xi_b)$ . Thus,

$$(\nabla_h \cdot (\epsilon_r E_h), \xi_h) = \frac{1}{\kappa^2} (F, \nabla \xi_h),$$

and the solution  $E_b \in X_b$  of (7.2) satisfies(7.41)

$$\begin{aligned} & \left( \mu_r^{-1} \nabla \times E_h, \nabla \times \varphi_h \right) + \gamma (\nabla_h \cdot \epsilon_r E_h, \nabla_h \cdot \epsilon_r \varphi_h) \\ & - \kappa^2 (\epsilon_r E_h, \varphi_h) = (F, \varphi_h) \end{aligned}$$

for all  $\varphi_b \in X_b$ . Naturally, we may be concerned that the stabilization term has ruined uniqueness or otherwise perturbed the solution. However, we have the following result:

**Lemma 7.26** *Provided  $b$  is sufficiently small and  $\gamma$  is sufficiently large (independent of  $b$ ), then(7.41)has a unique solution that is also a solution of (7.2).*

**Proof** We need only prove uniqueness since we already know that the solution of (7.2) is also a solution of (7.41). Thus, we assume  $F = 0$ . Using the discrete Helmholtz decomposition we can write  $E_b = \tilde{E}_{b,0} + \nabla \tilde{p}_b$  for some  $\tilde{p}_b \in S_b$ , and  $\tilde{E}_{b,0} \in X_{0,b}$  where  $(\epsilon_r \tilde{E}_{b,0}, \nabla \xi_b) = 0$  for all  $\xi_b \in S_b$  and hence  $\nabla_b \cdot (\epsilon_r \tilde{E}_{b,0}) = 0$ . Choosing  $\varphi_b = \nabla \xi_b$  for some  $\xi_b \in S_b$  in (7.41) shows that

$$\gamma (\nabla_h \cdot \epsilon_r \nabla \tilde{p}_h, \nabla_h \cdot \epsilon_r \nabla \xi_h) - \kappa^2 (\epsilon_r \nabla \tilde{p}_h, \nabla \xi_h) = 0 .$$

Now let  $qb = \nabla_b \cdot \epsilon_r \tilde{p}_b$ . Then using the definition of the discrete divergence  $\nabla_b \cdot$  to modify the second term on the left-hand side above, we obtain

$$\gamma (q_h, \nabla_h \cdot \epsilon_r \nabla \xi_h) + \kappa^2 (q_h, \xi_h) = 0 \quad \text{for all } \xi_h \in S_h .$$

Hence, again using the definition of the discrete divergence,(7.42)

$$(\overline{\epsilon_r \nabla q_h}, \nabla \xi_h) - \frac{\kappa^2}{\gamma} (q_h, \xi_h) = 0 \quad \text{for all } \xi_h \in S_h .$$

The Poincaré inequality applied to the first term above shows that

$$\left| (\overline{\epsilon_r \nabla q_h}, \nabla q_h) \right| \geq C \| \nabla q_h \|_{L^2(\Omega)}^2 \geq C \| q_h \|_{H^1(\Omega)}^2 .$$

Here  $C$  depends on the lower bound on  $\Re(\epsilon)$  assumed in Section 4.2 but is independent of  $b$ . So if  $\kappa^2/\gamma < C$  we conclude that (7.42) shows that  $qb = 0$ . But, by the definition of  $qb$  and the discrete divergence,

$$0 = (q_h, \tilde{p}_h) = (\nabla_h \cdot \in_r \nabla \tilde{p}_h, \tilde{p}_h) = -(\in_r \nabla \tilde{p}_h, \nabla \tilde{p}_h),$$

and another application of the Poincaré inequality shows that  $\tilde{p}_h = 0$ . Hence  $E_{h,0}$  satisfies (7.2) with zero data and so  $E_{h,0} = 0$ . This completes the uniqueness proof.  $\square$

One advantage of the formulation in (7.41) is that it stabilizes the edge element method for low frequency. In the standard edge element method given by (7.2), control over the discrete divergence of  $\epsilon_r E_h$  is via the lower-order term  $-\kappa^2(\epsilon_r E_h, \varphi_h)$ . As  $\kappa$  decreases this term becomes less significant and the conditioning of the discrete problem deteriorates [220]. With the stabilization term, the problem may be solved down to  $\kappa = 0$  (at least in a special case).

**Lemma 7.27** Suppose  $\Sigma = \emptyset$  and so  $\partial\Omega$  consists of one connected component  $\Gamma$ . Then for all  $E_h \in X_h$  there exists a constant  $C > 0$  independent of  $h$  and  $E_h$  such that

$$\|\nabla \times E_h\|_{(L^2(\Omega))^3} + \|\nabla_h \cdot \in_r E_h\|_{(L^2(\Omega))^3} \geq C \|E_h\|_{(L^2(\Omega))^3}.$$

**Remark 7.28** This result shows that for all  $\kappa$  small enough (i.e. for  $0 \leq \kappa^2 < C$ , where  $C$  is the constant in the above lemma) and if  $\Sigma = \emptyset$  we know that (7.41) is uniquely solvable (i.e. down to  $\kappa = 0$ ).

**Proof of Lemma 7.27** Using the discrete Helmholtz decomposition  $E_h = E_{h,0} + \nabla p_h$  for some  $p_h \in S_h$  and  $E_{h,0} \in X_{h,0}$  and  $\nabla_h \cdot \epsilon_r E_{h,0} = 0$ . By the discrete Friedrichs inequality (Lemma 7.20) there is a constant  $C > 0$  such that  $\|\nabla \times E_{h,0}\|_{(L^2(\Omega))^3} \geq C \|E_{h,0}\|_{(L^2(\Omega))^3}$ . But

$$(\nabla_h \cdot \in_r E_h, \xi_h) = (\nabla_h \cdot \in_r \nabla p_h, \xi_h) = -(\in_r \nabla p_h, \nabla \xi_h).$$

Thus, if  $\xi_h = p_h$ ,

$$(\nabla_h \cdot \in_r E_h, p_h) = -(\in_r \nabla p_h, \nabla p_h).$$

Via this equality, the bounds on  $\epsilon_r$  and the Poincaré inequality,

$$\begin{aligned} \|\nabla p_h\|_{(L^2(\Omega))^3}^2 &\leq C \|\nabla_h \cdot \in_r E_h\|_{(L^2(\Omega))^3} \|p_h\|_{L^2(\Omega)} \\ &\leq C \|\nabla_h \cdot \in_r E_h\|_{L^2(\Omega)} \|\nabla p_h\|_{(L^2(\Omega))^3}. \end{aligned}$$

Hence  $\|\nabla p_h\|_{(L^2(\Omega))^3} \leq C \|\nabla_h \cdot \epsilon_r E_h\|_{L^2(\Omega)}$ . Then, using the orthogonality of the Helmholtz decomposition in  $L_{\epsilon_r}^2(\Omega)$  we have

$$\|E_h\|_{(L^2(\Omega))^3} \leq C \|E_{h,0}\|_{(L^2(\Omega))^3} + \|\nabla p_h\|_{(L^2(\Omega))^3},$$

and using the above inequalities completes the proof.  $\square$

One obvious problem with the previous formulation in (7.41) is that  $\nabla_h \cdot$  is defined via (7.40) and so the evaluation of  $\nabla_h \cdot \epsilon_r E_h$  would require to invert the mass matrix. This can be avoided by using mass lumping (for a detailed discussion of mass lumping see [82]).

It is easier to do this if the data for the problem are divergence-free. Thus, we assume that  $p_b \in S_b$  has been computed via (7.24) and we now wish to compute  $E_{0,b} \in X_{0,b}$  that satisfies (7.25). Since  $\nabla_b \cdot \epsilon_r E_{0,b} = 0$ , we can approximate it in the following way. Suppose we are using linear edge elements (i.e.  $k = 1$ ) on tetrahedra. Then we can approximate the  $L^2(\Omega)$  inner product for functions in  $S_b$  by quadrature as follows:(7.43)

$$(p_h, \xi_h) = \sum_{K \in \tau_h} \int_K p \bar{\xi}_h dV \simeq \sum_{K \in \tau_h} \frac{\text{vol}(K)}{4} \sum_{j=1}^4 p_h(a_j^K) = (p_h, \xi_h)_h,$$

where  $\{a_j^K\}_{j=1}^4$  are the vertices of element  $K$  in the mesh  $\tau_b$ , and  $\text{vol}(K)$  is the volume of  $K$ . This quadrature is exact if  $p_b$  and  $\xi_b$  are piecewise constant. In any case, it defines an inner product  $(\cdot, \cdot)_h$  on  $S_b \times S_b$ . We can then define a new discrete divergence denoted  $\nabla_h^\mathcal{Q}$  using this quadrature. Thus, for  $u \in L^2(\Omega)$ , the function  $\nabla_h^\mathcal{Q} \cdot u \in S_h$  satisfies

$$(\nabla_h^\mathcal{Q} \cdot u, \xi_h)_h = - (u, \nabla \xi_h) \quad \text{for all } \xi_h \in S_h.$$

Since  $(\cdot, \cdot)_h$  is an inner product, the Lax–Milgram lemma proves that  $\nabla_h^\mathcal{Q} \cdot u$  is well defined. However, since the quadrature points in (7.43) are also at the degrees of freedom for the piecewise linear basis functions for  $S_b$ , we see that the mass matrix corresponding to this inner product is diagonal. Thus, we can compute the degrees of freedom of  $\nabla_h^\mathcal{Q} \cdot u$  by simply inverting a diagonal matrix. We can then compute  $E_{0,b} \in X_b$  by solving the following modified stabilized problem:(7.44)

$$\begin{aligned} & \left( \mu_r^{-1} \nabla \times E_{0,h}, \nabla \times \varphi_h \right) + \gamma (\nabla_h^\mathcal{Q} \cdot \in_r E_{0,h}, \nabla_h^\mathcal{Q} \cdot \in_r \varphi_h)_h - \kappa^2 (\in_r E_{0,h}, \varphi_h) \\ & - i\kappa \langle \lambda E_{0,h,T}, \varphi_{h,T} \rangle = (F, \varphi_h) + \langle g, \varphi_{h,T} \rangle + \kappa^2 (\in_r \nabla p_h, \varphi_h) \end{aligned}$$

for all  $\varphi_b \in X_b$ . In [220], Shangyou Zhang and myself showed how this problem (with  $\kappa = 0$ ,  $\sum = \emptyset$ ) could be solved by a multigrid method provided hexahedral elements with edges parallel to the coordinate axes are used.

Other methods of stabilizing the edge element method (7.2) are also possible. Demkowicz *et al.* [122, 286] have advocated the use of a stabilized problem in which the divergence constraint is explicitly enforced via a Lagrange multiplier. For example, in order to solve (7.44) we can compute  $E_{0,b} \in X_b$  and  $q_b \in S_b$  such that(7.45)

$$\begin{aligned} & \left( \mu_r^{-1} \nabla \times E_{0,h}, \nabla \times \varphi_h \right) - \kappa^2 (\in_r E_{0,h}, \varphi_h)_h - i\kappa \langle \lambda E_{0,h,T}, \varphi_{h,T} \rangle + (\in_r \varphi_h, \nabla q_h) \\ & = (F, \varphi_h) + \langle g, \varphi_{h,T} \rangle + \kappa^2 (\in_r \nabla p_h, \varphi_h), \\ & (\in_r E_{0,h}, \nabla \xi_h) = 0, \end{aligned} \tag{7.46}$$

for all  $\varphi_b \in X_b$  and  $\xi_b \in S_b$ . Choosing  $\varphi_b = \nabla q_b$  and using the definition of  $p_b$  shows that  $q_b = 0$ . Thus the solution of the above mixed system is exactly

the solution of (7.25). Given that  $q_b = 0$ , we can further modify the system by replacing (7.46) by  $(\epsilon_r E_{0,b}, \nabla \xi_b) = \bar{B}(q_b, \xi_b)$ , where  $\bar{B}(\cdot, \cdot)$  is a positive-semi-definite bilinear form defined on  $S_b \times S_b$  (i.e.  $\bar{B}(q_b, q_b) \geq 0$  for all  $q_b \in S_b$ ). Possible choices are

$$\hat{b}(q_h, \xi_h) = (q_h, \xi_h) \text{ or } \hat{b}(q_h, \xi_h) = (\nabla q_h, \nabla \xi_h).$$

The choice of  $\bar{B}(\cdot, \cdot)$  is usually motivated by a desire to improve the conditioning of the problem to speed suitable iterative solvers. No definitive conclusion concerning the best choice seems to have been reached so far.

## 7.5 The discrete eigenvalue problem

We now wish to show that we can approximate the cavity resonator problem of Section 4.7. Our presentation will barely scratch the surface of this important and interesting subject. Until recently, there has been a good deal of confusion surrounding finite element methods for this problem. In particular, standard piecewise-linear continuous finite elements methods usually produce spurious modes by which we mean that the computed eigenvalues, or the multiplicity of the eigenvalues, is incorrect regardless of the size of the mesh parameter  $b$ . The situation was clarified by the work of Boffi *et al.* [48, 46, 49], who showed that pointwise convergence of the solution of the discrete source problem is not sufficient to guarantee good convergence of eigenvalues. They proposed an extension to mixed method theory that provides a sufficient condition for eigenvalue convergence. In particular, the work in [49] provides a complete picture of why edge elements are successful for eigenvalue calculations.

Our approach to the cavity problem in the preceding sections has been based on discrete compactness to allow us to treat general coefficients. Thus, rather than take the mixed method approach to the eigenvalue problem, it is natural for us to invoke the theory of Section 2.3.4. Indeed Boffi [47] has shown that his new FortID property of mixed methods is equivalent to discrete compactness.

The problem we wish to approximate is given in Section 4.7. Thus, as in that section, we assume  $\sum = \emptyset$  and  $\Im(\epsilon) = 0$ , together with the standard assumptions from Section 4.2. It follows from Theorem 4.18 that there are nontrivial eigenfunction  $E \in X = H_0(\text{curl}; \Omega)$  and eigenvalues  $\kappa \in \mathbb{R}$  such that (4.24) is satisfied.

Now let  $X_b \subset X$  denote the edge finite element space given by (7.1). This is an edge element space of degree- $k$  piecewise polynomials on a regular tetrahedral mesh. Of course, we could also use edge elements on a mesh of hexahedra with edges parallel to the coordinate axes as described in Chapter 6.

The discrete eigenvalue problem is to find  $E_b \in X_b$ ,  $E_b \neq 0$  and  $\kappa_b \in \mathbb{R}$  such that (7.47)

$$\left( \mu_r^{-1} \nabla \times E_h, \nabla \times \varphi_h \right) = \kappa_h^2 ( \epsilon_r E_h, \varphi_h ), \quad \nabla \varphi_h \in X_h.$$

We immediately see that  $E_b = \nabla p_b \neq 0$  for any  $p_b \in S_b$  is an eigenfunction corresponding to the eigenvalue  $\kappa_b = 0$ . Under the assumption that  $\Omega$  has a

boundary  $\Gamma$  consisting of just one component, the discrete Friedrichs inequality from Lemma 7.20 shows that the only eigenfunctions corresponding to  $\kappa_b = 0$  lie in  $\nabla S_b$  (recall  $S_b$  is given by (7.7)).

If  $\kappa_b \neq 0$  then choosing  $\varphi_b = \nabla \xi_b$  for any  $\xi_b \in S_b$  in (7.47) shows that  $-\kappa_b^2 (\in_r E_h, \nabla \xi_b) = 0$ , for all  $\xi_b \in S_b$ . Thus,  $\in_r E_b$  is discrete divergence-free and so  $E_b$  lies in  $X_{0,b}$  (see (7.9)). For the purposes of analyzing the physically relevant eigenfunctions corresponding to  $\kappa_b > 0$ , it suffices to analyze the problem of finding  $E_b \in X_{0,b}$ ,  $E_b \neq 0$ , and  $\kappa_b \in \mathbb{R}$ ,  $\kappa_b > 0$ , such that

$$\left( \mu_r^{-1} \nabla \times E_h, \nabla \times \varphi_h \right) = \kappa_b^2 (\in_r E_h, \varphi_h) \text{ for all } \varphi_h \in X_{0,h}.$$

As in Section 4.7, we rewrite this variational problem as the problem of finding  $E_b \in X_{0,b}$ ,  $\kappa_b > 0$ , such that

$$\left( \mu_r^{-1} \nabla \times E_h, \nabla \times \varphi_h \right) + (\in_r E_h, \varphi_h) = (\kappa_b^2 + 1) (\in_r E_h, \varphi_h)$$

for all  $\varphi_b \in X_{0,b}$ .

Now we can define the operator  $\tilde{K}_h : L^2_{\in_r}(\Omega) \rightarrow L^2_{\in_r}(\Omega)$  analogously to  $K$  in (4.26), so that if  $F \in L^2_{\in_r}(\Omega)$  then  $\tilde{K}_h F \in X_{0,b}$  satisfies (7.48)

$$\left( \mu_r^{-1} \nabla \times \tilde{K}_h F, \nabla \times \varphi_h \right) + (\in_r \tilde{K}_h F, \varphi_h) = (\in_r F, \varphi_h) \text{ for all } \varphi_h \in X_{0,h}.$$

As is the case for  $K$  (using essentially the same proof), we can see that  $\tilde{K}_h$  is self-adjoint using the  $L^2_{\in_r}(\Omega)$  inner product. Furthermore, since  $\tilde{K}_h$  is related via a constant to the operator  $K_b$  defined in (7.26), Lemma 7.11 shows that  $\{\tilde{K}_h\}_{h \in \Lambda}$  is pointwise convergent. In addition, Theorems 7.14 and 7.18 show that  $\{\tilde{K}_h\}_{h \in \Lambda}$  is collectively compact (recall  $\Lambda$  is the discrete set of mesh sizes defined in Section 7.3.2). Thus, if we define  $\mu_h = 1 / (1 + \kappa_h^2)$  then the discrete eigenvalue problem is equivalent to finding  $\mu_b$  and  $E_b \in X_{0,b}$ ,  $E_b \neq 0$  such that (7.49)

$$\tilde{K}_h E_h = \mu_h E_h.$$

We can now apply Theorem 2.52 to prove the following result:

**Theorem 7.29** Suppose  $\mu$  is an eigenvalue of  $K$  of multiplicity  $m$ . Then, under the conditions on the data given in Section 4.2 and, in addition, under the restrictions noted in this section there are exactly  $m$  discrete eigenvalues  $\mu_{h,j}$ ,  $j = 1, \dots, m$  of (7.48) such that

$$|\mu - \mu_{h,j}| \rightarrow 0, \quad 1 \leq j \leq m, \quad \text{as } h \rightarrow 0.$$

**Remark 7.30** Using the error estimate for pointwise convergence in Theorem 7.11, we could derive the order of convergence estimates, depending on the regularity of the eigenfunctions. For example, if linear ( $k = 1$ ) edge elements are used and if all the eigenfunctions lie in  $(H^1(\Omega))^3$ , as occurs for a cube, we know that the error in the eigenvalues is  $O(h^2)$ . For a much more detailed analysis of eigenvalue problems see [49].

**Proof of Theorem 7.29** Using (4.7) and (7.48) we may write

$$\begin{aligned} (\tilde{K} - \tilde{K}_h) v_i, v_j &= (\nabla \times (\tilde{K} - \tilde{K}_h) v_i, \nabla \times \tilde{K} v_j) \\ &= (\nabla \times (\tilde{K} - \tilde{K}_h) v_i, \nabla \times (\tilde{K} - \tilde{K}_h) v_j). \end{aligned}$$

Hence, since  $E(\mu)$  is finite-dimensional, (2.23) may be rewritten as

$$|\mu - \mu_{h,j}| \leq C \left\{ \max_l \left\| \nabla \times (\tilde{K} - \tilde{K}_h) v_i \right\|_{(L^2(\Omega))^3}^2 + \max_l \left\| (\tilde{K} - \tilde{K}_h) v_i \right\|_{(L^2(\Omega))^3}^2 \right\}.$$

The pointwise convergence of  $\tilde{K}_h$  to  $\tilde{K}$  in  $H(\text{curl}; \Omega)$  shown in Lemma 7.11 proves that both terms on the right-hand side above vanish as  $h \rightarrow 0$ .  $\square$

In Fig. 7.2 we show a convergence study for computing eigenvalues using a tetrahedral finite element code base on edge elements with  $k = 1$  (from the Modulef library). The domain is  $\Omega = [0, 1]^3$  and it is then known analytically that the first three non-zero eigenvalues are as follows:

| Group index $l$ | Eigenvalue $\kappa_l$ | Multiplicity $m_l$ |
|-----------------|-----------------------|--------------------|
| 1               | $2(2\pi)^2$           | 3                  |
| 2               | $3(2\pi)^2$           | 2                  |
| 3               | $5(2\pi)^2$           | 6                  |

Having computed the first few (at least 11) non-zero eigenvalues, we plot the error (7.50)

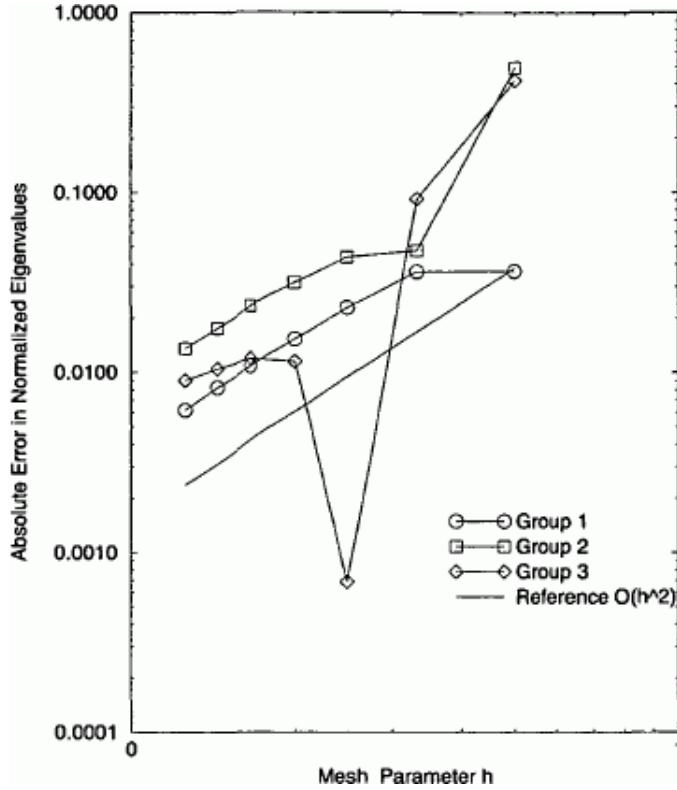
$$\frac{1}{\pi^2} \left| \kappa_l - \frac{1}{m_l} \sum_{j=1}^{m_l} \kappa_{j,l,h} \right|, \quad l = 1, 2, 3,$$

where the  $m_l$  discrete eigenvalues  $\kappa_{j,l,h}$ ,  $1 \leq j \leq m_l$ , converge to  $\kappa_l$ . We see that, since the eigenfunctions are smooth functions of position in this case, the weighted average of discrete eigenvalues converges at a rate  $O(h^2)$ . This is consistent with our theoretical prediction. For further numerical results, and details of implementation, see e.g. [205].

The result in Fig. 7.2 is from [289] where the  $h$  version of the finite element method (as described in this chapter) was compared to the  $p$ -version (see Section 8.4 for an introduction to the more general  $hp$  version of the finite element method). Unfortunately the theory in that paper was based on [210] which contains an error. Therefore the proof in [289] is not correct and convergence of the  $p$ -version is an open problem to date. For recent results on  $hp$  methods for the eigenvalue problem see [4].

This section has shown that edge elements may be used successfully to compute solutions of eigenvalue problems in electromagnetic applications. One area of considerable practical importance is the computation of modes in a waveguide. This problem reduces to a two-dimensional eigenvalue problem but with modified operators. Edge elements have proved to be a useful discretization tool (see,

Fig. 7.2. Error in approximating the first three normalized eigenvalues for the Maxwell eigenvalue problem on the unit cube. We show the error defined by (7.50) as a function of mesh size. Convergence with a rate  $O(h^2)$  is observed as expected for  $k = 1$  edge elements.<sup>2</sup>



# 8 TOPICS CONCERNING FINITE ELEMENTS

## 8.1 Introduction

The basic finite elements presented in Chapters 5 and 6 can be considerably elaborated to try to improve their performance. Of course, the edge elements presented in these chapters are not the only possible family of elements obeying the discrete de Rham diagram. In this section, we start by quickly reviewing another family of elements (termed here the “second family”) also due to Nédélec [235]. These elements are defined for a tetrahedral mesh and have been used extensively by Mur [231] for time domain electromagnetic modeling. Indeed, Mur independently discovered the lowest-order element of this type. These elements have the attractive property of offering superior approximation properties in the  $(L^2(\Omega))^3$  norm compared to the first family (see Theorems 8.15 and 5.41). On the other hand, for a given grid, the second family uses more degrees of freedom than the first family (for  $k = 1$  the second family on tetrahedra uses twice as many degrees of freedom as the first family). Having never directly compared the two families in the frequency domain, I cannot recommend one over the other.

Continuing our development of edge elements, we then discuss the approximation of curved boundaries. Our goal is very limited. In Chapters 10 and 11, we shall need to approximate Maxwell's equations on a domain with a spherical outer boundary. We need to have a method that will provide optimal convergence in this simple case. As we shall see, the results for more general curved boundaries are by no means complete.

We end this chapter with a brief introduction to  $hp$  finite element methods for Maxwell's equations. The elements we shall discuss (from [120]) underlie the  $hp$  code of Demkowicz and co-workers [286, 255, 257]. There are many subtleties involved in writing code for an  $hp$  finite element method [270] and this book, which focuses mainly on the  $h$  version, is not the place for discussing such details. We note that it is currently an open problem to prove error estimates for general  $hp$  methods for Maxwell's equations (see [218] for a start in this direction).

There are many more generalizations and modifications of edge elements in the literature. All are useful in certain circumstances, but are too specialized for this book. Examples include

- (1) *Enhanced elements* Various enhancements of edge elements have been proposed to obtain special properties. For example, the edge element of [135, 136] has additional basis functions designed to allow mass lumping for time dependent computations.

- (2) *Non-conforming elements* It is worth pointing out that standard non-conforming finite elements like those in [109] result in an inconsistent method when applied to Maxwell's equations [223]. A new family of non-conforming methods suitable for Maxwell's equations has been proposed in [131]. So far this has only been analyzed for low-frequency problems, but deserves to be looked at for the general Maxwell problem.
- (3) *Prism elements* These elements are due to Nédélec [235] (see also [247, 150]) and have been used by Nicaise [238] for building refined meshes near singularities in the solution along edges of the domain. For these elements the reference domain is a product domain formed by the reference triangle  $T$  with vertices  $(0, 0)^T, (1, 0)^T, (0, 1)^T$  and the unit interval  $\hat{I} = (0, 1)$  so  $\hat{K} = T \times \hat{I}$ . The elements on this domain are also tensor products. Let  $R_k(T)$  denote the set of edge basis functions on  $T$  of degree at most  $k$  (see Remark (5.29)). For example, if  $k = 1$  and the first family is used, then  $\hat{u} \in R_1(T)$  if and only if

$$\hat{u} = \begin{pmatrix} a + bx_2 \\ c - b\hat{x}_1 \end{pmatrix}$$

for some constants  $a, b$  and  $c$ . Then the finite element functions on  $\hat{K}$  are drawn from

$$\hat{P}_k = \left( R_k(\hat{T}) \otimes P_k(\hat{I}) \right) \times \left( P_k(\hat{T}) \otimes R_{k-1}(\hat{I}) \right).$$

Degrees of freedom may then be specified as suitable line and surface integrals motivated by the product structure of the element. We shall now give two examples.

The lowest order element has  $k = 1$ . In this case, we have  $\hat{u} \in \hat{P}_1$  provided

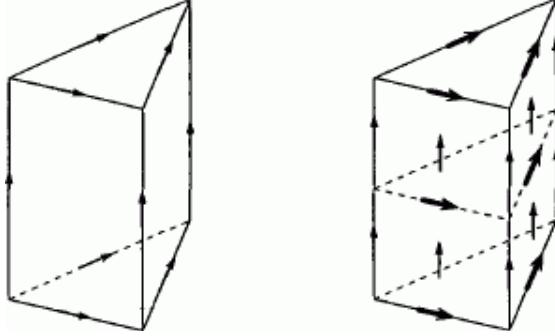
$$\hat{u} = \begin{pmatrix} (a_1 + a_2\hat{x}_3) + (a_3 + a_4\hat{x}_3)\hat{x}_2 \\ (a_5 + a_6\hat{x}_3) + (a_3 + a_4\hat{x}_3)\hat{x}_1 \\ a_7 + a_8\hat{x}_1 + a_9\hat{x}_2 \end{pmatrix}.$$

This gives nine degrees of freedom that are uniquely specified by the nine edge degrees of freedom

$$M_e(\hat{K}) = \left\{ \int_{\hat{e}} u \cdot \hat{\tau} ds \text{ for the nine edges } \hat{e} \text{ of } \hat{K} \text{ with unit tangent } \hat{\tau} \right\}.$$

If  $k = 2$ , there are 36 basis functions in  $\hat{P}_2$  with the usual degrees of freedom for  $R_2(T)$  on the upper and lower triangular faces of the prism and on the surface  $\hat{x}_3 = \frac{1}{2}$  (this gives a total of  $3 \times 8 = 24$  degrees of freedom). In addition there are 12 degrees of freedom along lines parallel to the  $\hat{x}_3$ -axis. More precisely, there are two integral degrees of freedom along each of 12 lines positioned at the Lagrange interpolation points for  $P_2(T)$  (see Fig. 8.1). Provided every prismatic element  $K \in \tau_b$  is a right prism (the triangular base is at right angles to the rectangular faces) a combination of the techniques

Fig. 8.1. Degrees of freedom for the two lowest order prism elements. *Left:*  $k = 1$ . *Right:*  $k = 2$ . The integral degrees of freedom are marked by arrows on the appropriate edges or faces. For the  $k = 2$  element, only the degrees of freedom on the visible faces of the prism are shown, and multiple degrees per edge are shown with bold arrows.



in Chapters 5 and 6 can be used to verify the existence of an interpolant with the usual approximation properties.

- (4) *Pyramidal elements* Grarinariu and Hiptmair [148] have shown that it is possible to derive edge elements on pyramids (here the reference domain is the pyramid with vertices  $(0, 0, 0)^T$ ,  $(1, 0, 0)^T$ ,  $(0, 1, 0)^T$ ,  $(1, 1, 0)^T$  and  $(\frac{1}{2}, \frac{1}{2}, 1)^T$ ). Such elements might be useful for combining tetrahedral and hexahedral elements in one computation.
- (5) *Super-convergence* The global convergence rate of a finite element method depends on the norm used for measuring the convergence rate and is limited by the smoothness of the solution and the maximum degree of the polynomial space contained in the element-wise basis. In general, when interpolating smooth functions, if the finite element space contains all polynomials of degree  $\ell$  on each element, the interpolation error in the  $H$  norm is  $O(h^{\ell+1-\delta})$  (this is a rule of thumb and not a theorem — but the statement can be made rigorous using the Bramble–Hilbert lemma, see [60]). For example, for edge elements with  $k = 1$  and for a smooth function  $u$ , Theorem 5.41 implies that we have  $\|u - r_b u\|_{L(\Omega)}^2 = O(h)$  since  $(P_0)^3 \subset R_1$  but  $(P_1)^3 \not\subset R_1$ . Similarly for  $k = 2$  we have  $(P_1)^3 \subset R_2$  but  $(P_2)^3 \not\subset R_2$ , so  $\|u - r_b u\|_{L(\Omega)}^2 = O(h^2)$  is the best possible rate. Frequently, however, there are certain points in the mesh at which the solution is “super-convergent”, or convergent to higher order than the global rate. In [219] (see also [211] for an improvement on this result), it was shown that edge elements of the first kind on cubes can be super-convergent on special surfaces within the element. Whether this holds for the time harmonic Maxwell system is not known. In addition, although super-convergence is known for triangular edge elements in two dimensions [59], there is no proof of super-convergence in  $R^3$  for tetrahedra. It is also not agreed if super-convergence is seen in computations, although Fig. 7.2 suggests that super-convergence may be observed at the

vertices of the mesh provided averaged values of the fields are used.

## 8.2 The second family of elements on tetrahedra

A potential disadvantage of the first-type elements presented in Chapters 5 and 6 is that using the degree- $k$  elements, it is only possible to obtain an  $O(b^k)$  error estimate for the interpolant in the  $(L^2(\Omega))^3$  norm for the divergence or curl conforming elements. By analogy to error estimates for scalar piecewise  $k$  degree, continuous, finite elements, we might hope for  $O(b^{k+1})$ . This can be obtained with the second family of curl or divergence conforming elements as described by Nédélec [235]. We shall denote the relevant finite element spaces by  $U_h^{(2)} \subset H^1(\Omega), V_h^{(2)} \subset H(\text{curl};\Omega)$  and  $W_h^{(2)} \subset H(\text{div};\Omega)$  to distinguish them from the first family spaces.

In keeping with our description of edge elements on hexahedra in Chapter 6, the presentation here is less detailed than for the first family of elements on tetrahedra presented in Chapter 5. It would be a good idea to be familiar with the material in Sections 5.2 (last part) and 5.3 before starting this chapter.

We assume a regular finite element mesh  $\tau_b$ ,  $b > 0$ , (see Section 5.3) of tetrahedra of maximum diameter  $b$ . As usual each element  $K \in \tau_b$  can be obtained from the reference element  $\hat{K}$  via an affine map  $F_{K\circ} = B_{K\circ} + b_K$  where  $B_K$  is an invertible matrix.

### 8.2.1 Divergence conforming element

Now we define the second family of divergence conforming finite elements.

**Definition 8.1** The *second family of divergence conforming elements* is defined as follows:

- $K$  is a tetrahedron;
- $P_K = (P_\circ)^3$ ;
- the degrees of freedom  $\Sigma_K$  is composed of two sets. For a vector function  $u$  on  $K$  such that  $u \in (H^{1/2+\delta}(K))^3$ ,  $\delta > 0$ , we define

(8.1)

$$M_f(u) = \left\{ \int_f (p \cdot v) q dA \text{ for all } q \in P_k(f) \text{ foreach face } f \text{ of } K \right\}, \quad (8.2)$$

$$M_K(u) = \left\{ \int_K u \cdot q dV \text{ for all } q \in R_{k-1} \right\},$$

where  $v$  is the unit normal to  $f$ .

As in the case of the first-type family analyzed previously, we shall use the standard reference tetrahedron  $\hat{K}$  and transform between  $\hat{K}$  and  $K$  using (5.20). We next prove the analogue of Lemma 5.18 for type 2 elements.

**Lemma 8.2** *The degrees of freedom (8.1) and (8.2) for  $u$  on  $K$  vanish if and only if the degrees of freedom for  $\hat{u}$  on  $\hat{K}$  vanish.*

**Proof** If we change variables in the degrees of freedom (5.23) using (5.20), we obtain, assuming that  $\det(B_K) > 0$ ,

$$\int_K u \cdot q dV = \int_{\hat{K}} (B_K \hat{u}) \cdot (q \circ F_K) d\hat{V} = \int_{\hat{K}} \hat{u} \cdot (B_K^T q) \circ F_K d\hat{V}.$$

Then using Lemma 5.32, we know that if  $q \in R_k$  on  $K$ , then  $B_K^T q \circ F_K \in R_k$  on  $\hat{K}$ . So we can be sure that if all the degrees of freedom (8.2) vanish on  $K$ , they also vanish on  $\hat{K}$  (and, of course, vice versa).

The proof of the invariance of the degrees of freedom (8.1) follows using the same argument as used to prove Lemma 5.18.  $\square$

Next we prove that the elements in Definition 8.1 are divergence conforming by proving the analogue of Lemma 5.20.

**Lemma 8.3** *If  $u \in (P_k)^3$  and all the degrees of freedom of type (8.1) for  $u$  associated with a given face  $f$  vanish, then  $u \cdot v = 0$  on  $f$ .*

**Proof** Since  $u \cdot v \in P_k(f)$ , picking  $q = u \cdot v$  in (8.1) proves the result.  $\square$

The remaining result we need in order to know that the finite elements are well-defined is that the degrees of freedom are unisolvant. As in the case of the first family of divergence conforming elements (see Section 5.4), the number of degrees of freedom and the dimension of  $P_k = (P_k)^3$  are the same and so it suffices to prove that the degrees of freedom uniquely determine a vector polynomial in  $P_k$ .

**Lemma 8.4** *If  $u \in (P_k)^3$  is such that all the degrees of freedom (8.1) and (8.2) vanish, then  $u = 0$ .*

**Proof** We have already shown (in Lemma 8.3) that  $u \cdot v = 0$  on each face  $f$  of  $K$ . As in the proof of Lemma 5.21, we can then use Green's formula to show that  $\nabla \cdot u = 0$  in  $K$ . More precisely, using (3.24) and the vanishing of the normal trace of  $u$  on  $\partial K$ , we obtain

$$\int_K \nabla \cdot u q dV = - \int_K u \cdot \nabla q dV = 0 \text{ for all } q \in P_{k-1},$$

so  $\nabla \cdot u = 0$ . Hence the last equality above holds for any polynomial  $q$  and in particular for  $q \in \tilde{P}_k$ . Now using Lemma 5.27 to write  $(P_{k-1})^3 = R_{k-1} + \nabla \tilde{P}_k$  we see that this implies (using also the the degrees of freedom (8.2)) that (8.3)

$$\int_K u \cdot q dV = 0 \text{ for all } q \in (P_{k-1})^3.$$

Mapping to the reference element and using the same argument as in proof of Lemma 5.21, the fact that  $\hat{u} = 0$  on the faces of  $\hat{K}$  now implies that  $u = (\mathcal{O}_1 \varphi_1, \mathcal{O}_2 \varphi_2, \mathcal{O}_3 \varphi_3)^T$  with  $(\varphi_1, \varphi_2, \varphi_3)^T \in (P_{k-1})^3$ . Choosing  $q = (\varphi_1, \varphi_2, \varphi_3)^T$  in (8.3) proves that  $u = 0$ .  $\square$

The previous lemmas show that the second family of elements in Definition 8.1 is divergence conforming and unisolvant as we summarize in the next theorem.

**Theorem 8.5** A vector field  $u \in (P_k)^3$  is entirely determined by the degrees of freedom (8.1) and (8.2). Moreover, the space  $W_h^{(2)}$  of finite element functions on a mesh  $\tau_h$  of  $\Omega$  defined by Definition 8.1 is divergence conforming so that  $W_h^{(2)} \subset H(\text{div}; \Omega)$ .

**Remark 8.6** We can summarize the definition of the second family divergence conforming space without reference to degrees of freedom by (8.4)

$$W_h^{(2)} = \left\{ u_h \in H(\text{div}, \Omega) \mid u_h|_K \in (P_k)^3 \text{ for all } K \in \mathcal{T}_h \right\}.$$

Having proved the unisolvence and divergence conforming properties of second kind divergence conforming elements, we can define a local interpolant using the degrees of freedom (8.1) and (8.2). Thus, if  $u \in H(\text{div}; \Omega)$ , we define  $w_K u \in (P_k)^3$  by requiring that

$$\begin{aligned} \int_f (u - w_K u) \cdot v q dA &= 0 \quad \text{for all } q \in P_k(f), \text{ for each face } f, \\ \int_K (u - w_K u) \cdot q dV &= 0 \quad \text{for all } q \in R_{k-1}. \end{aligned}$$

Then we can define a global interpolant  $w_h$  from  $(H^{1/2 + \delta}(\Omega))^3$ ,  $\delta > 0$ , onto  $W_h^{(2)}$  by

$$(w_h u)|_K = w_K u \text{ for all } K \in \mathcal{T}_h.$$

We have not used notation to distinguish between the interpolants for the two families of curl conforming finite element space, since in practice it will always be quite clear which space we use.

We have the following error estimate proved in the same way as for Lemma 5.25.

**Theorem 8.7** Let  $\tau_h$  be a regular mesh. Then there is a constant  $C$  independent of  $h$  and  $u$  such that

$$\|u - w_h u\|_{(L^2(\Omega))^3} \leq Ch^3 \|u\|_{(H^2(\Omega))^3},$$

for  $\delta \leq s \leq k+1$ ,  $\delta > 0$ .

**Remark 8.8** The case  $k = 1$  is of practical interest. In this case, assuming  $u$  is sufficiently regular,

$$\|u - w_h u\|_{(L^2(\Omega))^3} \leq Ch^2 \|u\|_{(H^2(\Omega))^3}.$$

This compares to a best case of  $O(h)$  for first kind elements in (5.25). On the other hand, this element has three times as many degrees of freedom as the first family on the same mesh.

**Proof of Theorem 8.7** The proof of this result follows the same outline as the proof of Theorem 5.25. We note that the invariance of the degrees of freedom in Lemma 8.2, and the invariance of  $(P_k)^3$  under the map (5.20), shows that  $\widehat{w_K u} = w_{\widehat{K}} \widehat{u}$ . Mapping to the reference element as before and using Theorem 5.5 proves the estimate.  $\square$

## 8.2.2 Curl conforming element

The lowest order space of the second family of curl conforming elements was discovered independently by Mur [231] and Nédélec [233]. Mur has used the elements extensively in electromagnetic computations (see, e.g. [230]). Nédélec [235] provides an analysis of the interpolation error and extends the construction to arbitrary order. It is his analysis we follow here.

As in the case of the divergence conforming elements in the previous section, it is possible to construct a space of curl conforming functions using vector polynomials of degree exactly  $k$ .

**Definition 8.9** This second family of curl conforming finite elements is defined as follows:

- (1) Each element  $K$  is a tetrahedron.
- (2)  $P_k = (P_k)^3$  for some positive integer  $k$ .
- (3) If  $e$  denotes an edge of  $K$  with unit tangent vector  $\tau$  and  $f$  denotes a face of  $K$  with unit normal  $v$ , we define, for  $u \in (H^{1/2+\delta}(K))^3$ ,  $\delta > 0$ , and  $\nabla \times u \in (L^q(K))^3$ ,  $q > 2$ ,

$$M_e(u) = \left\{ \int_e (u \cdot \tau) q \, ds \text{ for all } q \in P_k(e) \text{ and all edges } e \text{ of } K \right\}, \quad (8.6)$$

$$M_f(u) = \left\{ \int_f u_T \cdot q \, dA \text{ for all } q \in D_{k-1}(f) \text{ and all faces } f \text{ of } K \right\}, \quad (8.7)$$

$$M_K(u) = \left\{ \int_K u \cdot q \, dV \text{ for all } q \in D_{k-2}(K) \right\}.$$

Then

$$\sum_K = M_e(u) \cup M_f(u) \cup M_K(u).$$

Here  $D_{k-1}(f)$  is the analogue of  $D_{k-1}$  in two dimensions given by  $D_{k-1}(f) = (P_{k-2}(f))^2 \oplus P_{k-1}(f) \times$  of vector functions tangential to  $f$ , and  $u_T$  (in (8.6)) is interpreted as a vector with two components in the plane of  $f$ .

As in the case of the first family of edge elements, we establish conformance and unisolvence first, and later turn to the error estimates. Of course, the basis functions in  $(P_k)^3$  are invariant under the transformation (5.33). Furthermore, the number of basis functions is

$$\dim(P_k^3) = \frac{1}{2}(k+1)(k+2)(k+3),$$

and this is exactly the number of degrees of freedom in (8.9). Curl conformance and unisolvence are established by a series of results. First we show that even though the elements are not affine invariant it is still possible to handle the elements via affine maps.

**Lemma 8.10** *The degrees of freedom (8.5)–(8.7) for a function  $u$  on  $K$  vanish if and only if they vanish for  $\hat{u}$  on  $\hat{K}$ .*

**Proof** Using the change of variables (5.33) and (5.35), we have

$$\int_e u \cdot \tau q ds = \gamma_e \int_{\hat{e}} \hat{u} \cdot \hat{\tau} \hat{q} d\hat{s},$$

where  $\gamma_e = \pm 1$  is the constant resulting from the change of variables (see Chapter 5). Thus, the degrees of freedom (8.5) vanish on  $K$  and  $\hat{K}$  simultaneously.

Next we consider facial degrees of freedom of type (8.6). In this case we must use the fact that functions in  $D_{k-1}(\mathcal{f})$  must be transformed using (5.20). Hence, up to a constant factor  $\gamma_f$  depending on  $f$  only,

$$\int_f u \cdot q dA = \gamma_f \int_{\hat{f}} \hat{u} \cdot B_K^{-1} B_K \hat{q} d\hat{A} = \gamma_f \int_{\hat{f}} \hat{u} \cdot \hat{q} d\hat{A},$$

where  $\hat{f}$  is the usual two dimensional reference element in the same plane as  $f$ . From this, we see that the degrees of freedom (8.6) vanish simultaneously. The proof that the degrees of freedom (8.7) satisfy the lemma proceeds analogously.  $\square$

The next lemma establishes curl conformance, via Theorem 5.3, by showing that the tangential component of the finite element on a face is entirely determined by values of the degrees of freedom on the same face.

**Lemma 8.11** *Suppose  $u \in (P_3)^3$  is such that all degrees of freedom associated with a face  $f$  and the edges of  $f$  vanish. Then  $u \times v = 0$  on  $f$ .*

**Proof** This proof is similar to the proof of Lemma 5.35 and we shall only sketch the argument. The degrees of freedom (8.5) imply that  $u \cdot \tau = 0$  on  $\partial f$ . Then Green's theorem for the two-dimensional curl and the degrees of freedom (8.6) imply that if  $u_T$  is the tangential component of  $u$  on  $f$ , then

$$\nabla_f \times u_T = 0 \text{ on } f,$$

where  $\nabla_f \times$  is the surface curl defined in Section 3.4. This implies that  $u_T = \nabla_f p$  where  $\nabla_f$  is the surface gradient on  $f$  and  $p \in P_{k+1}(\mathcal{f})$ . Since  $u_T$  has vanishing tangential component on  $\partial f$ , we may choose  $p = 0$  on  $\partial f$  and so

$$p = \lambda_1 \lambda_2 \lambda_3 r,$$

where  $\lambda_i$ ,  $1 \leq i \leq 3$ , are the barycentric functions for  $f$  and  $r \in P_{k+2}(\mathcal{f})$ . Then the degrees of freedom (8.6) and the Divergence Theorem 3.19 in the plane imply that

$$\int_f \lambda_1 \lambda_2 \lambda_3 \tau \nabla \cdot q dA = 0 \text{ for all } q \in D_{k-1}(f) .$$

But, by Lemma 5.13,  $\nabla \cdot D_{k-1} = P_{k-2}$  and so choosing  $\nabla \cdot q = r$  we obtain  $r = 0$ . This proves the lemma.  $\square$

Now we establish unisolvence.

**Lemma 8.12** *If  $u \in (P_k)^3$  and all the degrees of freedom of type (8.5)–(8.7) vanish, then  $u = 0$ .*

**Proof** This proof is similar to the proof of Lemma 5.21. By the previous lemma,  $u \times v = 0$  on  $\partial K$ . Using the three dimensional Stokes formula (3.51) and the volume degrees of freedom

$$\int_K \nabla \times u \cdot q dV = \int_K u \cdot \nabla \times q dV = 0 \text{ for all } q \in (P_{k-1})^3,$$

since  $\nabla \times q \in D_{k-2}(K)$ . Selecting  $q = \nabla \times u$  implies that  $\nabla \times u = 0$ . Now we invoke Lemma 8.10 to conclude that all degrees of freedom vanish on  $K$  and hence  $\nabla \times \hat{u} = 0$  in  $K$  and  $\hat{u} \times \hat{v} = 0$ . The proof now follows the proof of Lemma 5.36 to show that  $\hat{u} = 0$  and hence  $u = 0$ .  $\square$

As in the case of first type elements, the preceding lemmas imply that we can define a finite element subspace of  $H(\text{curl}; \Omega)$  as follows: (8.8)

$$V_h^{(2)} = \left\{ u \in H(\text{curl}, \Omega) \mid u \Big|_K \in P_k^3 \text{ for all } K \in \mathcal{T}_h \right\} .$$

The associated interpolation operator defined using the degrees of freedom (8.5)–(8.7) is still denoted by  $r_b()$ . The second-family divergence conforming space from the previous section and the space  $V_h^{(2)}$  above are linked as before.

**Lemma 8.13** *Let  $W_b$  be defined by (5.28) and  $V_h^{(2)}$  be defined by (8.8) with their associated interpolation operators  $w_b$  and  $r_b$ . Then  $\nabla \times V_h^{(2)} \subset W_b$  and if  $u$  is smooth enough that  $r_b u$  is defined, then*

$$w_h \nabla \times u = \nabla \times r_b u .$$

**Remark 8.14** *This result holds if  $W_b$  and  $w_b$  are taken to be the second-kind spaces from this chapter (i.e.  $W_b^{(2)}$  from Section 8.2.1). For divergence-free functions, the interpolant is identical for these divergence conforming spaces.*

**Proof of Lemma 8.13** The assertion that  $\nabla \times V_h^{(2)} \subset W_b$  is proved in exactly the same way as in the proof of Lemma 5.40. For the facial degrees of freedom, using the two dimensional Stokes formula (3.28) on a given face  $f$ ,

$$\int_f \nabla_f \times (u - r_b u) q dA = \int_{\partial f} (u - r_b u) \cdot \tau q ds + \int_f (u - r_b u) \cdot \vec{\nabla} \times q dA$$

for all  $q \in P_{k-1}(f)$  where  $\nabla_f \times (u - r_h u) = v \cdot \nabla_x (u - r_h u)$ . Thus, using the definition of  $r_h$ , the face and edge degrees of freedom imply that the right-hand side vanishes, so that

$$\int_f \nabla \times (u - r_h u) \cdot V q dA = 0 \text{ for all } q \in P_k(f).$$

But by the Stokes theorem (3.51), for all  $q \in (P_{k-2})^3$ ,

$$\begin{aligned} \int_K \nabla \times (u - r_h u) \cdot q dV &= \int_K (u - r_h u) \cdot \nabla \times q dV \\ &\quad + \int_{\partial K} (u - r_h u) \times v \cdot q dA = 0. \end{aligned}$$

Hence, using the face and volume degrees of freedom,  $w_b \nabla_x (u - r_h u) = 0$  and so  $w_b \nabla_x u = w_b \nabla_x r_h u = \nabla r_h u$ .  $\square$

Finally, we can state an error estimate for this element.

**Theorem 8.15** Let  $\tau_h$  be a regular mesh of  $\Omega$ . Then if  $u \in (H^s(\Omega))^3$ ,  $1 \leq s \leq k$ ,

$$\|u - r_h u\|_{(L^2(\Omega))^3} + h \|\nabla \times (u - r_h u)\|_{(L^2(\Omega))^3} \leq Ch^{s-1} |u|_{(H^{s+1}(\Omega))^3}$$

In addition, both estimates of Theorem 5.41 hold.

**Remark 8.16** Note that if  $k = 1$ , the first error estimate of the theorem shows that the  $(L^2(\Omega))^3$  norm of the interpolation error for a smooth function is  $O(h^2)$  instead of  $O(h)$  for the first family of elements.

**Proof of Theorem 8.15** The proof of this theorem follows the same lines as the proof of Theorem 5.41. Here we have that (5.44) is replaced by

$$\|\hat{u} - \hat{r}_h \hat{u}\|_{(L^2(\Omega))^3} = \|(I - \hat{r}_h)(\hat{u} + \hat{\varphi})\|_{(L^2(\Omega))^3} \text{ for all } \hat{\varphi} \in P_k^3$$

and this allows us to obtain an  $O(h^{s+1})$  error bound in the  $(L^2(\Omega))^3$  norm.  $\square$

Let us again consider the case  $k = 1$  in more detail. There are now two degrees of freedom on each edge and the two basis functions associated with the edge from vertex  $i$  to vertex  $j$  are  $\lambda_{i,j}$  and  $-\lambda_{j,i}$ . Thus on a tetrahedron  $K$  the function  $u \in (P_1)^3$  may be written as

$$u = \sum_{1 \leq i < j \leq 4} \beta_{i,j} \lambda_i \nabla \lambda_j - \beta_{j,i} \lambda_j \nabla \lambda_i,$$

and

$$\nabla \times u = \sum_{1 \leq i < j \leq 4} (\beta_{i,j} + \beta_{j,i}) \nabla \lambda_i \times \nabla \lambda_j.$$

These expressions should be compared to those given for the first family of edge elements given in (5.47)–(5.48). Note that the second family here differs from the

first family only by the addition of gradients to the space (in particular  $\nabla \lambda, \lambda$ ). On a given mesh, when  $k = 1$ , the second family of elements has twice as many degrees of freedom as the first family, but the error in the  $(L^2(\Omega))^3$  norm is  $O(h^2)$  rather than  $O(h)$ .

### 8.2.3 Scalar functions and the de Rham diagram

We now need to discuss the appropriate scalar finite element space to complete the relevant parts of the de Rham diagram for second-type elements. This is easy since the spaces are the same as in Section 5.6. The difference now is that the curl conforming space consists of piecewise  $(P_k)^3$  polynomials, so the scalar space must consist, at least, of piecewise  $P_{k+1}$  polynomials. The appropriate scalar space turns out to be

$$U_h^{(2)} = \left\{ P_h \in H^1(\Omega) \mid p_h \Big|_K \in P_{K+1} \text{ for all } K \in \mathcal{T}_h \right\}.$$

With this choice of scalar space, and using the second-family edge space  $V_h^{(2)}$  defined above and the face space  $W_h$  defined in Section 5.4, we have the following discrete de Rham diagram:(8.9)

$$\begin{array}{ccccccc} H^1(\Omega) & \xrightarrow{\nabla} & H(\text{curl}; \Omega) & \xrightarrow{\nabla \times} & H(\text{div}, \Omega) & \xrightarrow{\nabla \cdot} & L^2(\Omega) \\ \cup & & \cup & & \cup & & \\ U & & V & & W & & \\ \pi_h \downarrow & & r_h \downarrow & & w_h \downarrow & & P_{0,h} \downarrow \\ U_h^{(2)} & \xrightarrow{\nabla} & V_h^{(2)} & \xrightarrow{\nabla \times} & W_h & \xrightarrow{\nabla \cdot} & Z_h \end{array}$$

where we use the first-type edge space for  $W_h$  and for  $Z_h$ . Of course,  $\pi_h$  and  $r_h$  are the  $U_h^{(2)}$  and  $V_h^{(2)}$  interpolants, respectively.

A final remark is necessary concerning the second-family. It is possible to define second family edge elements on hexahedra [235]. However, for this family, the discrete de Rham diagram does not hold and hence these elements do not fit the theory of this book. Despite this, elements of this type can be used (for an application in elasticity and a modification to mixed method theory to handle such elements, see [29, 30]). They have also been used in time domain Maxwell simulations [84] with considerable success, but there exists the possibility of spurious modes appearing in the computed solution.

## 8.3 Curved domains

In the majority of this book, we assume that the computational domain  $\Omega$  is a Lipschitz polyhedron. This is mainly so that we can assert that  $\Omega$  is exactly covered by a tetrahedral mesh. However, in Chapters 10 and 11 we wish to use an auxiliary boundary  $\Sigma$  that is a sphere. Thus, we need to discuss how to approximate a simple curved domain of this type.

For Laplace's equation and first-order scalar elements, the usual way to approximate a curved domain  $\Omega$  is to approximate the domain by a polyhedron  $\Omega_b$ , consisting of a union of regular tetrahedra with vertices in  $\bar{\Omega}$ . This mesh is chosen so that the skin  $(\Omega \setminus \Omega_b) \cup (\Omega_b \setminus \Omega)$  has small volume. In the case of Laplace's equation, it is then possible to show that the “variational crime” of not using the exact computational domain does not affect the order of accuracy of the method. Of course, for higher-order elements this crude boundary approximation adversely impacts the convergence rate. An analysis of this type has not been carried out for Maxwell's equations. It would be interesting because of the perturbation of the tangential boundary condition, the non-coercive nature of the Maxwell problem, and the perturbation of the divergence condition introduced by the approximation.

In view of the fact that the analysis is unavailable for edge element approximation on a perturbed domain, we now outline a strategy for dealing with a smooth curved boundary. This is certainly sufficient for the problems in this book where we only consider a spherical curved boundary. For piecewise smooth Lipschitz domains, a more advanced strategy would be needed (cf. Bernardi [40] for a start in this direction).

We start by presenting a method due to Dubois [132] for fitting smooth boundaries exactly. Error estimates are only proved for first-order edge elements of the first kind, and we shall summarize the known results in this area. However, we shall not use Dubois' method in the way he advocates, so we shall only sketch his theory. Instead, we make an essentially trivial observation that allows us to map the curvilinear domain, using Dubois methods, to a polyhedral domain (and a perturbed Maxwell system) and hence apply our theory of edge elements on polyhedral domains. It would be interesting, and of practical interest, to examine isoparametric type mapping methods for edge elements. In this regard, we should note a warning. The method we advocate will work for tetrahedral elements. In addition, if the polyhedral domain obtained by mapping the curvilinear domain can be meshed by rectilinear hexahedral elements as discussed in Chapter 6, these elements may also be used. However, if hexahedral elements are mapped to target elements that are not parallelepipeds, recent results of [17] suggest that there are situations under which non-optimal convergence rates (or even non-convergence) may be observed. Thus, mapped hexahedral elements are safest if used in the manner we shall outline in Section 8.3.2. Nevertheless, curvilinear hexahedral elements are in common use in engineering codes [110, 272, 177]. As we shall see, the theory of edge elements on curvilinear domains is not well developed. For example, the Dubois method, which is akin to an isoparametric technique, is only justified for smooth boundaries and the lowest-order vertex, edge and face elements described in Chapter 5 .

### 8.3.1 Locally mapped tetrahedral meshes

In this section we follow Dubois [132]. We suppose that the simply connected domain  $\Omega$  has a boundary consisting of two disjoint connected components, one

of which is denoted by  $\Gamma$  and is the boundary of a Lipschitz polyhedron and the other denoted by  $\Sigma$  which is  $C^2$  regular. By this we mean that  $\Sigma$  is such that the maps in Definition 3.1 are assumed to be  $C^2$  rather than just Lipschitz. We suppose  $\Omega$  is covered by a family of curvilinear meshes  $\tau_b$ ,  $b > 0$ , of disjoint elements  $K \in \tau_b$  of maximum diameter  $b$  such that

- (1) if  $K$  is interior to  $\Omega$  or  $K$  shares at most one point with  $\Sigma$  or  $K \cap \Gamma \neq \emptyset$ , then  $K$  is a tetrahedron (we assume that  $b$  is small enough that no element intersects both  $\Gamma$  and  $\Sigma$ );
- (2) if  $K$  shares a face or an edge with  $\Sigma$ , then it is the image of the reference tetrahedron under an invertible  $C^2$  map  $F_K : K \rightarrow K$ , which we shall give shortly;
- (3) the elements satisfy the usual finite element mesh geometric constraints given in Section 5.3.

We can summarize these properties by saying that  $\tau_b$  is a tetrahedral mesh except for elements sharing an edge or face with  $\Sigma$  which are allowed to be curvilinear tetrahedra. Provided the mesh is refined appropriately, we can assume that every curvilinear element satisfying condition (2) above has at most one face or one edge on  $\Sigma$ .

In order to specify the map  $F_K$  in part (2), we make use of a projection  $P_\Sigma$  which projects points close to  $\Sigma$  onto  $\Sigma$ . More precisely it is possible to show that there is a neighborhood  $\Omega_\Sigma \subset \Omega$  of  $\Sigma$  such that if  $x \in \Omega_\Sigma$ , then there is a unique point  $y \in \Sigma$  depending on  $x$  such that  $x - y$  is normal to  $\Sigma$  at  $y$ . Then we define  $P_\Sigma x = y$ , so  $P_\Sigma$  just projects points normally onto  $\Sigma$ . If  $\Sigma$  is the sphere of radius  $R$  (as will be the case for our applications), then(8.10)

$$P_\Sigma x = Rx / |x|$$

is well defined provided  $x \neq 0$ . In general, the use of  $P_\Sigma$  rules out piecewise smooth boundaries and this deficiency needs to be addressed in the future. We also need to assume that  $b$  is chosen sufficiently small so that all elements intersecting  $\Sigma$  lie in  $\Omega_\Sigma$ , and define  $\tau_{b,\Sigma} = \{K \in \tau_b \mid K \cap \Sigma \neq \emptyset\}$ . Then we suppose that each  $K \in \tau_{b,\Sigma}$  is such that  $K \subset \Omega_\Sigma$ . Since  $\Omega_\Sigma$  is open, this is always possible.

Now assume that  $K \in \tau_b$  has one edge on  $\Sigma$  and that the vertices  $a_1$  and  $a_2$  of  $K$  lie on  $\Sigma$ . This edge must be chosen to be  $P_\Sigma([a_1, a_2])$  where  $[a_1, a_2]$  is the straight line from  $a_1$  to  $a_2$ . Thus,

$$P_\Sigma([a_1, a_2]) = \left\{ x \in \sum \mid x = P_\Sigma(ta_1 + (1-t)a_2) \text{ for } 0 \leq t \leq 1 \right\}.$$

In the same way, if  $K \in \tau_b$  shares a face with  $\Sigma$ , this face is also described using the projection  $P_\Sigma$ . Suppose the face has vertices  $a_1, a_2$  and  $a_3$  all on  $\Sigma$ . Then the curvilinear face is given by

$$\begin{aligned} P_\Sigma([a_1, a_2, a_3]) = & \left\{ x \in \sum \mid x = P_\Sigma(\lambda_1 a_1 + \lambda_2 a_2 + \lambda_3 a_3), \right. \\ & \left. \lambda_j \geq 0, 1 \leq j \leq 3 \text{ and } \lambda_1 + \lambda_2 + \lambda_3 = 1 \right\}. \end{aligned}$$

The reason for these specific assumptions is to provide a concrete method for constructing  $F_K$  that ensures that elements do not overlap or leave some part of  $\Omega$  or  $\Sigma$  uncovered.

Using the notions just described, we can give Dubois' definition of  $F_K : K \rightarrow K$  for a curvilinear tetrahedron (see also [208]). Let  $K$  be the reference tetrahedron with barycentric coordinate functions  $\hat{\lambda}_j$ ,  $1 \leq j \leq 4$ . Then:

- (1) if  $K$  has a single curved edge between  $a_1$  and  $a_2$  (obtained by mapping the vertices  $\hat{a}_1$  and  $\hat{a}_2$  of the reference element), we define(8.11)

$$F_K(\hat{x}) = \left(1 - \hat{\lambda}_3 - \hat{\lambda}_4\right) P_{\Sigma} \left( \frac{\hat{\lambda}_1 a_1 + \hat{\lambda}_2 a_2}{\hat{\lambda}_1 + \hat{\lambda}_2} \right) + \hat{\lambda}_3 a_3 + \hat{\lambda}_4 a_4;$$

- (2) if  $K$  shares a single curvilinear face having vertices  $a_1$ ,  $a_2$  and  $a_3$  with  $\Sigma$ , then

(8.12)

$$F_K(\hat{x}) = \left(1 - \hat{\lambda}_4\right) P_{\Sigma} \left( \frac{\hat{\lambda}_1 a_1 + \hat{\lambda}_2 a_2 + \hat{\lambda}_3 a_3}{\hat{\lambda}_1 + \hat{\lambda}_2 + \hat{\lambda}_3} \right) + \hat{\lambda}_4 a_4 .$$

Now that we have this careful definition of  $F_K$ , we can show that  $\tau_b$  does cover  $\Omega$  and that the edges and faces are defined without needing to explicitly use the tetrahedra that contain them.

**Lemma 8.17** *The edges and faces of  $\tau_b$  on  $\Sigma$  are defined without explicit reference to the tetrahedra containing them.*

**Proof** Suppose elements  $K_1$  and  $K_2$  meet at a face  $f$  and that one edge of  $f$  lies on  $\Sigma$ . If the end points of this edge are  $a_1$  and  $a_2$ , then the edge is given by  $P_{\Sigma}(ta_1 + (1 - t)a_2)$  which does not depend on the choice of element.  $\square$

This is all of Dubois' theory that we need. But for completeness we outline his construction of edge elements on the curvilinear grid and give the main results of his theory. Recall from Section 3.9 that  $dF_K$  denotes the Jacobian matrix for  $F_K$ . Then using the definition of  $D_1$  in (5.17) and  $R_1$  in (5.32), the appropriate local subspaces for an element  $K \in \tau_b$  are obtained using the mappings (3.77) and (3.76) to obtain(8.13)

$$D_1(K) = \left\{ w \mid w \circ F_K = \frac{1}{\det(dF_K)} dF_K \hat{w} \text{ for some } \hat{w} \in D_1 \right\},$$

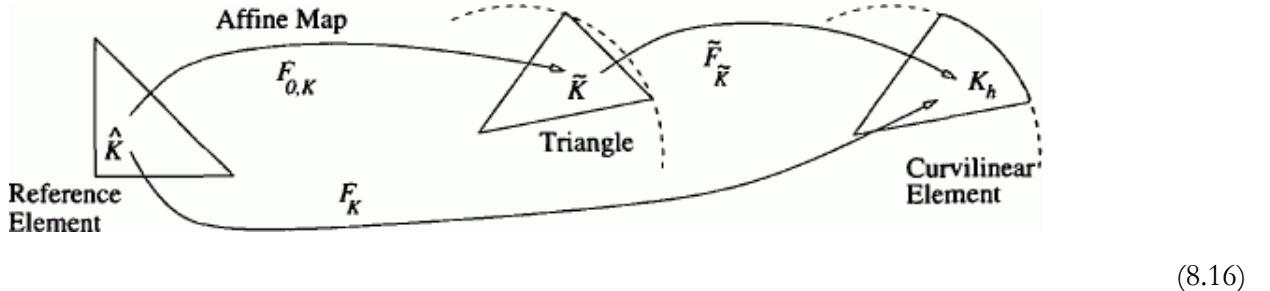
(8.14)

$$R_1(K) = \left\{ v \mid v \circ F_K = (dF_K)^{-T} \hat{v} \text{ for some } \hat{v} \in R_1 \right\} .$$

Note that it is vital to use the appropriate transformations here. With these subspaces in hand, the obvious generalization of the first kind spaces  $W_b$  and  $V_b$  defined in (5.28) and (5.40) to curvilinear domains is(8.15)

$$V_h = \{ v_h \in H(\text{curl}; \Omega) \mid v_h|_K \in R_1(K) \text{ for all } K \in \tau_h \},$$

Fig. 8.2. A summary of the various maps used in this section. For simplicity, we show the elements as triangles. The map  $F_{0,K}$  maps the reference triangle to a standard triangle  $\tilde{K}$  with two vertices on the curvilinear boundary (the boundary  $\Sigma$  is shown as a dashed line). The map  $\tilde{F}_K$  maps the triangle  $\tilde{K}$  to the curvilinear triangle  $K$ . The map  $F_K$  maps  $K$  to  $K_h$  directly.



$$W_h = \{ w_h \in H(\text{div}, \Omega) \mid v_h|_K \in D_1(K) \text{ for all } K \in \mathcal{T}_h \} .$$

In addition, we have the usual mapped scalar space(8.17)

$$U_h = \left\{ p_h \in H^1(\Omega) \mid p_h|_K \circ F_K = \hat{p} \text{ for some } \hat{v} \in P_1 \right\} .$$

Because we have used the appropriate transformations, the results of Section 3.9 show that the discrete de Rham diagram (5.59) still holds.

Now we need to define the interpolant in each case. For  $U_b \subset H^1(\Omega)$  we simply use standard point values at the vertices in the mesh. For  $V_b$  and  $W_b$ , we use the obvious degrees of freedom so that the degrees of freedom for  $V_b$  are

$$\Sigma_V = \left\{ \int_e v \cdot \tau ds \text{ for each edge } e \text{ of the mesh} \right\},$$

where  $\tau$  is the unit tangent to  $e$ , and the degrees of freedom for  $W_b$  are

$$\Sigma_W = \left\{ \int_f w \cdot v dA \text{ for each face } f \text{ of the mesh} \right\},$$

where  $v$  is a normal to  $f$ . Because of the use of the mappings to define  $D_1(K)$  and  $R_1(K)$ , and taking into account that  $v$  and  $\tau$  are related to  $\hat{v}$  and  $\hat{\tau}$  by (3.79) and (3.80), we see that these degrees of freedom are mapped (up to sign) to the corresponding degrees of freedom on the reference element  $\hat{K}$ . Thus, unisolvence and conformance follow using the same arguments as used in Sections 3.9, 5.4 and 5.5 (using eqns (3.81) and (3.82)). Furthermore, the same arguments show the commuting properties of the interpolants relative to the above degrees of freedom. Hence, we see that the discrete de Rham diagram (5.59) holds with the obvious definition for  $Z_b$  using the mapping (3.78).

To analyze the approximation properties of these spaces we need to analyze how the Jacobian  $dF_K$  and Hessian  $(\partial/\partial \mathcal{O}_l)dF_K(\mathcal{O})$ ,  $1 \leq l \leq 3$ , depend on  $b$ . This can be done by first defining  $F_{0,K}$  to be the affine map given by  $F_{0,K}(\mathcal{O}) = B_{K\mathcal{O}} +$

$b_K$  such that  $F_{0,K}$  maps the vertices of  $K$  to those of  $K$ . Then  $F_{0,K}(K) = K$  is a true tetrahedron and  $|B_K|$  and  $\det(B_K)$  can be estimated using Lemma 5.10. We define  $\tilde{F}_K = F_K \circ F_{0,K}^{-1}$ , so  $F_K = \tilde{F}_K \circ F_{0,K}$ . See Fig. 8.2 for a graphical summary of this notation.

For a true tetrahedron  $K$  the map  $\tilde{F}_K = I$ . For a curvilinear tetrahedron, if the mesh is regular and quasi-uniform (i.e.  $b_K \leq \sigma Q_K$  and  $b \leq \gamma b_K$  for fixed  $\sigma$  and  $\gamma$  and all  $K \in \tau_p$ ,  $b > 0$ ), Dubois proves the following lemma.

**Lemma 8.18** *The following estimates hold with constants independent of  $b$ : if  $d\tilde{F}_{K,i,j}$  is the  $(i,j)$  entry of  $d\tilde{F}_K$  then*

$$\begin{aligned} \sup_{\tilde{x} \in \tilde{K}} |(d\tilde{F}_{K,i,j})_{i,j} - \delta_{i,j}| &\leq Ch, \quad 1 \leq i, j \leq 3, \\ \sup_{\tilde{x} \in \tilde{K}} \left| \frac{\partial}{\partial \tilde{x}_l} (d\tilde{F}_{K,i,j})_{i,j} \right| &\leq C, \quad 1 \leq i, j, l \leq 3, \end{aligned}$$

This lemma allows the use of  $|B_K|$  in place of  $|dF_K|$  in the scaling estimates and allows Dubois to prove the following optimal approximation theorem (most likely the results of Theorems 5.41 and 5.25 also hold but this has not been verified).

**Theorem 8.19** *Assume that the mesh is regular and quasi-uniform. Then for all  $b$  small enough*

$$\begin{aligned} \|u - r_h u\|_{(L^2(\Omega))^3} + \|\nabla \times (u - r_h u)\|_{(L^2(\Omega))^3} \\ \leq Ch(\|u\|_{(H^1(\Omega))^3} + \|\nabla \times u\|_{(H^1(\Omega))^3}), \\ \|u - w_h u\|_{(L^2(\Omega))^3} \leq Ch\|u\|_{(H^1(\Omega))^3}, \end{aligned}$$

where  $r_b$  and  $w_b$  are the interpolation operators for the curvilinear spaces  $V_b$  and  $W_b$ , respectively.

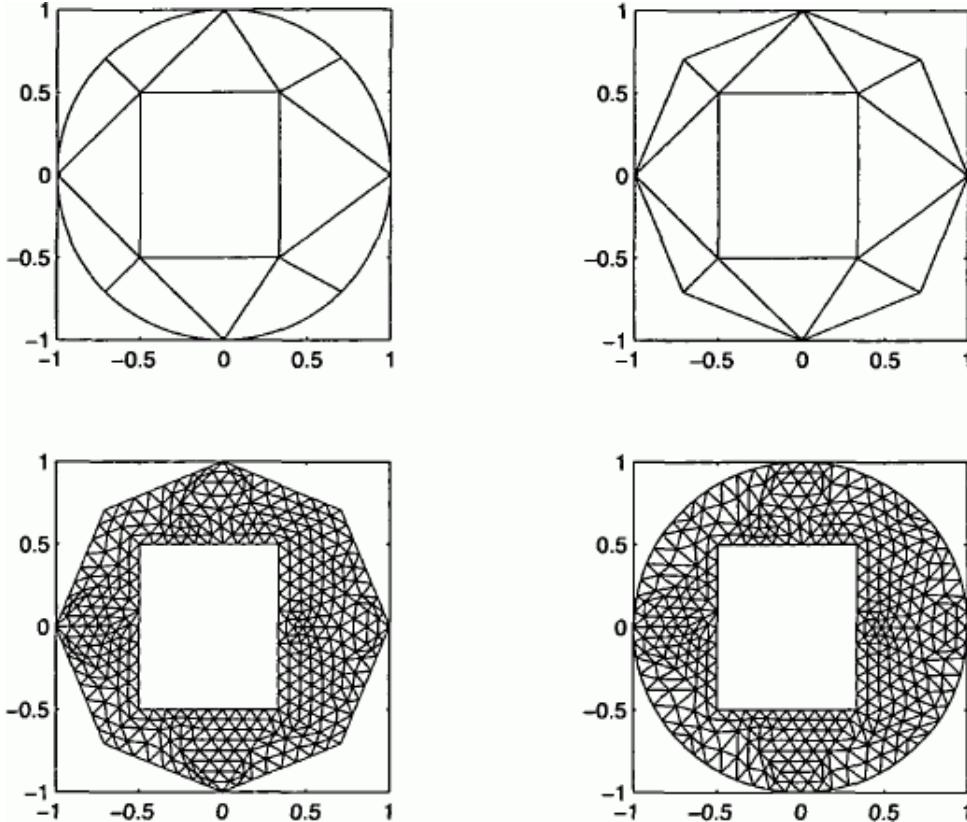
The problem with this approach is that only the lowest-order edge elements are covered by the theory. Since the mapping  $F_K$  depends on  $b$ , it is likely that the mapping procedure we have outlined would pollute error estimates for higher-order elements (see Ciarlet [80] for isoparametric element error estimates in the scalar case up to cubic elements).

Note also that this analysis is restricted to tetrahedra. For hexahedra, the corresponding mapping procedure (even if  $F_K$  is restricted to be trilinear) can destroy the optimal convergence properties of the element in  $H(\text{curl}; \Omega)$  (see [17] for the two-dimensional case). This problem occurs because local mappings  $F_K$ , dependent on  $b$ , are used. Since we wish to approximate Maxwell's equations in the interior of a sphere, we can use a mapping independent of  $b$  and hence maintain accuracy. This is done in the next section.

### 8.3.2 Large-element fitting of domains

We now discuss a method for exactly meshing a smooth curvilinear boundary which preserves the approximation properties of any finite element method. Suppose  $\Sigma$  is a smooth  $C^2$  curvilinear boundary (we have in mind the sphere) and

Fig. 8.3. Figure showing the steps in creating a curvilinear grid. First, the domain is covered by large curvilinear elements (*top left*) obtained by mapping standard elements on a crude approximation of the domain (*top right*). Then the crude approximate polygonal is meshed (*bottom left*) and the resulting grid mapped back to the curvilinear domain (*bottom right*).



$\Gamma$  is a polyhedral boundary for the domain  $\Omega$ . We first construct a curvilinear tetrahedral grid for  $\Omega$  of elements of diameter at most  $H$ . This can be done by Dubois' method from the previous section, provided  $P_\Sigma$  is known (see (8.10) for the case of a sphere).

Each curvilinear tetrahedron  $K \in \tau_H$  is the image of a true tetrahedron  $\tilde{K}$  having the same vertices via  $F_K$ . The element  $\tilde{K}$  is in turn obtained from the reference element by a standard affine map denoted by  $F_{0,K}$  as in the previous section (see Fig. 8.2). Let  $\tilde{\tau}_H$  denote the mesh of tetrahedra  $\tilde{K}$  covering the domain  $\Omega_H = \cup_{\tilde{K} \in \tilde{\tau}_H} \tilde{K}$ . Thus,  $\Omega_H$  is a crude tetrahedral mesh of  $\Omega$  that can be mapped by a piecewise smooth mapping element by element to cover  $\Omega$  with “large” elements. We now cover the polyhedral domain  $\Omega_H$  by regular elements on a finer mesh  $\tilde{\tau}_b$  such that each element  $\tilde{K}_b \in \tilde{\tau}_b$  is entirely contained in some  $\tilde{K} \in \tilde{\tau}_H$ . Mapping  $\tilde{K}$  to  $K$  transforms each  $\tilde{K}_b \subset \tilde{K}$  to a curvilinear element  $K_b$  on  $\Omega$  and the union of these elements gives a curvilinear mesh of  $\Omega$  in which

curvilinear elements are used even away from the boundary.

Fig. 8.3 shows the idea in the two-dimensional context. In this example, the annular domain is first meshed by a few curvilinear triangles as shown in the top left panel. These fit the boundary  $\Sigma$  exactly, thanks to the use of  $P_\Sigma$  and the two-dimensional version of the construction of  $\tilde{F}_K$  given in (8.11) and (8.12). The curvilinear triangles in  $\tilde{\tau}_h$  are the image under  $\tilde{F}_H$  of the true triangles in  $\tilde{\tau}_b$  forming  $\Omega_h$  and shown in the right-hand panel. The domain  $\Omega_h$  is then meshed as usual using a regular triangulation shown in the bottom left panel. Finally, the maps  $\tilde{F}_h$  are used to map back to  $\Omega$  and obtain a curvilinear triangulation of  $\Omega$  as shown in the bottom right panel. We may also form a quasi-uniform mesh of  $\Sigma$  by using a quasi-uniform mesh on the boundary of  $\Omega_h$  that is mapped to  $\Sigma$  (see Fig. 8.2).

For notational convenience, we can define  $\tilde{F}_h : \Omega_h \rightarrow \Omega$  to be the continuous piecewise smooth map such that if  $\tilde{x} \in K \in \tilde{\tau}_h$ , then  $\tilde{F}_h(\tilde{x}) = \tilde{F}_K(\tilde{x})$ . Then the curvilinear mesh  $\tilde{\tau}_h$  is defined by

$$\tilde{\tau}_h = \left\{ K_h \mid K = \tilde{F}_H(\tilde{K}_h) \text{ for some } \tilde{K}_h \in \tilde{\tau}_b \right\}.$$

Recalling that  $\Omega_h$  is a polyhedral domain, we may then define standard edge finite element spaces  $\tilde{U}_b$  (vertex elements in  $H^1(\Omega_h)$ ),  $\tilde{V}_b$  (edge finite elements in  $H(\text{curl}; \Omega_h)$ ) and  $\tilde{W}_b$  (face finite elements in  $H(\text{div}; \Omega_h)$ ) in the usual way as in Chapter 5 using the fine grid  $\tilde{\tau}_b$ . The corresponding space on  $\Omega$  is obtained by the usual mappings as follows:

$$\begin{aligned} U_h &= \left\{ p_h \in H^1(\Omega) \mid p_h \circ \tilde{F}_H = \bar{p}_h \text{ for some } \bar{p}_h \in \tilde{U}_b \right\}, \\ V_h &= \left\{ v_h \in H(\text{curl}; \Omega) \mid v_h \circ \tilde{F}_H = (d\tilde{F}_H) \text{ for some } \bar{v}_h \in \tilde{V}_b \right\}, \\ w_h &= \left\{ w_h \in H(\text{div}; \Omega) \mid w_h \circ \tilde{F}_H = \frac{1}{\det(d\tilde{F}_H)} d\tilde{F}_H \tilde{w}_h \text{ for some } \tilde{w}_h \in \tilde{W}_b \right\}. \end{aligned}$$

Because we used the continuous piecewise smooth map  $\tilde{F}_h$  in the way discussed in Section 3.9, we automatically have  $\nabla U_b \subset V_b$  and  $\nabla \times V_b \subset W_b$ . The appropriate continuity conditions are also satisfied as in the case of a simple affine map analyzed in Chapter 5.

To obtain error estimates we note that if  $v \in H(\text{curl}; \Omega)$  there is a function  $\tilde{v} \in H(\text{curl}; \Omega_h)$  such that  $v \circ \tilde{F}_h = (d\tilde{F}_h)^{-1} \tilde{v}$  and since  $\tilde{F}_h$  is piecewise smooth,  $\tilde{v}$  inherits the smoothness of  $v$  element by element. Thus, using the invariance of the degrees of freedom for  $V_b$  (see Lemma 5.34 for the affine case), we have  $r_b v = (d\tilde{F}_h)^{-1} \tilde{r}_b \tilde{v}$  provided  $v$  is smooth enough that  $\tilde{r}_b$  is well defined. This implies that

$$\begin{aligned} \|v - r_h v\|_{(L^2(\Omega))^3} &= \|\det(d\tilde{F}_H)^{\frac{1}{2}} (d\tilde{F}_H^{-\top}) (\tilde{v} - \tilde{r}_h \tilde{v})\|_{(L^2(\Omega_H))^3} \\ &\leq C \|\tilde{v} - \tilde{r}_h \tilde{v}\|_{(L^2(\Omega_H))^3}. \end{aligned}$$

The inequality holds since  $d\tilde{F}_H$  is independent of  $b$ . A similar result holds for the curl of  $v$ . Thus  $V_b$  (and in a corresponding way  $U_b$  and  $W_b$ ) on  $\Omega$  inherit all

the approximation properties from the corresponding space  $V_b$  ( $\tilde{U}_b$  or  $\tilde{W}_b$ ) on  $\Omega_H$ . We may summarize this as follows.

**Theorem 8.20** *Theorems 5.25, 5.41 and 5.48 hold for the spaces  $U_b$ ,  $V_b$  and  $W_b$  on the curvilinear grid  $\tau_b$  of  $\Omega$ .*

Of course, this construction requires us to evaluate various integrals of curvilinear finite element functions on  $\Omega$ . This can be done by mapping back to  $\Omega_H$ . So for example if  $u_b, v_b \in V_b$ , then

$$\begin{aligned} (\in_r u_h, v_h) &= \int_{\Omega} \in_r u_h \cdot \bar{v}_h dV \\ &= \int_{\Omega_H} \left[ (d\tilde{F}_H)^{-1} \in_r (d\tilde{F}_H)^{-T} \tilde{u}_h \right] \cdot \bar{v}_h \left| \det(d\tilde{F}_H) \right| d\tilde{V} \\ &= \int_{\Omega_H} \widetilde{\in}_r \tilde{u}_h \cdot \bar{v}_h d\tilde{V} \end{aligned}$$

where  $\widetilde{\in}_r = |\det(d\tilde{F}_H)| d\tilde{F}_H^{-1} (\in_r \circ \tilde{F}_H) d\tilde{F}_H^{-T}$  is defined element by element, and on each element inherits the smoothness of  $\in_r$ . In general, these integrals must be done numerically. Unfortunately, the analysis of integration rule accuracy needed to maintain the order of the approximation in the numerical scheme (i.e.  $O(h)$  when  $k = 1$ , etc.) has yet to be performed for edge elements. We presume that an integration rule that computes the integrals of polynomials of degree  $2k - 1$  exactly on each element is sufficient (see Ciarlet [80]). This is the second variational crime (the first being the isoparametric approximation of  $\Omega$ ) that needs to be investigated for the Maxwell system.

Let us remark that this technique is probably not of practical interest in general but does cover the case of spherical  $\Sigma$  considered in later chapters. In addition, we could also use hexahedral elements if  $\tau_H$  consists of curvilinear hexahedra and these can be mapped onto right hexahedra with edges parallel to the coordinate axis whose union forms  $\Omega_H$  (a limitation for sure!).

## 8.4hp finite elements

In the *hp* version of the finite element method [275], a finite element grid  $\tau_H$  is used, and in addition the degree of the polynomial approximation on each element is varied to produce a good approximation of the solution. Typically, where the solution is smooth (e.g. in Maxwell's equations away from boundaries, interfaces or regions of non-smooth  $\epsilon_r$  or  $\mu_r$ ), large elements can be used and high-degree piecewise polynomial basis functions ensure accuracy. In particular, analytic functions can be well approximated by high-order polynomials. In regions where the solution is less smooth (e.g. in the neighborhood of a re-entrant corner or edge of the domain), the mesh needs to be refined and the polynomial degree decreased (see [238] for an analysis of mesh refinement towards an edge singularity). The programming of an *hp* scheme is rather complex (see Demkowicz [255, 257]), but is rendered tractable by an appropriate definition of

the degrees of freedom for  $hp$  finite elements. These have to be chosen so that it is convenient to allow the degree of the piecewise polynomial to change from element to element. Here we give a construction of  $hp$  finite elements from [120], for another construction see [149, 7, 6]. Unfortunately, while we can analyze the interpolation error in these elements for fixed polynomial degree as the mesh is refined ( $h \rightarrow 0$ ), there is no analysis yet that provides error estimates like part (b) of Theorem 5.41 that includes the influence of the polynomial degree, and hence there is, as yet, no  $hp$  error analysis available for the time-harmonic Maxwell's equations. We start with the simplest case of elements in  $H^1(\Omega)$  and proceed to  $H(\text{curl};\Omega)$  and  $H(\text{div};\Omega)$ . As usual, we consider a Lipschitz polyhedral domain  $\Omega$  covered by a regular tetrahedral mesh  $\tau_h$ ,  $h > 0$ , where  $h$  is the maximum diameter of the elements in the mesh. Furthermore, we recall that  $K$  denotes the standard reference element.

### 8.4.1 $H^1(\Omega)$ conforming $hp$ element

This element is quite standard being essentially the element defined in Section 5.6 but modified to allow  $p$  to vary from element to element. The basis functions in the space  $P_K$  are defined by 11 integer parameters for each element  $K$ :  $k_K$ ,  $k_f$  for each face  $f$ , and  $k_e$  for each edge  $e$ , as follows:(8.18)

$$P_K = \left\{ q \in P_{k_K}(K) \mid q|_f \in P_{k_f}(f) \text{ for each face } f \text{ and} \right. \\ \left. q|_e \in P_{k_e}(e) \text{ for each edge } e \right\} .$$

We can now define the parameter  $p$  in the name “ $hp$  method”:  $p = \min_{K \in \tau_h} k_K$ . We choose the polynomial degree indices to satisfy  $k_f \leq k_K$  and  $k_e \leq k_K$ . We also require  $k_e \leq k_f$  for every edge  $e$  of a face  $f$ . Demkowicz *et al.* [255, 257] use the following *minimum rule* to assign the appropriate indices. First, the elemental polynomial degree  $k_K$  is assigned for each  $K \in \tau_h$ , then if  $f$  is a face of two tetrahedrons  $K_1$  and  $K_2$ , the index  $k_f$  is assigned to be the minimum of  $k_{K_1}$  and  $k_{K_2}$  (for boundary faces  $k_f = k_K$ ). The edge degrees  $k_e$  are then assigned to be the minimum of the face degrees  $k_f$  for faces adjacent to this edge.

Now we can define the  $hp$  degrees of freedom for  $q \in H^1(K)$ ,  $l > \frac{3}{2}$ , on a tetrahedral element  $K$  by using the following four sets of degrees of freedom corresponding to those in Definition 5.46:

- Vertex degrees: Let  $a_i$ ,  $1 \leq i \leq 4$ , be the vertices of  $K$ . Then

$$M_v(q) = \{q(a_i) \mid 1 \leq i \leq 4\} .$$

- Edge degrees: Let  $s$  denote arc length along  $e$ . Then

$$M_e(q) = \left\{ \int_e \frac{\partial q}{\partial s} \frac{\partial \varphi}{\partial s} ds \mid \text{for all } \varphi \in P_{k_e}(e) \text{ such that } \varphi \text{ vanishes at} \right. \\ \left. \text{the endpoints of } e \text{ for all edges } e \right\} .$$

- Face degrees:

$$M_f(q) = \left\{ \int_f \nabla_f q \cdot \nabla_f \varphi dA \text{ for all } \varphi \in P_{k_f}(f) \text{ such that} \right. \\ \left. \varphi = 0 \text{ on } \partial f \text{ for all face } f \right\} .$$

- Volume degrees:

$$M_K(q) = \left\{ \int_K \nabla q \cdot \nabla \varphi dV \text{ for all } \varphi \in P_{k_K}(K) \text{ such that } \varphi = 0 \text{ on } \partial K \right\} .$$

Then the element degrees of freedom are

$$\Sigma_K = M_v(q) \cup M_e(q) \cup M_f(q) \cup M_K(q) .$$

Note that  $M_k(q) = \emptyset$  if  $k_k < 4$ ,  $M_f(q) = \emptyset$  if  $k_f < 3$ , and  $M_e(q) = \emptyset$  if  $k_e < 2$ .

Using these degrees of freedom, we can prove unisolvence and  $H^1$  conformance as in the proof of Lemma 5.47. Suppose all degrees of freedom of a discrete function  $q_b \in P_K$  associated with vertices and edges of a face  $f$  vanish. The degrees  $M_v(q_b) = \{0\}$  imply that we can select  $\varphi = q_b$  in the degrees of freedom  $M_f(q_b)$  so that  $\int_e (\partial q_b / \partial s)^2 ds = 0$  and thus  $q_b = 0$  on each edge of  $f$ . But then we may choose  $\varphi = q_b$  in the degrees for  $M_e(q_b)$ , so  $\int_f |\nabla_{fqb}|^2 dA = 0$ . Using the fact that  $q_b = 0$  on  $\partial f$ , we have  $q_b = 0$  on  $f$  as required for unisolvence.

Using these degrees of freedom we may define an  $hp$  interpolant  $\pi_{h,p,K} : H(K) \rightarrow P_K$ , as in Chapter 5 and hence a global interpolant  $\pi_{h,p}$  via (5.56). The error estimates for this interpolant are technically challenging. It is possible to prove that if  $q \in H(\Omega)$ ,  $s \geq 2$ , and if a uniform degree  $p$  is used on all elements (i.e.  $p = k_k = k_f = k_e$  for all  $K, f$  and  $e$ ), then

$$\|q - \pi_{h,p}q\|_{H^1(\Omega)} \leq Ch^{\min(p+1,s)-1} p^{1-s} \|q\|_{H^s(\Omega)},$$

as  $h \rightarrow 0$  and  $p \rightarrow \infty$  [227]. More complex estimates near singularities and for variable degree are also known [270]. The set of all finite element functions of the type outlined in this section on a mesh  $\tau_h$  with maximum  $h$  and minimum polynomial degree is denoted by  $U_{h,p}$ .

## 8.4.2 hp curl conforming elements

Now we need to define a finite element subspace  $V_{h,p}$  of  $H(\text{curl}; \Omega)$  such that  $\nabla U_{h,p} \subset V_{h,p}$  and such that the appropriate part of the discrete de Rham diagram commutes. Recall that  $u_T$  denotes the tangential component of  $u$  on an appropriate surface. We define the space of basis functions  $P_K$  on an element  $K$  by

$$P_V = \left\{ u \in (P_{k_K-1}(K))^3 \mid u_T \in (P_{k_f-1}(f))^3 \text{ for each face } f \text{ and} \right. \\ \left. u \cdot t \in P_{k_e-1}(e) \text{ for each edge } e \text{ with unit tangent } t \right\} .$$

Here, as before,  $k_f \leq k_k$  and  $k_e \leq k_f$  for all edges  $e$  of  $f$  and each  $f$ . Of course, we use here the same indices  $k_k$ ,  $k_f$  and  $k_e$  as were used in (8.18).

Note that if  $k_e = k + 1$  and  $k_f = k_e = k + 1$ , then  $P_V = (P_e)^3$ , and we are thus dealing with a generalization of the second family of edge elements defined in Section 8.2. In general,

$$V_{h,p-1} = \{u_h \in H(\text{curl}; \Omega) | u_h|_K \in P_V \text{ for all } K \in \mathcal{T}_h\}.$$

Now we can define degrees of freedom for this space. As in Section 5.5 this involves edge, face and volume degrees in general. Let  $u \in (H^{1/2+\delta}(K))^3$ ,  $\delta > 0$ , and  $\nabla \times u \in (L^q(K))^3$ ,  $q > 2$ . Then the following degrees of freedom are well-defined.

- Edge degrees (recall that  $\tau$  is the unit tangent to edge  $e$ ):

$$M_e(u) = \left\{ \int_e u \cdot \tau \varphi ds \text{ for all } \varphi \in P_{k_e-1} \text{ for each edge } e \right\}.$$

- Face degrees of freedom:

$$\begin{aligned} M_f(u) = & \left\{ \int_f \nabla_f \times (u|_T) \nabla_f \times \varphi dA \text{ for all } \varphi \in (P_{k_f-1}(f))^3 \right. \\ & \text{with } \varphi \cdot \tau = 0 \text{ on } \partial f \text{ and } \int_f u_T \cdot \nabla_f \xi dA \text{ for all } \xi \in P_{k_f}(f) \\ & \left. \text{such that } \xi = 0 \text{ on } \partial K \right\}. \end{aligned}$$

- Volume degrees of freedom:

$$\begin{aligned} M_K(u) = & \left\{ \int_K \nabla \times u \cdot \nabla \times \varphi dV \text{ for all } \varphi \in (P_{k_K-1})^3 \text{ such that} \right. \\ & \varphi_T = 0 \text{ on } \partial K, \text{ and } \int_K u \cdot \nabla \xi dV \text{ for all } \xi \in P_{k_K} \text{ such that} \\ & \left. \xi = 0 \text{ on } \partial K \right\}. \end{aligned}$$

Then, as usual, the total degrees of freedom on  $K$  are

$$\Sigma_K = M_e(u) \cup M_f(u) \cup M_K(u).$$

Here  $M_e(u) = \emptyset$  if  $k_e < 4$  and  $M_f(u) = \emptyset$  if  $k_f < 3$ .

The degrees of freedom  $M_f(u)$  and  $M_K(u)$  look unusual and perhaps might appear overdetermined. This is not the case. When we compute the local edge interpolant  $r_{h,p-1}$  of  $u$ , we need (e.g. using the degrees of freedom in  $M_K(u)$ ) that

$$\int_K \nabla \times (u - r_{h,p-1}u) \cdot \nabla \times \varphi dV = 0 \text{ for all } \varphi \in (P_{k_K-1})^3 \cap H_0(\text{curl}; K),$$

$$\int_K (u - r_{h,p-1}u) \cdot \nabla \xi dV = 0 \text{ for all } \xi \in P_{k_K} \cap H_0^1(K).$$

If we introduce a Lagrange multiplier, we see that this is equivalent to finding  $r_{h,p-1}u \in P_K$  and  $\lambda \in P_{k_K} \cap H_0^1(K)$  such that

$$\int_K \nabla \times (u - v_{h,p-1}u) \cdot \nabla \times \varphi dV + \int_K \varphi \cdot \nabla \lambda dV = 0$$

for all  $\varphi \in (P_{k_K-1})^3 \cap H^0(\operatorname{curl}; K)$ ,

$$\int_K (u - r_{h,p-1}u) \cdot \nabla \xi dV = 0 \quad \text{for all } \xi \in P_{k_K} \cap H_0^1(K).$$

Choosing  $\varphi = \nabla \lambda$  shows that  $\lambda = 0$  so the above two problems are equivalent and the stated degrees of freedom do not overdetermine  $r_{h,p-1}u$ . Again unisolvence and conformance are easy to prove using arguments like those verifying Lemmas 5.36 and 5.37. However, nothing is proved about the approximation properties of this operator when  $p$  varies. Computational evidence [256] suggests that optimal convergence rates are obtained.

### 8.4.3 hp divergence conforming space

The next space needed to complete the de Rham diagram is much simpler than the previous spaces. Let

$$W_{h,p-2} = \{ \omega_h \in H(\operatorname{div}; \Omega) \mid \omega_h|_K \in (P_{k_K-2}(K))^3 \text{ and}$$

$$v_f \cdot \omega_h|_f \in (P_{k_f-2}(f))^2 \text{ for each face}$$

$$f \text{ in the mesh where } v_f \text{ is a unit normal to } f \} .$$

As in the case of the curl conforming space, the indices  $k_K$  and  $k_f$  are those for the  $H^1(\Omega)$  conforming space  $U_{h,p}$  and  $k_f \leq k_K$ . By using this choice, it is clear that  $\nabla \times V_{h,p-1} \subset W_{h,p-2}$ . To complete the definition of this space, we need to define the degrees of freedom. These are analogous to those for the second family face space  $W_h^{(2)}$  in Section 8.2. In particular, we have, for  $u \in (H^{1/2+\delta}(K))^3$ ,  $\delta > 0$ , the following degrees:

- Face degrees of freedom:

$$M_f(u) = \left\{ \int_f v_f \cdot u \varphi dA \text{ for all } \varphi \in P_{k_f-2}(f) \text{ and each face } f \text{ of } K \right\} .$$

- Volume degrees of freedom:

$$M_K(u) = \left\{ \int_K \nabla \cdot u \nabla \cdot \varphi dV \text{ for all } \varphi \in (P_{k_f-2}(K))^3 \text{ such that}$$

$$v_f \cdot \varphi = 0 \text{ on all faces of } K \text{ and}$$

$$\int_K u \cdot \nabla \times \xi dV \text{ for all } \xi \in P_{k_K-1}(K) \text{ such that}$$

$$v_f \times \xi = 0 \text{ on each face of } f \} .$$

The element degrees of freedom are then  $\sum_K = M(\mathcal{U}) \cup M_K(\mathcal{U})$ . As in the case of the curl conforming elements, the degrees of freedom  $M_K(\mathcal{U})$  are unusual and might appear to overdetermine a function  $u_b \in W_{h,p,2}$ . However, using the appropriate mixed problem, we see that this is not the case. The interpolant  $w_{h,p,2} u$  can now be defined using these degrees of freedom in the usual way. This definition only makes sense for  $k_K > 2$ . Note that  $M_K(\mathcal{U}) = \emptyset$ , if  $k_K < 3$ .

For completeness, we shall also define

$$Z_{h,p-3} = \left\{ z_h \in L^2(\Omega) \mid z_h|_K \in P_{k_K-3}(K) \text{ for all } K \in \tau_h \right\}$$

which is well defined for  $k_K \geq 3$  and denote by  $P_{0,h,p-3}$  the  $L^2(\Omega)$  orthogonal projection into  $Z_{h,p-3}$ .

#### 8.4.4 de Rham diagram for hp elements

Suppose  $k_K \geq 3$  for all  $K \in \tau_b$  (it is possible using first-type edge elements to extend this construction below  $k_K = 3$ ). Then we have the following theorem.

**Theorem 8.21** *Provided  $k_K \geq 3$  for all  $K \in \tau_b$  (i.e.  $p \geq 3$ ), the following discrete de Rham diagram commutes:* (8.19)

$$\begin{array}{ccccccc} H^1(\Omega) & \xrightarrow{\nabla} & H(\text{curl}; \Omega) & \xrightarrow{\nabla \times} & H(\text{div}; \Omega) & \xrightarrow{\nabla \cdot} & L^2(\Omega) \\ \cup & & \cup & & \cup & & \\ U & & V & & W & & \\ \pi_{h,p} \downarrow & & r_{h,p-1} \downarrow & & w_{h,p-2} \downarrow & & P_{0,h,p-3} \downarrow \\ U_{h,p} & \xrightarrow{\nabla} & V_{h,p-1} & \xrightarrow{\nabla \times} & W_{h,p-2} & \xrightarrow{\nabla \cdot} & Z_{h,p-3} \end{array}$$

where  $U$ ,  $V$  and  $W$  are subspaces on which the appropriate interpolation operators are defined.

**Proof** We have already seen that the bottom row of the diagram is satisfied by the design of the spaces. The remainder of the proof mirrors that of Theorems 5.49 and 5.40 (see [120] for details). For example to prove that

$$\nabla \pi_{h,p} q = r_{h,p-1} \nabla q$$

for all  $q \in H^{3/2+\delta}(\Omega)$ ,  $\delta > 0$ , we proceed to check that the degrees of freedom  $r_{h,p-1} \nabla q$  agree with those of  $\nabla \pi_{h,p} q$  element by element. We start with edge degrees. Let  $e$  be an edge of  $K \in \tau_b$  with unit tangent  $\tau$ . Then using the edge degrees for  $r_{h,p-1} \nabla q$ , we have that for all  $\varphi \in P_{k_e-1}$  (8.20)

$$\int_e (\nabla \pi_{h,p} q - r_{h,p-1} \nabla q) \cdot \tau \varphi ds = \int_e \frac{\partial}{\partial s} (\pi_{h,p} q - q) \varphi ds .$$

Choosing  $\varphi = 1$  we have, if  $a$  and  $b$  are the end points of  $e$ ,

$$\int_e \frac{\partial}{\partial s} (\pi_{h,p} q - q) ds = (\pi_{h,p} q(b) - q(b)) - (\pi_{h,p} q(a) - q(a)) = 0$$

where we have used the vertex degrees of freedom for  $\pi_{h,p} q$ .

Now we write  $\varphi = \varphi_0 + \varphi_1$  where  $\varphi_0 \in P_0$  and  $\int_e \varphi_0 ds = \int_e \varphi_1 ds$ . We have  $\varphi_1 = \partial \xi / \partial s$  for some  $\xi \in P_{k_e}$  and we may choose  $\xi(a) = 0$ . Then the fact that  $\int_e \varphi_1 ds = 0$  shows that  $\xi(b) = 0$ . Using  $\varphi = \partial \xi / \partial s$  in (8.20) and the edge degrees of freedom for  $\pi_{h,p} q$  shows that the edge degrees of freedom for  $r_{h,p-1\nabla} q$  and  $\nabla \pi_{h,p} q$  agree.

Now we proceed to the face degrees of freedom for  $r_{h,p-1\nabla} q$ . Using the degrees of freedom for  $r_{h,p-1\nabla} q$ , for  $\varphi \in (P_{k_f-1}(f))^2$  with vanishing tangential component on  $\partial f$ , we have

$$\begin{aligned} & \int_f \nabla_f \times (\nabla \pi_{h,p} q - r_{h,p-1} \nabla q) \cdot \nabla_f \times \varphi dA \\ &= \int_f \nabla_f \times (\nabla \pi_{h,p} q - \nabla q) \cdot \nabla_f \times \varphi dA = 0. \end{aligned}$$

Next, we use the face degrees of freedom for  $r_{h,p-1\nabla} q$  and the face degrees of freedom of  $\pi_{h,p} q$  to show that for  $\xi \in P_{k_f}$  with  $\xi = 0$ , we have

$$\begin{aligned} & \int_e (\nabla \pi_{h,p} q - r_{h,p-1} \nabla q) \cdot \nabla_f \xi dA \\ &= \int_e (\nabla \pi_{h,p} q - \nabla q) \cdot \nabla_f \xi dA \\ &= \int_e (\nabla_f \pi_{h,p} q - \nabla_f q) \cdot \nabla_f \xi dA = 0. \end{aligned}$$

Finally, we need to check the volume degrees of freedom. At this stage we know, from the curl conforming condition for  $V_{h,p-1}$ , that (8.21)

$$(\nabla \pi_{h,p} q - r_{h,p-1} \nabla q) \times v = 0 \text{ on } \partial K,$$

where  $v$  is the unit outward normal to  $K$ . In the same way as for the face degrees, using the degrees of freedom for  $r_{h,p-1}$ , for all  $\varphi \in (P_{k_K}(K))^3 \cap H_0(\text{curl}, K)$  we have

$$\begin{aligned} & \int_K \nabla \times (\nabla \pi_{h,p} q - r_{h,p-1} \nabla q) \cdot \nabla \times \varphi dV \\ &= \int_K \nabla \times (\nabla \pi_{h,p} q - \nabla q) \cdot \nabla \times \varphi dV = 0. \end{aligned}$$

Indeed, choosing  $\varphi = \nabla \pi_{h,p} q - r_{h,p-1} \nabla q$  we see that  $\nabla \times (\nabla \pi_{h,p} q - r_{h,p-1} \nabla q) = 0$  in  $K$ , so there is a function  $\xi \in P_{k_K}(K)$  with  $\xi = 0$  on  $\partial K$  such that  $\nabla \xi =$

$\nabla \pi_{h,p} q - r_{h,p-1} \nabla q$  (the boundary condition on  $\xi$  is necessary for (8.21) to hold). Hence, using the volume degrees for  $r_{h,p-1}$  and  $\pi_{h,p}$ , we have

$$\begin{aligned} \|\nabla \pi_{h,p} q - r_{h,p-1} \nabla q\|_{(L^2(K))^3}^2 &= \int_K (\nabla \pi_{h,p} q - r_{h,p-1} \nabla q) \cdot \nabla \xi dV \\ &= \int_K (\nabla \pi_{h,p} q - \nabla q) \cdot \nabla \xi dV = 0. \end{aligned}$$

This completes the proof that  $\nabla \pi_{h,p} q = r_{h,p-1} \nabla q$ . The proof that  $\nabla \times r_{h,p-1} u = w_{h,p-2} \nabla \times u$  is similar.  $\square$

# 9 CLASSICAL SCATTERING THEORY

## 9.1 Introduction

In this chapter we shall present some material from classical scattering theory. The first part of the chapter is devoted to establishing an integral representation, called the Stratton–Chu formula, for the scattered field in a homogeneous isotropic exterior domain. For us this has three uses. First, we use it in Section 9.3.3 to prove that a separation-of-variables solution of Maxwell's equations converges. Second, in the same section, it will be used to derive a representation for the electromagnetic field at large distance from the scatterer. In radar applications this is often the desired output from a numerical simulation. Finally, it will be used in Chapter 12.2 as the basis for another integral representation of the scattered field that can be coupled to a finite element code. This material is entirely classical (except that the formula needs to be verified for a Lipschitz domain), and can be found in many books. Our presentation roughly follows [94].

After deriving the Stratton–Chu formula, we will analyze a scattering problem. In this case the scatterer is a sphere. We shall develop a standard separation-of-variables solution using vector spherical harmonics. Again, this material is standard in most textbooks, and we broadly follow [94]. Having established the separation-of-variables solution, we can then use it to verify mapping properties of the exterior Calderon operator which maps electric boundary data to magnetic boundary data. In particular, we verify the mapping properties in Sobolev spaces. This presentation follows [190, 192]. For a more general discussion of the exterior Calderon operator, see, for example, [73]. Finally, we derive the famous Mei series solution for the scattered field due to the interaction of a plane wave with a perfectly conducting sphere.

## 9.2 Basic integral identities

We start by deriving a classical representation formula for solutions of the Maxwell system in a uniform, homogeneous and isotropic medium called the Stratton–Chu formula. Our presentation follows loosely that of Colton and Kress [94], which is a good reference for more details about proving these results. Note that in this section, in anticipation of later applications, we are dealing with general Lipschitz domains.

Before starting our analysis, we recall that the fundamental solution of the Helmholtz equation is given by(9.1)

$$\Phi(x, y) = \frac{\exp(i\kappa|x-y|)}{4\pi|x-y|}, \quad x \neq y.$$

This function satisfies the Helmholtz equation(9.2)

$$\nabla_y \Phi + \kappa^2 \Phi = -\delta_x \text{ in } \mathbb{R}^3,$$

where  $\Delta_y$  is the Laplacian with respect to  $y$  and  $\delta_x$  is the Dirac-delta function concentrated at  $x$ , so that for any  $u \in C_0^\infty(\mathbb{R}^3)$

$$\int_{\mathbb{R}^3} \delta_x u dV = u(x)$$

(in other words,  $\delta_x \in (C_0^\infty(\mathbb{R}^3))'$ ). This can be extended to continuous functions by a density argument. In addition,  $\Phi$  satisfies the *Sommerfeld radiation condition*, the appropriate radiation condition for the Helmholtz equation, as follows:(9.3)

$$\lim_{\rho_y \rightarrow \infty} \rho_y \left( \frac{\partial \Phi}{\partial \rho_y} - i\kappa \Phi \right) = 0,$$

where  $\rho_y = |y|$  and the limit is uniform in  $j = y/|y|$  for  $x$  in a compact subset of  $\mathbb{R}^3$ . This can be derived using the asymptotic estimates in the proof of Corollary 9.5. In general, we shall use the notation  $\nabla_y$ ,  $\nabla_j$ , and  $\nabla_j \times$  to denote gradient, divergence and curl with respect to  $y$ , and  $dA(y)$  and  $dV(y)$  to remind the reader that the appropriate integrals are with respect to  $y$ .

Clearly,  $\nabla_x \Phi = -\nabla_j \varphi$  if  $x \neq j$ . In addition, using asymptotic estimates (see [94] or the proof of Corollary 9.5), we can show that(9.4)

$$\nabla_y \Phi \times \hat{y} = o\left(\frac{1}{\rho_y^2}\right) \text{ as } \rho_y \rightarrow \infty.$$

We start with a simple representation theorem for a suitably smooth vector function on a bounded Lipschitz domain  $G$  (see, e.g. Theorem 6.1 of [94]). For this theorem the vector functions do not need to satisfy Maxwell's equations.

**Theorem 9.1** Let  $G$  be a bounded Lipschitz domain in  $\mathbb{R}^3$  with boundary  $\partial G$ . Let  $v$  denote the unit outward normal on  $\partial G$ . If  $E, H \in C(G) \cap C^1(G)$  we have, for any  $x \in G$ ,(9.5)

$$\begin{aligned} E(x) = & -\nabla \times \int_{\partial G} (v \times E)(y) \Phi(x, y) dA(y) + \nabla \int_{\partial G} (v \cdot E)(y) \Phi(x, y) dA(y) \\ & -i\kappa \int_{\partial G} (v \times H)(y) \Phi(x, y) dA(y) - \nabla \int_G (v \cdot E)(y) \Phi(x, y) dV(y) \\ & + \nabla \times \int_G (\nabla \times E - i\kappa H)(y) \Phi(x, y) dV(y) \\ & + i\kappa \int_G (\nabla \times E - i\kappa H)(y) \Phi(x, y) dV(y). \end{aligned}$$

**Proof** Let  $e \in \mathbb{R}^3$  be a constant vector. Then, using the properties of the fundamental solution for the Helmholtz equation in (9.1)

$$-e \cdot E(x) = \int_G \left( \nabla_y \Phi + \kappa^2 \Phi \right)(x, y) (e \cdot E)(y) dV(y).$$

But, using (B.6),  $\nabla_y(e\Phi) = -\nabla_y \times (\nabla_y \times (e\Phi)) + \nabla_y \times (\nabla_y \cdot (e\Phi))$ , so, using the Green's formulae (3.27) and (3.24), we have (9.6)

$$\begin{aligned} -e \cdot E(x) &= \int_G \left\{ -\nabla_y \times (e\Phi)(x, y) \cdot (\nabla \times E)(y) - \nabla_y \cdot (e\Phi)(x, y) (\nabla \cdot E)(y) \right. \\ &\quad \left. + \kappa^2 e \cdot E(y) \Phi(x, y) \right\} dV(y) \\ &= -\int_{\partial G} (\mathbf{v}(y) \times (\nabla_y \times (e\Phi))(x, y)) \cdot E(y) dA(y) \\ &\quad + \int_{\partial G} \nabla_y \cdot (e\Phi)(x, y) (\mathbf{v} \cdot E)(y) dA(y). \end{aligned}$$

But regrouping terms and applying the integration by parts formula (3.27) we have

$$\begin{aligned} &\int_G \left\{ \kappa^2 e \cdot E(y) \Phi(x, y) - (\nabla_y \times (e\Phi))(x, y) (\nabla \times E)(y) \right\} dV(y) \\ &= \int_G \left\{ \kappa^2 e \cdot E(y) \Phi(x, y) - i\kappa \nabla_y \times (e\Phi)(x, y) \cdot H(y) \right\} dV(y) \\ &\quad + \int_G (i\kappa \nabla_y \times (e\Phi)(x, y) H(y) - \nabla_y \times (e\Phi)(x, y) (\nabla \times E)(y)) dV(y) \\ &= -i\kappa \int_G \Phi(x, y) \cdot (\nabla \times H + i\kappa E)(y) dV(y) \\ &\quad + \int_G \nabla_y \times (e\Phi)(x, y) \cdot (i\kappa H - \nabla \times E)(y) dV(y) \\ &\quad - \int_{\partial G} i\kappa \mathbf{v}(y) \times (e\Phi(x, y)) \cdot H(y) dA(y). \end{aligned}$$

Using this in (9.6), together with the fact that for any continuously differentiable vector function  $\xi$  on  $G$  we have

$$\nabla_y \times (e\Phi(x, y)) \cdot \xi(y) = e \cdot \nabla_x \times (\Phi(x, y) \xi(y))$$

and

$$\nabla_y \times (e\Phi)(x, y) = -e \cdot (\nabla_x \Phi)(x, y),$$

we obtain

$$\begin{aligned} e \cdot E(x) &= -e \cdot \left\{ \nabla \times \int_G \Phi(x, y) (i\kappa H - \nabla \times E)(y) dV(y) \right. \\ &\quad + \nabla \int_G \Phi(x, y) \nabla \cdot E(y) dV(y) \\ &\quad - i\kappa \int_G \Phi(x, y) (\nabla \times H + i\kappa E)(y) dV(y) \\ &\quad - i\kappa \int_{\partial G} \Phi(x, y) H(y) \times \mathbf{v}(y) dA(y) \\ &\quad - \nabla \times \int_{\partial G} \Phi(x, y) H(y) \times \mathbf{v}(y) dA(y) \\ &\quad \left. - \nabla \int_{\partial G} \Phi(x, y) \mathbf{v} \cdot E(y) dA \right\}. \end{aligned}$$

Since this holds for any  $\epsilon$ , we can remove  $\epsilon$  from the above equation and reordering the triple products gives the conclusion of this theorem.  $\square$

Now we assume that  $E$  and  $H$  satisfy the homogeneous isotropic Maxwell system(9.7a)

$$\nabla \times E - i\kappa H = 0, \quad (9.7b)$$

$$\nabla \times H + i\kappa E = 0,$$

in the sense of distributions in a Lipschitz domain  $G$ . At first,  $G$  is assumed to be bounded, but later we shall generalize to the unbounded complement of a bounded domain.

Under the assumption that  $E, H$  satisfy the Maxwell system (9.7), we obtain the following simplification of the previous theorem, which states the famous Stratton–Chu formula on a bounded domain.

**Theorem 9.2** *Let  $G$  be a bounded Lipschitz domain with unit outward normal  $v$ . Let  $E, H \in H(\text{curl}; G)$  be solutions of (9.7) in  $G$ . Then, for any  $x \in G$ ,* (9.8)

$$\begin{aligned} E(x) = & -\nabla \times \int_{\partial G} (v \times E)(y) \Phi(x, y) dV(y) \\ & + \frac{1}{i\kappa} \nabla \times \int_{\partial G} (v \times H)(y) \Phi(x, y) dV(y). \end{aligned}$$

**Remark 9.3** A similar formula holds for  $H$ :

$$\begin{aligned} H(x) = & -\nabla \times \int_{\partial G} (v \times H)(y) \Phi(x, y) dA(y) \\ & - \frac{1}{i\kappa} \nabla \times \nabla \times \int_{\partial G} v(y) \times E(y) \Phi(x, y) dA(y). \end{aligned}$$

(see [94] for a derivation).

**Proof of Theorem 9.2** If we eliminate  $H$  from (9.7), we obtain(9.9)

$$\nabla \times (\nabla \times E) - \kappa^2 E = 0 \quad \text{in } G.$$

Taking the divergence of this equation, we see that  $\nabla \cdot E = 0$  in  $G$ . Now using (B.6) we obtain(9.10)

$$\nabla \times (\nabla \times E) - \kappa^2 E = -\Delta E - \kappa^2 E,$$

so that each component of  $E$  satisfies the Helmholtz equation on compact subsets of  $G$ . Hence using interior regularity estimates for solutions of second-order elliptic problems (see, e.g. Theorem 4.16 of McLean [215]) we see that the components of  $E$  are smooth functions of position away from the boundary (i.e. on compact subsets of  $G$ ).

Let  $\{G_n\}_{n=1}^\infty$  be a sequence of Lipschitz domains that are compactly contained in  $G$  and expand to fill  $G$  as  $n \rightarrow \infty$ . The previous theorem holds on each  $G_n$ , and since Maxwell's equations hold in the classical sense in  $G_n$ , we can take the divergence of each equation to show that  $\nabla \cdot H = \nabla \cdot E = 0$  in  $G_n$ . Using this fact and the fact that  $E, H$  satisfy (9.7), we have(9.11)

$$\begin{aligned} E(x) &= -\nabla \times \int_{\partial G_n} v(y) \times E(y) \Phi(x, y) dA(y) \\ &\quad + \nabla \int_{\partial G_n} v(y) \cdot E(y) \Phi(x, y) dA(y) \\ &\quad - i\kappa \nabla \times \int_{\partial G_n} (v(y) \times H(y)) \Phi(x, y) dA(y). \end{aligned}$$

Using the fact that  $x \notin \partial G_n$  for  $n$  large enough, together with (9.2) and (9.10), we have(9.12)

$$\begin{aligned} -i\kappa \int_{\partial G_n} v(y) \times H(y) \Phi(x, y) dA(y) \\ &= -\frac{1}{i\kappa} \int_{\partial G_n} (v \times H)(y) \Delta_x \Phi(x, y) dA(y) \\ &= i\kappa \nabla \times \nabla \times \int_{\partial G_n} (v(y) \times H(y)) \Phi(x, y) dA(y) \\ &\quad - \frac{1}{i\kappa} \nabla \nabla \cdot \int_{\partial G_n} (v \times H)(y) \Phi(x, y) dA(y). \end{aligned}$$

By integration by parts using (3.27) and using Maxwell's equations (9.7) together with the fact that  $E$  is divergence free, we have(9.13)

$$\begin{aligned} \nabla \cdot \int_{\partial G_n} (v \times H)(y) \Phi(x, y) dA(y) &= - \int_{\partial G_n} (v \times H)(y) \cdot \nabla_y \Phi(x, y) dA(y) \\ &= - \int_{\partial G} \nabla \times H(y) \cdot \nabla_y \Phi(x, y) dV(y) \\ &= i\kappa \int_{G_n} E(y) \cdot \nabla_y \Phi(x, y) dV(y) \\ &= i\kappa \int_{\partial G_n} \Phi(x, y) (v \cdot E)(y) dA(y). \end{aligned}$$

Using (9.13) in (9.12) and the result in (9.11) proves the desired identity on  $G_n$ . Another application of Greens formula and the Lebesgue dominated convergence theorem shows that

$$\nabla \times \int_{\partial G_n} (\mathbf{v} \times \mathbf{E})(y) \Phi(x, y) dA(y) \rightarrow \nabla \times \int_{\partial G} (\mathbf{v} \times \mathbf{E})(y) \Phi(x, y) dA(y)$$

as  $n \rightarrow \infty$ , with a similar convergence result for the other term. Hence, we have proved the desired result on  $G$ .  $\square$

Now we wish to extend this result to unbounded domains. This involves a limiting argument in which the outer boundary of  $G$  tends to infinity. The radiation condition can then be used to show that the contribution of this boundary vanishes in the limit.

Let  $D$  be a bounded Lipschitz domain in  $\mathbb{R}^3$  whose complement is connected. A solution  $E, H$  of the Maxwell system (9.7) in  $\mathbb{R}^3 \setminus D^-$  is said to be *radiating* if it satisfies the Silver–Müller radiation condition,(9.14)

$$\lim_{\rho \rightarrow \infty} \rho (H \times \hat{\mathbf{x}} - E) = 0,$$

where  $\rho = |x|$  and the limit holds uniformly in all directions  $\hat{\mathbf{x}} = \mathbf{x}/\rho$ .

Let  $H_{loc}(\text{curl}; \mathbb{R}^3 \setminus D^-)$  denote the space of functions  $u \in H(\text{curl}; B_R \setminus D^-)$  for every ball  $B_R$  containing  $D$  in its interior. The following theorem gives the Stratton–Chu formula for an unbounded domain.

**Theorem 9.4** Let  $v$  denote the exterior normal to  $D$  (i.e. interior to  $\mathbb{R}^3 \setminus D^-$ ), where  $D$  is the domain defined prior to this theorem. Let  $E, H \in H_{loc}(\text{curl}; \mathbb{R}^3 \setminus D^-)$  satisfy the Maxwell system (9.7) in  $\mathbb{R}^3 \setminus D^-$  and, in addition, suppose  $E, H$  are radiating. Then for each  $x \in \mathbb{R}^3 \setminus D^-$ (9.15)

$$\begin{aligned} E(x) = & \nabla \times \int_{\partial D} (\mathbf{v} \times \mathbf{E})(y) \Phi(x, y) dA(y) \\ & - \frac{1}{ik} \nabla \times \nabla \times \int_{\partial D} (\mathbf{v} \times \mathbf{H})(y) \Phi(x, y) dA(y). \end{aligned}$$

**Proof** Let  $\Omega_R$  denote  $\Omega \cap B_R$ , where  $B_R$  is the ball of radius  $R$  centered at the origin. We choose  $R$  sufficiently large that  $D^- \subset \Omega_R$ . The normal  $v$  is chosen to point into  $\Omega_R$  on  $\partial\Omega$  and out of  $\Omega_R$  on  $\partial B_R$ . Using the previous two theorems, taking into account the direction of the normal, for each  $x \in \Omega_R$ ,(9.16)

$$\begin{aligned} E(x) = & \nabla \times \int_{\partial D} (\mathbf{v} \times \mathbf{E})(y) \Phi(x, y) dA(y) \\ & - \frac{1}{ik} \nabla \times \nabla \times \int_{\partial D} (\mathbf{v} \times \mathbf{H})(y) \Phi(x, y) dA(y) \\ & - \nabla \times \int_{\partial B_R} (\mathbf{v} \times \mathbf{E})(y) \Phi(x, y) dA(y) \\ & + \nabla \int_{\partial B_R} (\mathbf{v} \cdot \mathbf{E})(y) \Phi(x, y) dA(y) \\ & - ik \int_{\partial B_R} (\mathbf{v} \times \mathbf{H})(y) \Phi(x, y) dA(y), \end{aligned}$$

where we have used the transformed boundary terms of Theorem 9.2 on  $\partial\Omega$  and the boundary term of Theorem 9.1 on  $\partial B_R$ . We shall now show that the boundary terms on  $\partial B_R$  vanish as  $R \rightarrow \infty$ .

First we show that(9.17)

$$\int_{\partial B_R} |E(y)|^2 dA(y) = o(1) \quad \text{as } R \rightarrow \infty.$$

From the radiation condition (9.14) we have

$$\int_{\partial B_R} |H \times v|^2 + |E|^2 - 2\Re(v \times E \cdot \bar{H}) dA = \int_{\partial B_R} |H \times v - E|^2 dA \rightarrow 0.$$

as  $R \rightarrow \infty$ . But (recalling the choice of the direction of the normal vectors) using (3.27) and the Maxwell system (9.7)(9.18)

$$\begin{aligned} & \int_{\partial B_R} v \times E \cdot \bar{H} dA - \int_{\partial D} v \times E \cdot \bar{H} dA \\ &= \int_{\Omega_R} \nabla \times E \cdot \bar{H} - E \cdot \nabla \times \bar{H} dV \\ &= i\kappa \int_{\Omega_R} |H|^2 - |E|^2 dV, \end{aligned}$$

so we obtain the conservation of energy result,(9.19)

$$\Re \int_{\partial B_R} v \times E \cdot \bar{H} dA = \Re \int_{\partial D} v \times E \cdot \bar{H} dA,$$

and thus we conclude that

$$\lim_{R \rightarrow \infty} \int_{\partial B_R} |H \times v|^2 + |E|^2 dA = 2\Re \int_{\partial D} (v \times E) \cdot \bar{H} dA.$$

Hence

$$\int_{\partial B_R} |H \times v|^2 dA = o(1) \quad \text{and} \quad \int_{\partial B_R} |E|^2 dA = o(1)$$

as  $R \rightarrow \infty$ .

Now using this result and the Cauchy–Schwarz inequality, we have that(9.20)

$$\begin{aligned} & \left| \nabla \times \int_{\partial B_R} (\mathbf{v} \times \mathbf{E})(y) \Phi(x, y) dA(y) \right| \\ &= \left| \int_{\partial B_R} (\mathbf{v} \times \mathbf{E})(y) \times \nabla_x \Phi(x, y) dA(y) \right| \\ &\leq \| \mathbf{E} \|_{L^2(B_R)}^3 \| \nabla_x \Phi \times \mathbf{v} \|_{L^2(B_R)}^3 \xrightarrow{R \rightarrow \infty} 0 \end{aligned}$$

as  $R \rightarrow \infty$ , where we have also used the asymptotic estimate (9.4).

The remaining terms on  $\partial B_R$  in (9.16) are estimated by expanding as follows:(9.21)

$$\begin{aligned} & -\nabla \times \int_{\partial B_R} (\mathbf{v} \cdot \mathbf{E})(y) \Phi(x, y) dA(y) + i\kappa \int_{\partial B_R} (\mathbf{v} \times \mathbf{H})(y) \times \Phi(x, y) dA(y) \\ &= -\nabla \int_{\partial B_R} (\mathbf{v} \cdot \mathbf{E})(y) \Phi(x, y) dA(y) - \int_{\partial B_R} i\kappa \mathbf{E}(y) \Phi(x, y) dA(y) \\ &\quad + i\kappa \int_{\partial B_R} (\mathbf{v} \times \mathbf{H} + \mathbf{E})(y) \times \Phi(x, y) dA(y). \end{aligned}$$

The last term on the right-hand side of this expansion tends to zero as  $R \rightarrow \infty$  since  $|\Phi(x, y)| = O(1/R)$ , and the radiation condition (9.14) estimates the term  $|\mathbf{v} \times \mathbf{H} + \mathbf{E}| = o(1/R)$ . The remaining terms in this expansion need even more manipulation using standard vector identities,(9.22)

$$\begin{aligned} & -\nabla \times \int_{\partial B_R} (\mathbf{v} \cdot \mathbf{E})(y) \Phi(x, y) dA(y) - \int_{\partial B_R} i\kappa \mathbf{E}(y) \times \Phi(x, y) dA(y) \\ &= -\nabla \int_{\partial B_R} (\mathbf{v} \cdot \mathbf{E})(y) \nabla_y \Phi(x, y) dA(y) - i\kappa \mathbf{E}(y) \Phi(x, y) dA(y) \\ &= \int_{\partial B_R} \mathbf{v}(y) \times ((\nabla_y \Phi(x, y)) \times \mathbf{E}(y)) dA(y) \\ &\quad + \int_{\partial B_R} \mathbf{E}(y) \cdot (\mathbf{v}(y) \cdot \nabla_y \Phi(x, y) - i\kappa \Phi(x, y)) dA(y). \end{aligned}$$

The first term on the right-hand side above tends to zero as  $R \rightarrow \infty$  as can be seen by using the estimates (9.4), (9.17) and the Cauchy–Schwarz inequality as in the proof of (9.20). The second term also vanishes using the Cauchy–Schwarz inequality, (9.17) and the Sommerfeld radiation condition for  $\Phi$  in (9.3). Combining (9.20)–(9.22) shows that the integrals over  $\partial B_R$  in (9.16) vanish as  $R \rightarrow \infty$  and we have proved the desired result.  $\square$

Let us note that we could equally well have used the integral radiation condition

$$\int_{\partial B_R} |H \times v - E|^2 dA \rightarrow 0 \text{ as } R \rightarrow \infty.$$

Note also that the result on conservation of energy (9.19), has a physical interpretation. Reordering the triple products, we have

$$\Re \int_{\partial B_R} \mathbf{v} \cdot \mathbf{E} \times \bar{\mathbf{H}} dA = \Re \int_{\partial D} \mathbf{v} \cdot \mathbf{E} \times \bar{\mathbf{H}} dA .$$

The quantity  $\Re(\mathbf{E} \times \mathbf{H})$  is the classical Poynting vector and  $\Re(\mathbf{v} \cdot \mathbf{E} \times \mathbf{H})$  gives the flux density of energy transport through a surface normal to  $\mathbf{v}$  (see p. 79 of [73]).

As a corollary of the previous result, we have the following asymptotic representation for  $\mathbf{E}$  at great distances from the scatterer. This result expresses the intuitive fact that, far from a scatterer, the scattered field is an expanding spherical wave with amplitude modulated in different directions.

**Corollary 9.5** Every radiating solution of Maxwell's equations (9.7) in  $\mathbb{R}^3 \setminus D^-$  has the asymptotic behavior (9.23)

$$\mathbf{E}(\mathbf{x}) = \frac{\exp(i\kappa|\mathbf{x}|)}{|\mathbf{x}|} \left\{ \mathbf{E}_\infty(\hat{\mathbf{x}}) + o\left(\frac{1}{|\mathbf{x}|}\right) \right\} \quad \text{as } |\mathbf{x}| \rightarrow \infty$$

uniformly in  $\mathcal{O} = \mathbf{x}/|\mathbf{x}|$ . Furthermore, (9.24)

$$\begin{aligned} \mathbf{E}_\infty(\hat{\mathbf{x}}) = & \frac{i\kappa}{4\pi} \hat{\mathbf{x}} \times \int_{\partial D} ((\mathbf{v} \times \mathbf{E})(y) \\ & + (\mathbf{v} \times \mathbf{H})(y) \times \hat{\mathbf{x}}) \exp(-i\kappa \hat{\mathbf{x}} \cdot y) dA(y) . \end{aligned}$$

**Remark 9.6** There is a corresponding expansion for the magnetic field in terms of  $\mathbf{H}_\infty(\hat{\mathbf{x}})$ . However, there is no need to compute both  $\mathbf{E}_\infty$  and  $\mathbf{H}_\infty$  since  $\mathbf{H}_\infty = \hat{\mathbf{x}} \times \mathbf{E}_\infty$ . Note that  $\mathbf{E}_\infty$  is tangential to the unit sphere, that is  $\mathcal{O} \cdot \mathbf{E}_\infty(\hat{\mathbf{x}}) = 0$ .

**Definition 9.7** The vector function  $\mathbf{E}_\infty(\hat{\mathbf{x}})$  is called the *electric far field pattern* or the far field pattern of the electric field.

Frequently, particularly for radar computations, it is the far field pattern that is the desired output from a Maxwell solver. We shall discuss this more in Chapter 13.

**Proof of Corollary 9.5** Using the asymptotic expansion

$$|\mathbf{x} - \mathbf{y}| = \sqrt{|\mathbf{x}|^2 - 2\mathbf{x} \cdot \mathbf{y} + |\mathbf{y}|^2} = |\mathbf{x}| - \hat{\mathbf{x}} \cdot \mathbf{y} + o\left(\frac{1}{|\mathbf{x}|}\right),$$

we can derive the following estimates. For any constant vector  $a$ ,

$$\nabla_{\mathbf{x}} \times (a\Phi(\mathbf{x}, \mathbf{y})) = \frac{i\kappa \exp(i\kappa|\mathbf{x}|)}{4\pi|\mathbf{x}|} \left\{ \exp(-i\kappa \hat{\mathbf{x}} \cdot \mathbf{y}) \hat{\mathbf{x}} \times a + o\left(\frac{1}{|\mathbf{x}|}\right) \right\}$$

and (9.25)

$$\begin{aligned} \nabla_{\mathbf{x}} \times \nabla_{\mathbf{x}} \times (a\Phi(\mathbf{x}, \mathbf{y})) = & \kappa^2 \frac{\exp(i\kappa|\mathbf{x}|)}{4\pi|\mathbf{x}|} \left\{ \exp(-i\kappa \hat{\mathbf{x}} \cdot \mathbf{y}) \hat{\mathbf{x}} \times (a \times \hat{\mathbf{x}}) \right. \\ & \left. + o\left(\frac{1}{|\mathbf{x}|}\right) \right\} \end{aligned}$$

as  $|x| \rightarrow \infty$  uniformly for all  $y \in \partial\Omega$ . Using these formulae in (9.15) and taking limits we obtain the desired result.  $\square$

## 9.3 Scattering by a sphere

We start our discussion of electromagnetic scattering problems by restricting ourselves to a particularly simple geometry: scattering by a sphere. Besides the fact that we shall need results for this problem in order to provide a truncation scheme for one of the finite element methods discussed in a later chapter, we can also use this problem to prove some preliminary theorems concerning uniqueness for the general scattering problem.

Let  $D = B_R$ , i.e. the scatterer is a sphere of radius  $R$  centered at the origin. Then, given an incident field  $E^i$  that satisfies the homogeneous isotropic Maxwell's equations in all of  $\mathbb{R}^3$ , we seek to find a total electric field  $E$  and scattered field  $E^s$  such that (9.26a)

$$\begin{aligned} \nabla \times (\nabla \times E) - \kappa^2 E &= 0 \quad \text{in } \mathbb{R}^3 \setminus \bar{B}_R, \\ \hat{x} \times E &= 0 \quad \text{on } \partial B_R, \end{aligned} \tag{9.26b}$$

$$E = E^i + E^s \quad \text{in } \mathbb{R}^3 \setminus \bar{B}_R, \tag{9.26c}$$

$$\lim_{\rho \rightarrow \infty} \rho(\nabla \times E^s \times \hat{x} - i\kappa E^s) = 0. \tag{9.26d}$$

Here  $\hat{x} = x/|x|$  and  $\rho = |x|$ . The limit in (9.26d) is uniform in  $\hat{x}$ .

In this case it will be convenient to solve for  $E^s$ , so we eliminate  $E$  using (9.26c) so that  $E^s$  satisfies (9.27a)

$$\begin{aligned} \nabla \times (\nabla \times E^s) - i\kappa E^s &= 0 \quad \text{in } \mathbb{R}^3 \setminus \bar{B}_R, \\ \widehat{\hat{x}} \times E^s &= g = -\hat{x} \times E^i \quad \text{on } \partial B_R, \end{aligned} \tag{9.27b}$$

$$\begin{aligned} \widehat{\hat{x}} \times E^s &= g = -\hat{x} \times E^i \quad \text{on } \partial B_R, \\ \lim_{\rho \rightarrow \infty} \rho(\nabla \times E^s \times \hat{x} - i\kappa E^s) &= 0. \end{aligned} \tag{9.27c}$$

We shall use separation of variables to solve this problem in a classical way. Our presentation is taken from [94], but the material is very well known and can be found, for example, in [236, 284]. It turns out that the appropriate functions to use in our solution are spherical harmonics and spherical Bessel functions. For complete details, see [203].

The idea is to use solutions of the simpler Helmholtz equation to build solutions of Maxwell's equations. In particular, let  $u$  be a classical solution of the Helmholtz equation (9.28)

$$\Delta u + \kappa^2 u = 0 \quad \text{in } \mathbb{R}^3 \setminus \bar{B}_R.$$

Then the function  $u$  given by

(9.29)

$$\mathbf{u} = \nabla \times (\mathbf{u} \mathbf{x})$$

is a solution of (9.27a). The function  $u$  is called a *Debye potential*. To see that  $u$  in (9.29) is a solution of Maxwell's equations, we rewrite (9.29), using the expression for the curl in spherical polar coordinates  $(\rho, \theta, \varphi)$ , noting that  $xu = \rho u e_\rho$ , where  $e_\rho$  is a unit vector in the  $x$  direction (see Section A.2) to obtain (9.30)

$$\mathbf{u} = -\frac{\partial u}{\partial \theta} e_\varphi + \frac{1}{\sin \theta} \frac{\partial u}{\partial \varphi} e_\theta,$$

where  $e_\theta$  and  $e_\varphi$  are unit vectors for spherical polar coordinates. Then, using the expression for the curl in these coordinates given by (A.1)(9.31)

$$\begin{aligned} \nabla \times \mathbf{u} &= -\frac{1}{\rho \sin \theta} (\Delta_{\partial B_1} u) e_\rho + \frac{1}{\rho} \frac{\partial}{\partial \rho} \left( \rho \frac{\partial u}{\partial \theta} \right) e_\varphi \\ &\quad + \frac{1}{\rho \sin \theta} \frac{\partial}{\partial \rho} \left( \rho \frac{\partial u}{\partial \varphi} \right) e_\theta, \end{aligned}$$

where the Laplace–Beltrami operator  $\Delta_{\partial B_1}$  for the surface of the unit sphere is given by

$$\Delta_{\partial B_1} u = \frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial u}{\partial \theta} \right) + \frac{1}{\sin^2 \theta} \frac{\partial^2 u}{\partial \varphi^2}.$$

Hence, again using the formula for curl in spherical coordinates,

$$\Delta \times (\nabla \times \mathbf{u}) = -\frac{1}{\sin \theta} \frac{\partial}{\partial \varphi} (\Delta u) e_\theta + \frac{\partial}{\partial \theta} (\Delta u) e_\varphi$$

and, using this expansion in Maxwell's equations, we obtain

$$\Delta \times (\nabla \times \mathbf{u}) - \kappa^2 u = -\frac{1}{\sin \theta} \frac{\partial}{\partial \varphi} (\Delta u + \kappa^2 u) e_\theta + \frac{\partial}{\partial \theta} (\Delta u + \kappa^2 u) e_\varphi.$$

Thus,  $u$  satisfies Maxwell's equations if  $u$  satisfies the Helmholtz equation. We have not discussed the radiation condition, but later it will turn out that if  $u$  satisfies the Sommerfeld radiation condition (9.3) then  $u$  will satisfy the Silver–Müller radiation condition (9.27c).

In order to use separation of variables to solve (9.28), we use spherical coordinates  $(\rho, \theta, \varphi)$  and rewrite (9.28) as

$$\frac{1}{\rho^2} \frac{\partial}{\partial \rho} \left( \rho^2 \frac{\partial u}{\partial \rho} \right) + \frac{1}{\rho^2} \Delta_{\partial B_1} u + \kappa^2 u = 0.$$

If  $u = u_1(\rho)u_2(\theta, \varphi)$ , it follows that

$$\frac{1}{u_1} \left( \frac{\partial}{\partial \rho} \left( \rho^2 \frac{\partial u_1}{\partial \rho} \right) + \kappa^2 \rho^2 u_1 \right) + \frac{1}{u_2} \Delta_{\partial B_1} u_2 = 0.$$

Thus, we need to solve  $\Delta_{\partial B_1} u_2 = \delta u_2$  on  $\partial B_1$ , for  $\delta$  constant, or, equivalently, find the eigenvalues and eigenfunctions of the Laplace–Beltrami operator on  $\partial B_1$ .

These turn out to be the classical spherical harmonics, and we present a summary of their properties in the next section. Once  $u_2$  is known, we have(9.32)

$$\frac{\partial}{\partial \rho} \left( \rho^2 \frac{\partial u_1}{\partial \rho} \right) + \kappa^2 \rho^2 u_1 + \delta u_1 = 0.$$

Introducing  $t = \kappa \rho$ , this can be rewritten as

$$\frac{\partial}{\partial t} \left( t^2 \frac{\partial u_1}{\partial t} \right) + t^2 u_1 + \delta u_1 = 0.$$

For appropriate choices of  $\delta$ , this is the spherical Bessel differential equation. Solutions of this equation are the spherical Bessel functions, and these are studied in Section 9.3.2. Using these functions we shall obtain a series solution of (9.27b) and (9.27c) in Section 9.3.3. The properties of this solution are studied in the remainder of the chapter.

### 9.3.1 Spherical harmonics

On the surface of a sphere, the eigenfunctions of the Laplace–Beltrami operator turn out to be the classical spherical harmonics which we define next. This section provides no proofs, since the results are rather standard. Good references are [94, 236]. Recall that  $\mathcal{P}_n$  denotes the space of homogeneous polynomials of degree  $n$  in  $x_1, x_2$  and  $x_3$ .

**Definition 9.8** The trace on  $\partial B_1$  of a function  $u \in \mathcal{P}_n$  such that  ${}^\Delta u = 0$  in  $\mathbb{R}^3$  is called a *spherical harmonic* of order  $n$  (recall that  $\mathcal{P}_n$  is the space of homogeneous polynomials of degree exactly  $n$ , see (5.7)).

It turns out that there are exactly  $2n + 1$  linearly independent spherical harmonics of order  $n$ . Now we seek to write down explicit formulae for the spherical harmonics using spherical polar coordinates  $(\rho, \theta, \varphi)$  (see Section A.2). In this coordinate system, any polynomial  $u \in \mathcal{P}_n$  must have the form  $u = \rho^n Y_n(\theta, \varphi)$ . Using the standard expansion for the Laplacian in spherical polar coordinates, we see that  ${}^\Delta u = 0$  implies(9.33)

$$\frac{1}{\sin \theta} \frac{\partial}{\partial \theta} (\sin \theta \frac{\partial}{\partial \theta} Y_n) + \frac{1}{\sin^2 \theta} \frac{\partial^2 Y_n}{\partial \varphi^2} + n(n+1)Y_n = 0,$$

or, using the Laplace–Beltrami operator for the surface of the unit sphere,

$$\Delta_{\partial B_1} Y_n + n(n+1)Y_n = 0 \quad \text{on } \partial B_1.$$

This tells us that the eigenvalues of  $\Delta_{\partial B_1}$  on  $\partial B_1$  are  $-n(n+1)$ . Since  $\Delta_{\partial B_1}^{-1}$  is a self-adjoint compact operator from  $L^2(\partial B_1)$  to  $L^2(\partial B_1)$ , we can apply the Hilbert–Schmidt theory (Theorem 2.36) to conclude that

$$\int_{\partial B_1} Y_n \bar{Y}_m dA = 0 \quad \text{for } n \neq m.$$

Alternatively, this can be proved directly via Green's theorem [94].

The easiest spherical harmonics to obtain explicitly are those that do not depend on the angle  $\varphi$ . In this case, using the substitution  $t = \cos\theta$ , and denoting the  $\varphi$ -independent spherical harmonic by  $P_n(t)$ , we see that (9.33) becomes the *Legendre differential equation*(9.34)

$$\frac{\partial}{\partial t} \left( 1 - t^2 \right) \frac{\partial}{\partial t} P_n + n(n+1)P_n = 0.$$

The solutions of this equation that are polynomials of degree  $n$  in  $t$  are called *Legendre polynomials*. We summarize the relevant results concerning these polynomials next. The proof of this theorem can be found in [94] except for Rodrigues' formula (9.35). This is proved in Lebedev [203].

**Theorem 9.9A** *A family of solutions of (9.34) is provided by the Legendre polynomials denoted  $P_n(t)$  and given by Rodrigues' formula(9.35)*

$$P_n(t) = \frac{(-1)^n}{2^n n!} \frac{d^n}{dt^n} (1 - t^2)^n, \quad n = 0, 1, 2, \dots$$

*These polynomials satisfy the recurrence relation*

$$(n+1)P_n(t) - (2n+1)tP_n(t) + nP_{n-1}(t) = 0, \quad n = 1, 2, \dots$$

*and have the orthogonality property*

$$\int_{-1}^1 P_n(t)P_m(t) dt = \frac{2}{2n+1} \delta_{nm}.$$

*Finally for  $-1 \leq t \leq 1$ ,  $|P_n(t)| \leq 1$ ,  $n = 0, 1, 2, \dots$*

**Remark 9.10** *Rodrigues' formula implies that*

$$P_0(t) = 1, \quad P_1(t) = t, \quad P_2(t) = \frac{1}{2} \left( 3t^2 - 1 \right).$$

*In fact, for  $n$  odd,  $P_n(t)$  is a polynomial of odd powers of  $t$  and, for  $n$  even, is a polynomial of even powers.*

Using the Legendre polynomials we can now seek solutions of (9.33) that depend on  $\theta$  and  $\varphi$ . Using separation of variables in (9.33) we see that the  $m$ th spherical harmonic of order  $n$ , denoted  $Y_n^m(\theta, \varphi)$ , must have the form

$$Y_n^m(\theta, \varphi) = f(\cos \theta) \exp(im\varphi)$$

for some function  $f$ . Setting  $t = \cos\theta$  in (9.33) we see that  $f$  must satisfy the *associated Legendre differential equation*(9.36)

$$\left( 1 - t^2 \right) f'' - 2tf'(t) + \left[ n(n+1) - \frac{m^2}{1-t^2} \right] f(t) = 0.$$

By writing  $f(t) = (1 - t^2)^{n/2}g(t)$ , we can prove (see [94] for details) that the solutions of (9.36) are given by  $f = P_n^m(t)$ , where  $P_n^m(t)$  denotes the  $m$ th associated Legendre function of order  $n$  given by the following extension of Rodrigues formula

$$P_n^m(t) = \left(1 - t^2\right)^{m/2} \left(\frac{d}{dt}\right)^m P_n(t), \quad m = 0, 1, 2, \dots, n.$$

Note that there does not seem to be a unique normalization for  $P_n^m$  in the literature and we have adopted the one used in [94]. Thus, the corresponding spherical harmonic is

$$Y_n^m(\theta, \varphi) = \gamma_n^m P_n^m(\cos \theta) \exp(im\varphi) \quad n = 0, 1, 2, \dots, \quad m = 0, 1, 2, \dots, n.$$

where  $\gamma_n^m$  is a normalization constant given in the next theorem.

**Theorem 9.11** *The spherical harmonics (9.37)*

$$Y_n^m(\theta, \varphi) = \sqrt{\frac{2n+1}{4\pi} \frac{(n-|m|)!}{(n+|m|)!}} P_n^{|m|}(\cos \theta) \exp(im\varphi)$$

for  $m = -n, \dots, n$  and  $n = 0, 1, 2, \dots$  form a complete orthonormal system in  $L^2(\partial B_1)$ .

**Remark 9.12** *It might appear that this theorem is a consequence of the Hilbert–Schmidt theory. However, this would only be a complete argument if we could show that (9.37) gives all eigenfunctions of  $\Delta_{\partial B_1}$ . A direct verification of the theorem is given in [94].*

Using spherical polar coordinates, we have denoted by  $Y_n^m(\theta, \varphi)$  the  $m$ th spherical harmonic of order  $n$ . We will find it convenient to also use the notation  $Y_n^m(\hat{x})$ , where  $\hat{x}$  is the unit vector with spherical polar coordinates  $(\theta, \varphi)$ .

A useful expansion using spherical harmonics is given by the *addition theorem*:

$$\sum_{m=-1}^n Y_n^m(\hat{x}) \overline{Y_n^m(\hat{y})} = \frac{2n+1}{4\pi} P_n(\cos \xi),$$

where  $\xi$  denotes the angle between the unit vectors  $\hat{x}$  and  $\hat{y}$ .

### 9.3.2 Spherical Bessel functions

Having determined that the spherical harmonics are the eigenfunctions of the Laplace–Beltrami operator on the unit sphere we can now determine the radial dependence of solutions of the Helmholtz equation (i.e. the function  $u_1$  in (9.32)). Using the fact that  $\delta = -n(n + 1)$  we can rewrite equation (9.32) as (9.38)

$$\frac{\partial}{\partial \rho} \left( \rho^2 \frac{\partial u_1}{\partial \rho} \right) + \kappa^2 \rho^2 u_1 - n(n + 1) u_1 = 0.$$

Using the change of variable  $t = \kappa Q$ , this can be rewritten as the spherical Bessel differential equation

$$\frac{\partial}{\partial t} \left( t^2 \frac{\partial u_1}{\partial t} \right) + \left( t^2 - n(n+1) \right) u_1 = 0.$$

The direct series solution of this differential equation gives two families of solutions:(9.39)

$$j_n(t) = \sum_{l=0}^{\infty} \frac{(-1)^l t^{n+2l}}{2^l l! 1 \cdot 3 \cdots (2n+2l+1)}$$

and(9.40)

$$y_n(t) = -\frac{(2n)!}{2^n n!} \sum_{l=0}^{\infty} \frac{(-1)^l t^{2l-n-1}}{2^l l! (-2n+1)(-2n+3)\cdots(2n+2l+1)}.$$

The function  $j_n$  is called the spherical Bessel function of order  $n$  and is analytic for all  $t \in \mathbb{R}$ . The function  $y_n$  is the spherical Neumann function and is analytic for  $t \in (0, \infty)$ . The corresponding spherical Hankel functions  $h_n^{(1)}$  and  $h_n^{(2)}$  are defined by

$$h_n^{(1)} = j_n + iy_n, \quad \text{and} \quad h_n^{(2)} = j_n - iy_n.$$

In fact,  $j_n$  and  $y_n$  can be expressed in terms of trigonometric functions. In particular,

$$j_0(t) = \frac{\sin(t)}{t}, \quad y_0(t) = -\frac{\cos(t)}{t},$$

and we see that

$$h_0^{(1)}(t) = \frac{\exp(it)}{it}, \quad h_0^{(2)}(t) = -\frac{\exp(-it)}{it}.$$

We have the following theorem summarizing the asymptotic properties of the spherical Hankel functions. The proof may be found in [94, 203].

**Theorem 9.13** If  $f_n = j_n$ ,  $f_n = y_n$ ,  $f_n = h_n^{(1)}$  or  $f_n = h_n^{(2)}$ , the following recurrence relations hold(9.41)

$$\begin{aligned} f_{n+1}(t) + f_{n-1} &= \frac{2n+1}{t} f_n(t), \quad n = 1, 2, \dots, \\ f_{n+1}(t) &= -t^n \frac{d}{dt} \left\{ t^{-n} f_n(t) \right\}, \quad n = 0, 1, 2, \dots, \end{aligned} \tag{9.42}$$

$$f'_n(t) = f_{n-1}(t) - \frac{n+1}{t} f_n(t), \quad n = 1, 2, \dots. \tag{9.43}$$

For fixed  $n$ , the following asymptotic expansions hold:(9.44)

$$h_n^{(1)}(t) = \frac{1}{t} \exp(i(t - n\pi/2)) \left\{ 1 + o\left(\frac{1}{t}\right) \right\},$$

$$h_n^{(1)}(t) = \frac{1}{t} \exp(i(t - n\pi/2)) \left\{ 1 + o\left(\frac{1}{t}\right) \right\}, \quad (9.45)$$

as  $t \rightarrow \infty$ . The following asymptotic relations hold for large  $n$ : (9.46)

$$h_n^{(1)}(z) = \frac{(2n+1)!!}{iz^{n+1}} \left( 1 + o\left(\frac{1}{n}\right) \right), \quad (9.47)$$

$$\left( h_n^{(1)} \right)'(z) = -\frac{n+1}{z} h_n^{(1)}(z) + h_{n-1}^{(1)}(z), \quad (9.47)$$

$$\left( h_n^{(1)} \right)'(z) = -(n+1) \frac{(2n-1)!!}{iz^{n+2}} \left( 1 + o\left(\frac{1}{n}\right) \right), \quad (9.48)$$

$$j_n(z) = n \frac{z^n}{(2n+1)!!} \left( 1 + o\left(\frac{1}{n}\right) \right), \quad (9.49)$$

$$(j_n)'(z) = n \frac{z^{n-1}}{(2n+1)!!} \left( 1 + o\left(\frac{1}{n}\right) \right), \quad (9.50)$$

where  $(2n-1)!! = 1 \cdot 3 \cdot 5 \cdots (2n-1)$ . We also have the Wronskian identity: (9.51)

$$h_n^{(1)}(z) \overline{h_n^{(1)'}(z)} - \overline{h_n^{(1)}(z)} h_n^{(1)'}(z) = -\frac{2i}{z^2}$$

Finally,

$$h_n^{(1)}(t) = o\left(\frac{2n}{et}\right)^n, \quad n \rightarrow \infty$$

uniformly for  $t$  on compact subsets of  $(0, \infty)$ .

Note that the asymptotic expansions show that

$$\frac{\partial}{\partial r} \left( h_n^{(1)}(\kappa r) \right) - i\kappa h_n^{(1)}(\kappa r) = o\left(\frac{1}{\kappa^2 r^2}\right).$$

Hence  $u(r) = h_n^{(1)}(\kappa r)$  satisfies the Sommerfeld radiation condition, (9.52)

$$\lim_{r \rightarrow \infty} r \left( \frac{\partial u}{\partial r} - i\kappa u \right) = 0.$$

As we have seen, this is the radiation condition for the Helmholtz equation corresponding to the Silver-Müller condition for Maxwell's equations.

We can now summarize our discussion of spherical harmonics and Bessel functions by the following theorem.

**Theorem 9.14** The function  $\tilde{v}_n^m(x) = j_n(\kappa|x|) Y_n^m(\hat{x})$  satisfies the Helmholtz equation  $\Delta v + \kappa^2 v = 0$  in all  $\mathbb{R}^3$ . The function  $v_n^m(x) = h_n^{(1)}(\kappa|x|) Y_n^m(\hat{x})$  satisfies the Helmholtz equation in  $\mathbb{R}^3 \setminus \{0\}$  and the Sommerfeld radiation condition (9.52).

Finally, we note two standard expansions. The first is for plane waves and is termed the Jacobi–Anger expansion:(9.53)

$$\exp(i\kappa\rho\cos\theta) = \sum_{n=0}^{\infty} i^n (2n+1) j_n(\kappa\rho) P_n(\cos\theta).$$

For a proof of this result, see [94]. The second is the Funk–Hecke formula(9.54)

$$\int_{\partial B_1} \exp(-i\kappa\rho\hat{x}\cdot\hat{z}) Y_n^m(\hat{z}) dA(\hat{z}) = \frac{4\pi}{i^n} (\kappa\rho) Y_n^m(\hat{x})$$

for  $\hat{x} \in \partial B_1$ ,  $\rho > 0$  and all  $n \geq 0$ ,  $-n \leq m \leq n$ .

### 9.3.3 Series solution of the exterior Maxwell problem

In this section we shall use the separation-of-variables solutions of the Helmholtz equation developed in the previous two sections to solve Maxwell's equations in the exterior of the ball of radius  $R$ , i.e. in  $\mathbb{R}^3 \setminus B_R$ . The material parallels the presentation in [94], but with the emphasis on Sobolev spaces.

In particular, we want to solve the problem of finding  $E^s$  such that(9.55a)

$$\begin{aligned} \nabla \times \nabla \times E^s - \kappa^2 E^s &= 0 \quad \text{in } \mathbb{R}^3 \setminus B_R, \\ (9.55b) \end{aligned}$$

$$\nabla \times E^s = \lambda \quad \text{on } \partial B_R,$$

(9.55c)

$$\rho(\nabla \times E^s \times \hat{x} - i\kappa E^s) \rightarrow 0 \quad \text{as } \rho \rightarrow \infty,$$

where  $\lambda$  is a suitable given tangential vector field on  $\partial B_R$ . Choosing  $\lambda = g$  provides the solution of (9.27a)–(9.27c).

We start by expanding the boundary data in terms of suitable vector basis functions on  $\partial B_R$ . Let  $Y_n^m(\hat{x})$ ,  $m = -n, \dots, n$ ,  $n = 0, 1, \dots$ , denote an orthonormal sequence of spherical harmonics on the unit sphere normalized so that

$$\int_{\partial B_1} Y_n^m(\hat{x}) \overline{Y^{m'n'}(\hat{x})} dA = \delta_{nn'} \delta_{m,m'}.$$

We have in mind the explicit spherical harmonics given in Theorem 9.11. The basis functions for tangential fields on  $\partial B_R$  are then the *vector spherical harmonics* of order  $n$  given by(9.56)

$$U_n^m = \frac{1}{\sqrt{n(n+1)}} \nabla_{\partial B_1} Y_n^m \quad \text{and} \quad V_n^m = \hat{x} \times U_n^m$$

for  $n = 1, 2, \dots$  and  $m = -n, \dots, n$ . Here, as usual,  $\nabla_{\partial B_1}$  denotes the surface gradient on the surface of the unit sphere  $\partial B_1$  (see Section 3.4). That these vector spherical harmonics are a good choice is shown next.

**Lemma 9.15** *The vector spherical harmonics defined in(9.56)are a complete orthonormal basis for  $L_t^2(\partial B_1)$ .*

**Proof** From the fact that  $\Delta_{\partial B_1} Y_n^m = -n(n+1)Y_n^m$ , we see, using integration by parts, that

$$\begin{aligned} \int_{\partial B_1} \nabla_{\partial B_1} Y_n^m \cdot \nabla_{\partial B_1} \overline{Y_{n'}^{m'}} dA &= n(n+1) \int_{\partial B_1} Y_n^m \overline{Y_{n'}^{m'}} dA \\ &= n(n+1) \delta_{n,n'} \delta_{m,m'}, \end{aligned}$$

so  $\{U_n^m\}$  is an orthonormal set in  $L_t^2(\partial B_1)$ . In addition, again using integration by parts,

$$\begin{aligned} \int_{\partial B_1} U_n^m V_{n'}^{m'} dA &= \frac{1}{\sqrt{n(n+1)n'(n'+1)}} \int_{\partial B_1} \nabla_{\partial B_1} Y_n^m \cdot \left( \vec{\nabla}_{\partial B_1} \times \overline{Y_{n'}^{m'}} \right) dA \\ &= 0, \end{aligned}$$

where we have used the fact that  $\nabla_{\partial B_1} \cdot (\vec{\nabla}_{\partial B_1} \times p) = 0$  for any sufficiently smooth function  $p$ .

Finally,

$$\begin{aligned} \int_{\partial B_1} V_{n'}^{m'} \overline{V_n^m} dA &= \int_{\partial B_1} \frac{1}{\sqrt{n(n+1)n'(n'+1)}} \longrightarrow \nabla \\ &= \int_{\partial B_1} \frac{1}{\sqrt{n(n+1)n'(n'+1)}} \nabla_{\partial B_1} \times \left( \vec{\nabla}_{\partial B_1} \times Y_n^m \right) \cdot \overline{Y_{n'}^{m'}} dA \\ &= 0, \end{aligned}$$

since  $\nabla_{\partial B_1} \times (\vec{\nabla}_{\partial B_1} \times Y_n^m) = \Delta_{\partial B_1} Y_n^m = -n(n+1)Y_n^m$ . Thus, we have verified that the vector spherical harmonics are orthonormal.

Now we want to show that if  $a \in L_t^2(\partial B_1)$  we can expand  $a$  using the orthonormal set in (9.56). We do this by proving a Helmholtz decomposition for  $a$ . First define  $\alpha \in H^1(\partial B_1)/\mathbb{R}$  as the solution of

$$\int_{\partial B_1} \nabla_{\partial B_1} \alpha \cdot \nabla_{\partial B_1} \bar{\xi} dA = \int_{\partial B_1} \alpha \cdot \nabla_{\partial B_1} \bar{\xi} dA \quad \text{forall } \xi \in H^1(\partial B_1) / \mathbb{R}.$$

This has a unique solution by the Lax–Milgram Lemma 2.21. We may expand  $\alpha$  as

$$\alpha(\hat{x}) = \sum_{n=1}^{\infty} \sum_{m=-n}^n a_{n,m} Y_n^m(\hat{x}),$$

where  $a_0^0$  since  $\alpha$  has vanishing average value on  $\partial B_1$ . Furthermore, since

$$\int_{\partial B_1} \nabla_{\partial B_1} \alpha \cdot \nabla_{\partial B_1} \bar{\alpha} dA = \sum_{n=1}^{\infty} \sum_{m=-n}^n n(n+1) |a_{n,m}|^2 < \infty$$

we may be sure that

$$\nabla_{\partial B_1} \alpha = \sum_{n=1}^{\infty} \sum_{m=-n}^n \alpha_{n,m} \nabla_{\partial B_1} Y_n^m .$$

Similarly, we define  $\beta \in H^1(\partial B_1)/R$  by

$$\int_{\partial B_1} \vec{\nabla}_{\partial B_1} \times \beta \cdot \vec{\nabla}_{\partial B_1} \times \bar{\xi} dA = \int_{\partial B_1} \alpha \cdot \vec{\nabla}_{\partial B_1} \times \bar{\xi} dA$$

for all  $\xi \in H^1(\partial B_1)/R$ . Again this has a unique solution by the Lax–Milgram lemma. We may write

$$\beta(\hat{x}) = \sum_{n=1}^{\infty} \sum_{m=-n}^n \beta_{n,m} Y_n^m(\hat{x})$$

and

$$(\vec{\nabla}_{\partial B_1} \times \beta)(\hat{x}) = \sum_{n=1}^{\infty} \sum_{m=-n}^n \beta_{n,m} \hat{x} \times \nabla_{\partial B_1} Y_n^m(\hat{x}).$$

Then, since  $\nabla_{\partial B_1} \times (\alpha - \vec{\nabla}_{\partial B_1} \times \beta - \nabla_{\partial B_1} \alpha) = 0$ , we see that  $\alpha - \vec{\nabla}_{\partial B_1} \times \beta - \nabla_{\partial B_1} \alpha = \nabla_{\partial B_1} \varphi$  for some  $\varphi \in H^1(\partial B_1)/R$  and, since

$$\nabla_{\partial B_1} \cdot (\alpha - \vec{\nabla}_{\partial B_1} \times \beta - \nabla_{\partial B_1} \alpha) = \nabla_{\partial B_1} \varphi = 0,$$

we see that  $\alpha - \vec{\nabla}_{\partial B_1} \times \beta - \nabla_{\partial B_1} \alpha = 0$ , which completes the proof.  $\square$

By the preceding lemma we can expand any function  $\lambda \in L_t^2(\partial B_R)$  by (9.57)

$$\lambda = \sum_{n=1}^{\infty} \sum_{m=-n}^n a_{n,m} U_n^m + b_{n,m} V_n^m,$$

and Parseval's theorem shows that we can define (this differs by a constant factor  $R^2$  from the definition using integrals, and is thus strictly just equivalent to the usual definition)

$$\|\lambda\|_{L_t^2(\partial B_R)}^2 = \sum_{n=1}^{\infty} \sum_{m=-n}^n (|a_{n,m}|^2 + |b_{n,m}|^2).$$

Furthermore, if  $H_t^1(\partial B_R) = L_t^2(\partial B_R) \cap (H^1(\partial B_R))^3$ , it follows that the norm on the space  $H_t^1(\partial B_R)$  can be characterized, equivalently, by

$$\|\lambda\|_{H_t^1(\partial B_R)}^2 = \sum_{n=1}^{\infty} \sum_{m=-n}^n n(n+1) (|a_{n,m}|^2 + |b_{n,m}|^2).$$

Still more generally, if

$$H_t^s(\partial B_R) = \left\{ u \in (H^s(\partial B_R))^3 \mid u \cdot v = 0 \text{ on } \partial B_R \right\}$$

then

$$\|\lambda\|_{H_t^s(\partial B_R)}^2 = \sum_{n=1}^{\infty} \sum_{m=-n}^n (n(n+1))^s (|a_{n,m}|^2 + |b_{n,m}|^2).$$

For our purposes, we shall find that the correct space for the tangential vector field  $\lambda$  is  $H^{-1/2}(\text{Div}; \partial B_R)$ , which is the completion of  $L_t^2(\partial B_R)$  in the norm

$$\|\lambda\|_{H^{-1/2}(\text{Div}; \partial B_R)}^2 = \|\lambda\|_{(H^{-1/2}(\partial B_R))^3}^2 + \|\nabla_{\partial B_R} \cdot \lambda\|_{H^{-1/2}(\partial B_R)}^2.$$

As we shall see, for a smooth boundary, in particular for  $\partial B_R$ , we have

$$Y(\partial B_R) = H^{-1/2}(\text{Div}; \partial B_R)$$

where  $Y(\partial B_R)$  is the trace space defined in (3.50) and

$$Y(\partial B_R)' = H(\text{Curl}; \partial B_R).$$

If  $\lambda = \sum_{n=1}^{\infty} \sum_{m=-n}^n a_{n,m} U_n^m + b_{n,m} V_n^m$  and

$$\begin{aligned} \nabla_{\partial B_R} \cdot \lambda &= \frac{1}{R} \sum_{n=1}^{\infty} \sum_{m=-n}^n \frac{a_{n,m}}{\sqrt{n(n+1)}} \Delta_{\partial B_R} Y_n^m \\ &= \frac{1}{R} \sum_{n=1}^{\infty} \sum_{m=-n}^n a_{n,m} \sqrt{n(n+1)} Y_n^m, \end{aligned}$$

we may equivalently express this norm (again ignoring factors of  $R$ ) as (9.58)

$$\|\lambda\|_{H^{-1/2}(\text{Div}; \partial B_R)}^2 = \sum_{n=1}^{\infty} \sum_{m=-n}^n \sqrt{n(n+1)} |a_{n,m}|^2 + \frac{1}{\sqrt{n(n+1)}} |b_{n,m}|^2.$$

Notice we use  $\text{Div}$  in place of  $\text{div}$  to indicate a surface divergence. Although it will turn out that suitable traces of functions in  $H(\text{curl}; B_R)$  lie in  $H^{1/2}(\text{Div}; \partial B_R)$ , we can go further and define  $H^s(\text{Div}; \partial B_R)$  for any  $s$  as follows: (9.59)

$$\begin{aligned} H^s(\text{Div}, \partial B_R) &= \left\{ u \in (H^s(\partial B_R))^3 \mid u \cdot v = 0 \text{ and} \right. \\ &\quad \left. \nabla_{\partial B_R} \cdot u \in H^s(\partial B_R) \right\} \end{aligned}$$

with norm

$$\|\lambda\|_{H^s(\text{Div}; \partial B_R)}^2 = \|\lambda\|_{(H^s(\partial B_R))^3}^2 + \|\nabla_{\partial B_R} \cdot \lambda\|_{H^s(\partial B_R)}^2$$

or, equivalently, if  $\lambda$  is expanded as in (9.57),

$$\|\lambda\|_{H^s(\text{Div}, \partial B_R)}^2 = \sum_{n=1}^{\infty} \sum_{m=-n}^n \left\{ (n(n+1))^{s+1} |a_{n,m}|^2 + (n(n+1))^s |b_{n,m}|^2 \right\}.$$

Now that we have an explicit representation for  $\lambda$ , as well as for norms on  $\partial B_R$ , we need to develop a corresponding series solution for the radiating solutions of Maxwell's equations (i.e. those satisfying (9.55a) and (9.55c)). Motivated by the discussion of Debye potentials in Section 9.3, we define the *vector wave functions* (9.60)

$$M_n^m = \nabla \times \left\{ x h_n^{(1)}(\kappa|x|) Y_n^m(\hat{x}) \right\} \quad \text{and} \quad N_n^m = \frac{1}{i\kappa} \nabla \times M_n^m,$$

for  $n = 1, 2, \dots$  and  $m = -n, \dots, n$ , where  $h_n^{(1)}$  is the spherical Hankel function of first kind and order  $n$  presented in Section 9.3.2 (note that this notation *differs* from that in [94]).

**Theorem 9.16** *The functions  $N_n^m$  and  $M_n^m$  defined in (9.60) are radiating solutions of Maxwell's equations in  $\mathbb{R}^3 \setminus \{0\}$  (i.e. they satisfy (9.27a) except at  $x = 0$  and (9.27c)).*

**Proof** The argument in the introduction of this chapter, together with the fact that  $N_n^m$  satisfies the Helmholtz equation in  $\mathbb{R}^3 \setminus \{0\}$  shows that  $h_n^{(1)}(\kappa|x|) Y_n^m(\hat{x})$  satisfies Maxwell's equations in  $\mathbb{R}^3 \setminus \{0\}$ . Taking the curl of Maxwell's equations shows that  $M_n^m$  is also a solution.

Using (B.4), we may compute  $M_n^m = h_n^{(1)}(\kappa|x|) \nabla_{\partial B_1} Y_n^m(\hat{x}) \times \hat{x}$ , so

$$(\nabla \times M_n^m) \times \hat{x} = \left\{ \frac{1}{|\kappa|} h_n^{(1)}(\kappa|x|) + \kappa h_n^{(1)'}(\kappa|x|) \right\} \nabla_{\partial B_1} Y_n^m(\hat{x}) \times \hat{x}.$$

Hence (9.61)

$$\begin{aligned} (\nabla \times M_n^m) \times \hat{x} - i\kappa M_n^m &= \left( \frac{1}{|\kappa|} h_n^{(1)}(\kappa|x|) \right. \\ &\quad \left. + \kappa h_n^{(1)'}(\kappa|x|) - i\kappa h_n^{(1)}(\kappa|x|) \right) \nabla_{\partial B_1} Y_n^m(\hat{x}) \times \hat{x}. \end{aligned}$$

The Silver–Müller radiation condition follows from the decay of  $M_n^m = h_n^{(1)}(\kappa|x|) \nabla_{\partial B_1} Y_n^m(\hat{x}) \times \hat{x}$  for large  $|x|$  and the fact that  $h_n^{(1)}(\kappa|x|)$  satisfies the Sommerfeld radiation condition (see Theorem 9.14 and the discussion in Section 9.3.2). A similar computation verifies this for  $h_n^{(1)}(\kappa|x|)$  by taking the curl of (9.61).  $\square$

Before we state and prove the main theorem of this section, we need to recall another well-known result for vector spherical harmonics. Let us define (9.62)

$$\widetilde{M}_n^m = \nabla \times \left\{ x j_n(\kappa|x|) Y_n^m(\hat{x}) \right\} \quad \text{and} \quad \widetilde{N}_n^m = \frac{1}{i\kappa} \nabla \times \widetilde{M}_n^m.$$

These are the *interior vector spherical harmonics* used to expand a solution of Maxwell's equations inside the sphere. We also need suitable solutions of the Helmholtz equation as defined in Theorem 9.14. Now recalling that  $\Phi$  is the fundamental solution of the Helmholtz equation (see (9.1)), and letting  $p$  be any fixed vector, we have the following formula, usually referred to as the *vector addition theorem* (9.63)

$$\begin{aligned}\Phi(x, y)p = & \sum_{n=1}^{\infty} \frac{i\kappa}{n(n+1)} \sum_{m=-n}^n N_n^m(x) \overline{\widetilde{N}_n^m(y)} \cdot p \\ & - \sum_{n=1}^{\infty} \frac{i\kappa}{n(n+1)} \sum_{m=-n}^n M_n^m(x) \overline{\widetilde{M}_n^m(y)} \cdot p \\ & + \frac{i}{\kappa} \sum_{n=1}^{\infty} \sum_{m=-n}^n \nabla V_n^m(x) \overline{\widetilde{V}_n^m(y)} \cdot p\end{aligned}$$

convergent either with respect to  $y$  for fixed  $x$  or with respect to  $x$  for fixed  $y$  provided  $|x| > |y|$ . In addition this series and its term by term derivatives in  $x$  or  $y$  are uniformly and absolutely convergent on compact subsets of  $|x| > |y|$  [194].

**Theorem 9.17** Let  $E^s$  be a radiating solution of Maxwell's equations for  $|x| > R > 0$ . Then  $E^s$  has the representation (9.64)

$$E^s(x) = \sum_{n=1}^{\infty} \sum_{m=-n}^n \{ \alpha_{n,m} M_n^m(x) + \beta_{n,m} N_n^m(x) \} .$$

The series converges uniformly (together with its derivatives) on compact subsets of  $|x| > R$ . Conversely, if the tangential component of the series converges in  $L_t^2(\partial B_R)$  then the series converges uniformly on compact subsets of  $\mathbb{R}^3 \setminus B_R$  and represents a radiating solution of Maxwell's equations.

**Remark 9.18** The corresponding series for  $H^s = (1/i\kappa)^\nabla \times E^s$  is

$$H^s(x) = \sum_{n=1}^{\infty} \sum_{m=-n}^n \{ \alpha_{n,m} N_n^m(x) - \beta_{n,m} M_n^m(x) \} .$$

There is also an analogue of this theorem for interior problems. In this case, for any Lipschitz domain  $D$ , a solution of the interior Maxwell system can be represented on a ball  $B$  contained in  $D$  by (9.64) with  $M_n^m$  replaced by  $\widetilde{M}_n^m$  and  $N_n^m$  replaced by  $\widetilde{N}_n^m$ . Convergence is uniform on compact subsets of  $B$ . This expansion is useful for representing incident electromagnetic fields in the neighborhood of the scatterer (see Section 9.5.2) or for computing a series solution of scattering from a dielectric sphere [284].

**Proof of Theorem 9.17** This proof follows Kress [194]. Suppose  $E^s$  is a radiating solution of Maxwell's equations, and let  $H^s = (1/i\kappa)^\nabla \times E^s$ . Using these functions, and substituting the expansion (9.63) for  $\Phi$  in the Stratton-Chu formula (9.15) proves the expansion.

The converse of the theorem is proved by showing that the convergence of the series  $L_t^2(\partial B_R)$  implies the uniform convergence of tangential components of the series on any sphere of strictly larger radius. Applying the first part of the theorem to the solution outside the sphere of radius  $R_i > R$  proves the result (see the proof of Theorem 6.25 in [94] for more details).  $\square$

Now let us suppose the boundary data  $\lambda \in H^{1/2}(\text{Div}; \partial B_R)$  has the representation (9.57). We want to compute the scattered field satisfying (9.55) in

terms of the coefficients of this expansion. Theorem 9.17 shows that any radiating solution of Maxwell's equations can be written (for  $|x| > R$ ) in the form(9.65)

$$E^s(x) = \sum_{n=1}^{\infty} \sum_{m=-n}^n \{ \alpha_{n,m} M_n^m(x) - \beta_{n,m} N_n^m(x) \},$$

with uniform convergence on compact subsets of  $|x| > R$ . The corresponding magnetic field  $H^s$  is given by

$$H^s = \frac{1}{ik} \nabla \times E^s = \sum_{n=1}^{\infty} \sum_{m=-n}^n \{ \alpha_{n,m} N_n^m(x) - \beta_{n,m} M_n^m(x) \}.$$

We need to express  $\hat{x} \times E^s$  and  $\hat{x} \times H^s$  on  $|x| = R$  in terms of the coefficients of the expansion for  $E^s$ . Using (9.65), we have

$$\hat{x} \times E^s = \sum_{n=1}^{\infty} \sum_{m=-n}^n \{ \alpha_{n,m} \hat{x} \times M_n^m(x) + \beta_{n,m} \hat{x} \times N_n^m(x) \},$$

and using the definition of  $M_n^m$  in (9.60) and the fact that for a suitably smooth vector function  $u$  and scalar function  $\varphi$  the identity  $\nabla \times (\varphi u) = (\nabla \varphi) \times u + \varphi \nabla \times u$  holds, we obtain

$$\begin{aligned} \hat{x} \times M_n^m(x) &= \hat{x} \times \left( \nabla \left\{ h_n^{(1)}(k|x|) Y_n^m(\hat{x}) \right\} \times x \right) \\ &= \nabla_{\partial B_1} \left\{ h_n^{(1)}(k|x|) Y_n^m(\hat{x}) \right\}. \end{aligned}$$

It follows from the definition of  $U_n^m$  in (9.56) that(9.66)

$$\hat{x} \times M_n^m(x) = h_n^{(1)}(kR) \sqrt{n(n+1)} U_n^m(\hat{x}) \quad 0|x|=R.$$

Next we want to compute  $\hat{x} \times N_n^m$  on  $|x|=R$ . Using the vector identity (B.8), the definition of  $N_n^m$ , the fact that  $\hat{x}$  and  $x$  are parallel, and (3.14), we obtain

$$\begin{aligned} \hat{x} \times N_n^m(x) &= \frac{1}{ik} \hat{x} \times \nabla \left\{ h_n^{(1)}(k|x|) Y_n^m(\hat{x}) + |x| \frac{\partial}{\partial r} \left[ \left( h_n^{(1)}(k|x|) \right) Y_n^m(\hat{x}) \right] \right\} \\ &= \frac{1}{ikR} \left\{ h_n^{(1)}(k|x|) + |x| \frac{\partial}{\partial r} h_n^{(1)}(k|x|) \right\} (\hat{x} \times \nabla_{\partial B_R} Y_n^m(\hat{x})). \end{aligned}$$

Using the definition of  $V_n^m$ , we have shown that(9.67)

$$\hat{x} \times N_n^m(x) = \frac{1}{ikR} \left\{ h_n^{(1)}(k|x|) + |x| \frac{\partial}{\partial r} h_n^{(1)}(k|x|) \right\} \sqrt{n(n+1)} V_n^m(\hat{x}).$$

Using this equality and (9.66), shows that, on  $|x|=R$ ,(9.68)

$$\begin{aligned} \hat{x} \times E^s &= \sum_{n=1}^{\infty} \sum_{m=-n}^n \alpha_{n,m} h_n^{(1)}(kR) \sqrt{n(n+1)} U_n^m \\ &\quad + \frac{1}{ikR} \sum_{n=1}^{\infty} \sum_{m=-n}^n \beta_{n,m} \left\{ h_n^{(1)}(kR) + \right. \\ &\quad \left. kR \left( h_n^{(1)} \right)'(kR) \right\} \sqrt{n(n+1)} V_n^m. \end{aligned}$$

We also need to obtain the series for  $\hat{x} \times H^s$  on  $|x| = R$ . But  $H^s$  has the same form as  $E^s$ , where now  $\alpha_{n,m}$  plays the role of  $\beta_{n,m}$  and  $-\beta_{n,m}$  the role of  $\alpha_{n,m}$ . This yields(9.69)

$$\begin{aligned}\hat{x} \times H^s &= \frac{1}{ikR} \sum_{n=1}^{\infty} \sum_{m=-n}^n \alpha_{n,m} \left\{ h_n^{(1)}(kR) + kR \left( h_n^{(1)} \right)'(kR) \right\} \sqrt{n(n+1)} V_n^m \\ &\quad - \sum_{n=1}^{\infty} \sum_{m=-n}^n \beta_{n,m} h_n^{(1)}(kR) \sqrt{n(n+1)} U_n^m.\end{aligned}$$

Now we can solve the boundary value problem (9.55) for arbitrary tangential boundary data  $\lambda$ . For given  $\lambda \in H^{1/2}(\text{Div}; \partial B_R)$ , let  $(E^s, H^s)$  satisfy(9.70a)

$$ikE^s + \nabla \times H^s = 0 \quad \text{in } \mathbb{R}^3 \setminus \bar{B}_R,$$

$$ikH^s - \nabla \times E^s = 0 \quad \text{in } \mathbb{R}^3 \setminus \bar{B}_R, \tag{9.70b}$$

$$\hat{x} \times E^s = \lambda \quad \text{on } \partial B_R, \tag{9.70c}$$

$$\lim_{\rho \rightarrow \infty} \rho (H^s \times \hat{x} - E^s) = 0. \tag{9.70d}$$

Equating terms in (9.57) and (9.68), we obtain a series for each field which converges in  $H_{\text{loc}}(\text{curl}; \mathbb{R}^3 \setminus \bar{B}_R)$  as follows.

**Lemma 9.19** For  $\lambda \in H^{1/2}(\text{Div}; \partial B_R)$  given by (9.57), the unique solution  $E^s, H^s \in H_{\text{loc}}(\text{curl}; \mathbb{R}^3 \setminus \bar{B}_R)$  of (9.70) is given by

$$\begin{aligned}E^s &= \sum_{n=1}^{\infty} \sum_{m=-n}^n \left[ \frac{a_{n,m} M_n^m}{h_n^{(1)}(kR) \sqrt{n(n+1)}} \right. \\ &\quad \left. + \frac{i k R b_{n,m} N_n^m}{\left[ h_n^{(1)}(kR) + kR \left( h_n^{(1)} \right)'(kR) \right] \sqrt{n(n+1)}} \right], \\ H^s &= \sum_{n=1}^{\infty} \sum_{m=-n}^n \left[ \frac{a_{n,m} N_n^m}{h_n^{(1)}(kR) \sqrt{n(n+1)}} \right. \\ &\quad \left. - \frac{i k R b_{n,m} M_n^m}{\left[ h_n^{(1)}(kR) + kR \left( h_n^{(1)} \right)'(kR) \right] \sqrt{n(n+1)}} \right].\end{aligned}$$

## 9.4 Electromagnetic Calderon operators

We now wish to define analogs of the famous Dirichlet-to-Neumann (DtN) map for Maxwell's equations. These maps are referred to as Calderon operators by Cessenat [73] and as boundary component maps (either electric-to-magnetic, or magnetic-to-electric) by Colton and Kress [94].

### 9.4.1 The electric-to-magnetic Calderon operator

The electric-to-magnetic Calderon operator takes electric field boundary data to magnetic field boundary data. In particular, for a given tangential vector field  $\lambda$  on  $\partial B_R$  we define  $G_e \lambda = \hat{x} \times H$ , where  $E$  and  $H$  satisfy (9.70). We can use (9.69) to obtain an explicit representation for the map  $G_e$  from  $\lambda$  to  $\hat{x} \times H$ . For  $\lambda \in H^{1/2}(\text{Div}; \partial B_R)$  given by (9.57) and using (9.68), (9.69) and Lemma 9.19, we have (9.71)

$$G_e \lambda = \hat{x} \times H^s = \sum_{n=1}^{\infty} \sum_{m=-n}^n \left\{ -i\kappa R \frac{b_{n,m}}{\delta_n} U_n^m + \frac{a_{n,m} \delta_n}{i\kappa R} V_n^m \right\},$$

where (9.72)

$$\delta_n = \kappa R \frac{h_n^{(1)}(z)(\kappa R)}{h_n^{(1)}(z)} + 1.$$

We now proceed to analyze  $G_e$  in a fashion similar to that followed by Masmoudi [213] in his analysis of series solutions of the Helmholtz equation. First we prove bounds for the coefficients  $\delta_n$  defined in (9.72) which appear in the expansion (9.71) of  $G_e \lambda$ .

**Lemma 9.20** *There exist positive constants  $c_1$  and  $c_2$  such that, for all  $n$ ,*

$$c_1 n \leq |\delta_n| \leq c_2 n.$$

**Proof** This lemma is proved in [94]. Set  $z = \kappa R$ . From the recurrence relation (9.43) with  $f_n = h_n^{(1)}$ , we see that

$$\delta_n + n = \frac{z h_n^{(1)}(z)}{h_n^{(1)}(z)} + n + 1 = \frac{z h_{n-1}^{(1)}}{h_n^{(1)}(z)}.$$

Now we use the formula for the asymptotic behavior of the Hankel function, given by (9.46) for large values of  $n$ , and obtain

$$\begin{aligned} \delta_n + n &= \frac{z h_{n-1}^{(1)}(z)}{h_n^{(1)}(z)} = \frac{z(1 \cdot 3 \cdots (2n-3)/iz^n)(1+o(\frac{1}{n}))}{(1 \cdot 3 \cdots (2n-1)/iz^{n+1})(1+o(\frac{1}{n}))} \\ &= \frac{z^2}{2n-1} \left( 1 + o\left(\frac{1}{n}\right) \right). \end{aligned}$$

For small  $n$  we note that  $h_n^{(1)}(z)$  and  $z(h_n^{(1)})'(z) + h_n^{(1)}(z)$  have no real roots. This completes the proof.  $\square$

Our next result shows that  $G_e$  is continuous as a map

$$G_e : H^{-1/2}(\text{Div}; \partial B_R) \rightarrow H^{-1/2}(\text{Div}; \partial B_R).$$

**Theorem 9.21** There exists a constant  $C$  such that the following inequality holds:

$$\|G_e \lambda\|_{H^{-1/2}(\text{Div}; \partial B_R)} \leq C \|\lambda\|_{H^{-1/2}(\text{Div}; \partial B_R)} \quad \text{for all } \lambda \in H^{-1/2}(\text{Div}; \partial B_R).$$

**Proof** By (9.58), (9.71) and Lemma 9.20,

$$\begin{aligned} \|G_e \lambda\|_{H^{-1/2}(\text{Div}; \partial B_R)}^2 &= \sum_{n=1}^{\infty} \sum_{m=-n}^n \left\{ \sqrt{1+n(n+1)} \frac{|\kappa R b_{n,m}|^2}{|\delta_n|^2} \right. \\ &\quad \left. + \frac{1}{\sqrt{1+n(n+1)}} \frac{|b_{n,m} \delta_n|^2}{\kappa^2 R^2} \right\} \\ &= \sum_{n=1}^{\infty} \sum_{m=-n}^n \left\{ \frac{1}{\sqrt{1+n(n+1)}} \frac{1+n(n+1)}{|\delta_n|^2} |\kappa R b_{n,m}|^2 \right. \\ &\quad \left. + \frac{|\delta_n|^2}{1+n(n+1)} \frac{\sqrt{1+n(n+1)} |b_{n,m}|^2}{\kappa^2 R^2} \right\} \\ &\leq C \sum_{n=1}^{\infty} \sum_{m=-n}^n \left\{ \frac{1}{\sqrt{1+n(n+1)}} |b_{n,m}|^2 \right. \\ &\quad \left. + \sqrt{1+n(n+1)} |a_{n,m}|^2 \right\}, \end{aligned}$$

where  $C$  depends on  $\kappa R$ ,  $c_1$  and  $c_2$ , from Lemma 9.20.  $\square$

Next we wish to analyze the operator  $G_e$  for a purely imaginary wavenumber. We start by analyzing  $\delta_n$  for  $\kappa = i$ .

**Lemma 9.22** Let (9.73)

$$\tilde{\delta}_n = iR \frac{\left( h_n^{(1)} \right)'(iR)}{h_n^{(1)}(iR)} + 1.$$

Then  $\tilde{\delta}_n$  is real and strictly negative for all  $n$ .

**Proof** In the proof of Lemma 9.20, we have derived the representation (now for  $\kappa = i$ )

$$\tilde{\delta}_n + n = \frac{iR h_{n-1}^{(1)}(iR)}{h_n^{(1)}(iR)}.$$

But  $h_n^{(1)}(iR) = j_n(iR) + iy_n(iR)$  and using the expansions for  $j_n$  and  $y_n$  in (9.39) and (9.40), respectively, we see that  $i^n h_n^{(1)}(iR) \in \mathbb{R}$ . But  $i^n h_n^{(1)}(|x|) Y_n^m(x)$  is a non-trivial solution of the Helmholtz equation  $\Delta u - u = 0$  in  $\mathbb{R}^3 \setminus B_R$  with appropriate decay at infinity and hence can have no real zeros as a function of  $r$ . Thus  $i^n h_n^{(1)}(iR)$  has one sign as a function of  $R$  for each  $n$ . But then

$$\text{sign} \left( \frac{iR h_{n-1}^{(1)}(iR)}{h_n^{(1)}(iR)} \right) = \lim_{r \rightarrow \infty} \text{sign} \left( \frac{ir h_{n-1}^{(1)}(ir)}{h_n^{(1)}(ir)} \right).$$

From the large- $r$ , asymptotics of the Hankel function given by (9.44) we obtain

$$\frac{ih_{n-1}^{(1)}(ir)}{h_n^{(1)}(ir)} = ir\exp(i\pi/2)\left(1 + o\left(\frac{1}{r}\right)\right) = -r\left(1 + o\left(\frac{1}{r}\right)\right).$$

Thus,  $\delta_n^+ < 0$  for each  $n$ , as claimed.  $\square$

Now we analyze the operator  $G_e$  for purely imaginary wavenumbers. Let

$$\tilde{G}_e : H^{-1/2}(\text{Div}; \partial B_R) \rightarrow H^{-1/2}(\text{Div}; \partial B_R)$$

be defined by (9.71) with  $\kappa = i$ , so that if  $\lambda$  is given by (9.57) then (9.74)

$$\tilde{G}_e \lambda = \sum_{n=1}^{\infty} \sum_{m=-n}^n \left\{ R \frac{b_{n,m}}{\tilde{\delta}_n} U_n^m - \frac{a_{n,m} \tilde{\delta}_n}{R} V_n^m \right\},$$

where  $\delta_n^+$  is given by (9.73).

**Lemma 9.23** *The operator  $\tilde{G}_e$  is negative definite in the sense that*

$$\langle \tilde{G}_e \lambda, \lambda \times \hat{x} \rangle < 0$$

for any  $\lambda \in H^{1/2}(\text{Div}; \partial B_R)$  with  $\lambda \neq 0$ . Furthermore,

$$|\langle \tilde{G}_e \lambda, \lambda \times \hat{x} \rangle| \geq c \|\lambda\|_{H^{-1/2}(\text{Div}; \partial B_R)}^2 \quad \text{for all } \lambda \in H^{-1/2}(\text{Div}; \partial B_R).$$

**Proof** Recall that if  $\lambda$  is given by (9.57) then

$$\begin{aligned} \lambda \times \hat{x} &= \sum_{n=1}^{\infty} \sum_{m=-n}^n \{a_{n,m} U_n^m \times \hat{x} + b_{n,m} V_n^m \times \hat{x}\} \\ &= \sum_{n=1}^{\infty} \sum_{m=-n}^n \{-a_{n,m} V_n^m + b_{n,m} U_n^m\}. \end{aligned}$$

Thus,

$$\langle \tilde{G}_e \lambda, \lambda \times \hat{x} \rangle = \sum_{n=1}^{\infty} \sum_{m=-n}^n \left\{ R \frac{|b_{n,m}|^2}{\tilde{\delta}_n} + \frac{|a_{n,m}|^2 \tilde{\delta}_n}{R} \right\} < 0.$$

Now, using the fact that  $\delta_n^+ = -n + O(1/n)$  and using the characterization of the norm on  $H^{1/2}(\text{Div}; \partial B_R)$  given in (9.58) we can prove the coercivity estimate of the lemma.  $\square$

Our final lemma of this section shows that a suitable combination of  $G_e$  and  $\tilde{G}_e$  is compact on a suitable set of functions on  $\partial B_R$ . Let

$$\begin{aligned} H_{\text{Div}}^{-1/2}(\text{Div}; \partial B_R) &= \left\{ \lambda = \sum_{n=1}^{\infty} \sum_{m=-n}^n b_{n,m} V_n^m \mid \sum_{n=1}^{\infty} \sum_{m=-n}^n \frac{1}{\sqrt{1+n(n+1)}} |b_{n,m}|^2 < \infty \right\}. \end{aligned}$$

**Lemma 9.24** Then the following operator is well defined and bounded:

$$G_e + i\kappa \tilde{G}_e \Big|_{H_{\text{Div}}^{-1/2}(\text{Div}; \partial B_R)} : H_{\text{Div}}^{-1/2}(\text{Div}; \partial B_R) \rightarrow H_t^{3/2}(\partial B_R),$$

where  $H_t^{3/2}(\partial B_R) = \left\{ u \in (H^{3/2}(\partial B_R))^3 \mid u \cdot v \right\}$ . Hence

$$G_e + i\kappa \tilde{G}_e : H_{\text{Div}}^{-1/2}(\text{Div}; \partial B_R) \rightarrow H^{-1/2}(\text{Div}; \partial B_R)$$

is compact.

**Proof** From the series expansion of  $G_e$  and  $\tilde{G}_e$  we know that if

$$\lambda = \sum_{n=1}^{\infty} \sum_{m=-1}^n \{a_{n,m} V_n^m + b_{n,m} U_n^m\}$$

then

$$(G_e + i\kappa \tilde{G}_e)\lambda = \sum_{n=1}^{\infty} \sum_{m=-n}^n \left[ \left( \frac{\delta_n}{i\kappa} - i\kappa \tilde{\delta}_n \right) \frac{a_{n,m}}{R} V_n^m + i\kappa R \left( \frac{1}{\tilde{\delta}_n} - \frac{1}{\delta_n} \right) b_{n,m} U_n^m \right].$$

From the asymptotics of  $\delta_n$  and  $\tilde{\delta}_n$  we obtain

$$\frac{1}{\delta_n} - \frac{1}{\tilde{\delta}_n} = o\left(1/n^2\right)$$

(in fact  $\delta_n^{-1} - \tilde{\delta}_n^{-1} = o(1/n^3)$  but we do not need this improved estimate).

If  $\lambda \in H_{\text{Div}}^{-1/2}(\text{Div}; \partial B_R)$ , then  $a_{n,m} = 0$  for all  $n, m$  and

$$\begin{aligned} & \| (G_e + i\kappa \tilde{G}_e)\lambda \|_{H_t^{3/2}(\partial B_R)}^2 \\ &= \kappa^2 R^2 \sum_{n=1}^{\infty} \sum_{m=-n}^n \left| \frac{1}{\delta_n} - \frac{1}{\tilde{\delta}_n} \right|^2 |b_{n,m}|^2 (1 + n(n+1))^{3/2} \\ &= \kappa^2 R^2 \sum_{n=1}^{\infty} \sum_{m=-n}^n \frac{1}{\sqrt{1+n(n+1)}} |b_{n,m}|^2 \left[ \left| \frac{1}{\delta_n} - \frac{1}{\tilde{\delta}_n} \right|^2 (1 + n(n+1))^2 \right] \\ &\leq c \|\lambda\|_{H_t^{3/2}(\partial B_R)}^2. \end{aligned}$$

This ends the proof.  $\square$

## 9.4.2 The magnetic-to-electric Calderon operator

The magnetic-to-electric Calderon operator is the analogue of the Neumann-to-Dirichlet (NtD) map. The properties of the exterior operator of this type follow

directly from Theorem 9.21, given the symmetry between electric and magnetic fields for Maxwell's equations. However, we provide a direct proof since we need some slightly different properties when we apply this operator. We also analyze the interior magnetic-to-electric Calderon operator.

Proceeding formally, suppose that  $\lambda \in H^{1/2}(\text{Div}; \partial B_R)$  is a tangential vector field on  $\partial B_R$ , then we define the exterior Calderon operator  $g_e$  by(9.75)

$$\mathcal{G}_e \lambda = \hat{x} \times u|_{\partial B_R},$$

where  $u$  is the solution of(9.76a)

$$\nabla \times \nabla \times u - \kappa^2 u = 0 \text{ in } \mathbb{R}^3 \setminus \overline{B_R},$$

$$\hat{x} \times \frac{1}{i\kappa} \nabla \times u = \lambda \text{ on } \partial B_R, \quad (9.76b)$$

$$\lim_{\rho \rightarrow \infty} \rho((\nabla \times u) \times \hat{x} - i\kappa u) = 0. \quad (9.76c)$$

The interior Calderon operator is defined in a similar way by(9.77)

$$\mathcal{G}_i \lambda = \hat{x} \times w|_{\partial B_R}$$

where  $w$  is the solution of(9.78a)

$$\nabla \times \mu_r^{-1} \nabla \times w - \kappa^2 \in {}_r w = 0 \text{ in } B_R,$$

$$\hat{x} \times \frac{1}{i\kappa} \nabla \times w = \lambda \text{ on } \partial B_R. \quad (9.78b)$$

We have the following lemma summarizing the basic mapping properties of  $g$ :

**Lemma 9.25** If  $\lambda \in H(\text{Div}; \partial B_R)$ ,  $s \in \mathbb{R}$ , is given by(9.57)then the solution  $u$  of (9.76a)–(9.76c) is given by(9.79)

$$u = \sum_{n=1}^{\infty} \sum_{m=-n}^n \frac{i\kappa R b_n^m}{h_n^{(1)}(\kappa R) + \kappa R (h_n^{(1)})'(\kappa R)} \frac{M_n^m}{\sqrt{n(n+1)}} - \sum_{n=1}^{\infty} \sum_{m=-n}^n \frac{a_n^m}{h_n^{(1)}(\kappa R)} \frac{N_n^m}{\sqrt{n(n+1)}},$$

and the exterior Calderon operator has the representation(9.80)

$$G_e \lambda = \sum_{n=1}^{\infty} \sum_{m=-n}^n \left[ \frac{b_n^m}{\delta_n} U_n^m - \delta_n a_n^m V_n^m \right],$$

where

$$\delta_n = \frac{1}{ikR} \left( 1 + \kappa R \frac{h_n^{(1)'}(\kappa R)}{h_n^{(1)}(\kappa R)} \right).$$

&gt;

In particular,  $G_\epsilon : H^s(\text{Div}; \partial B_R) \rightarrow H^s(\text{Div}; \partial B_R)$  and  $g_\epsilon$  is invertible.

**Proof** The representation of  $u$  in (9.79) in terms of  $M_n^m$  and  $N_n^m$  is proved in Theorem 9.19. The representation of  $u$  and  $g\lambda$  in terms of the coefficients of  $\lambda$  then follows using the basis functions in (9.66) and (9.67). Finally, the mapping properties follow from the definition of the norm on  $H^s(\text{Div}; \partial B_R)$  and the estimate for  $\delta_n$  from Lemma 9.20.  $\square$

Now let  $\tilde{g}_i$  denote the interior Calderon operator in the case when  $\epsilon_r = \mu_r = 1$ . We have the following lemma:

**Lemma 9.26** Let

$$\hat{\delta}_n = \frac{1}{ikR} \left( 1 + \kappa R \frac{(j_n)'(\kappa R)}{j_n(\kappa R)} \right),$$

and suppose  $R$  is chosen so that  $0 < |\hat{\delta}_n| < \infty$ . When  $\epsilon_r = \mu_r = 1$  we have

$$\tilde{g}_i \lambda = \sum_{n=1}^{\infty} \sum_{m=-n}^n \left[ \frac{b_n^m}{\hat{\delta}_n} U_n^m - \hat{\delta}_n a_n^m V_n^m \right].$$

The operator  $\tilde{g}_i : H^s(\text{Div}; \partial B_R) \rightarrow H^s(\text{Div}; \partial B_R)$  is bounded linear operator for any  $s$ .

**Remark 9.27** The condition that  $0 < |\hat{\delta}_n| < \infty$  implies a restriction on  $R$ , which can be checked a priori from a knowledge of the spherical Bessel functions.

**Proof of Lemma 9.26** The proof is the same as for Lemma 9.25 once the conditions on  $\hat{\delta}_n$  are satisfied.  $\square$

## 9.5 Scattering of a plane wave by a sphere

In this section we shall derive a series solution for a particular scattering problem to illustrate the techniques used in the preceding sections of this chapter. In particular, we shall solve the problem of scattering of a plane wave from a perfectly conducting sphere of radius  $R$ . The equations are given by (9.27a)–(9.27c), with (9.81)

$$g = -\hat{x} \times E^i \text{ on } \partial B_R, \text{ where } E^i = e_1 \exp(i\kappa x_3).$$

In the first part of this section we shall prove that this scattering problem has at most one solution by proving a uniqueness theorem for scattering by a sphere. This will be used in later sections to prove uniqueness of solutions of general scattering problems. Then, in Section 9.5.2, we will actually derive the famous Meier series solution for this problem. For us, it is valuable as a test problem for finite element scattering codes.

### 9.5.1 Uniqueness and Rellich's lemma

Our first result, termed Rellich's lemma, is a tool used to prove that scattering problems involving a bounded scatterer have a unique solution.

**Lemma 9.28** (Rellich's lemma) Suppose that  $E^s$  is a solution of Maxwell's equations (9.27a) in the exterior of a sphere of radius  $R$  subject to the radiation condition (9.27c). Let  $H^s = (1/ik)^\nabla \times E^s$ . If

$$\Re \left( \int_{\partial B_\rho} (\hat{x} \times E^s) \cdot \bar{H^s} dA \right) \leq 0$$

for all  $\rho > R$ , then  $E^s = H^s = 0$  in  $\mathbb{R}^3 \setminus B_R$ .

**Proof** Using the series for  $\hat{x} \times E^s$  in (9.68) and computing  $(\hat{x} \times H^s) \times \hat{x}$  using (9.69) and (9.56) on a sphere of radius  $\rho$ , we obtain

$$\begin{aligned} & \int_{\partial B_\rho} (\hat{x} \times E^s) \cdot \bar{H^s} dA \\ &= \rho^2 \sum_{n=1}^{\infty} \sum_{m=-n}^n \left[ |\alpha_{n,m}|^2 h_n^{(1)}(\kappa\rho) \frac{n(n+1)}{-ik\rho} \overline{\left\{ h_n^{(1)}(\kappa\rho) + \kappa\rho h_n^{(1)'}(\kappa\rho) \right\}} \right. \\ & \quad \left. + |\beta_{n,m}|^2 \overline{h_n^{(1)}(\kappa\rho)} \frac{n(n+1)}{ik\rho} \left\{ h_n^{(1)}(\kappa\rho) + \kappa\rho h_n^{(1)'}(\kappa\rho) \right\} \right] \end{aligned}$$

Now, taking the real part of both sides, we obtain

$$\Re \int_{\partial B_\rho} (\hat{x} \times E^s) \cdot \bar{H^s} dA = \rho^2 \sum_{n=1}^{\infty} \sum_{m=-n}^n \left[ |\alpha_{n,m}|^2 + |\beta_{n,m}|^2 \right] \frac{n(n+1)}{-i} W$$

where the Wronskian  $W$  is given by

$$W = \left( h_n^{(1)}(\kappa\rho) \overline{h_n^{(1)'}(\kappa\rho)} - \overline{h_n^{(1)}(\kappa\rho)} h_n^{(1)'}(\kappa\rho) \right).$$

From (9.51) we know that  $W = -2i/(\kappa Q)^2$ , so that

$$\Re \int_{\partial B_\rho} (\hat{x} \times E^s) \cdot \bar{H^s} dA = \frac{2}{\kappa^2} \sum_{n=1}^{\infty} \sum_{m=-n}^n n(n+1) \left[ |\alpha_{n,m}|^2 + |\beta_{n,m}|^2 \right].$$

Hence, under the assumption of this lemma,

$$\frac{2}{\kappa^2} \sum_{n=1}^{\infty} \sum_{m=-n}^n n(n+1) \left\{ |\alpha_{n,m}|^2 + |\beta_{n,m}|^2 \right\} \leq 0$$

and thus  $\alpha_{n,m} = \beta_{n,m} = 0$  for all appropriate  $n$  and  $m$ .  $\square$

We also have the following result that follows from this proof.

**Corollary 9.29** Let  $E$  be a radiating solution of Maxwell's equations in the complement of  $B_R$ . If  $E_\infty = 0$  then  $E = 0$  in  $\mathbb{R}^3 \setminus B_R$ .

With the aid of the Lemma 9.28 we can show that (9.27a)–(9.27c) have a unique solution.

**Corollary 9.30** *Given  $g \in H^{1/2}(\text{Div}; \partial B_R)$ , problem (9.27a)–(9.27c) has at most one solution  $E^s \in H_{\text{loc}}(\text{curl}; \mathbb{R}^3 \setminus \hat{B}_R)$ .*

**Proof** By the linearity of the problem we only need prove that there is a unique solution to (9.27a)–(9.27c) when  $g = 0$ . Multiplying (9.27a) by  $\bar{E}^s$  and integrating over the annulus  $\Omega_{R,R_1} = B_{R_1} \setminus \bar{B}_R$ ,  $R_1 > R$ , and using (3.27), we obtain

$$\begin{aligned} 0 &= \int_{\Omega_{R,R_1}} (\nabla \times \nabla \times E^s - \kappa^2 E^s) \cdot \bar{E}^s dV \\ &= \int_{\Omega_{R,R_1}} |\nabla \times E^s|^2 - \kappa^2 |E^s|^2 dV + \int_{\partial \Omega_{R,R_1}} (\nu \times \nabla \times E^s) \cdot \bar{E}^s dA, \end{aligned}$$

where  $\nu$  is the unit outward normal to  $\Omega_{R,R_1}$ . Using the vanishing boundary condition on  $\partial B_R$  and the definition of  $H$ , we obtain

$$\int_{\Omega_{R,R_1}} |\nabla \times E^s|^2 - \kappa^2 |E^s|^2 dV + i\kappa \int_{\partial B_{R_1}} (\nu \times H^s) \cdot \bar{E}^s dA = 0.$$

Taking the imaginary part of this equation (using the fact that  $\kappa$  is real) shows that

$$J \left( i\kappa \int_{\partial B_{R_1}} (\nu \times H^s) \cdot \bar{E}^s dA \right) = 0.$$

Hence

$$\Re \left( \kappa \int_{\partial B_{R_1}} (\nu \times H^s) \cdot \bar{E}^s dA \right) = 0.$$

Taking complex conjugates, we see that the condition of Lemma 9.28 is satisfied for all  $R < R_1 < \infty$  and hence  $E^s = 0$ . This completes the proof.  $\square$

## 9.5.2 Series solution

In this subsection we shall derive the Mei series solution of (9.27a)–(9.27c) using  $g$  given by (9.81). We follow Ishimaru [172]. For the more general case of a dielectric sphere, see [172, 283]. For a more sophisticated treatment, see [194].

Our first goal is to determine a series representation for the incident field that is suitable for matching to the series for  $E^s$  given in (9.64) on  $\partial B_R$ . In this case we seek a representation in the same form as (9.64) but using basis functions appropriate for the interior problem (see remark 9.18) since  $E^i$  is an analytic function of position in the neighborhood of  $B_R$ . Thus, we write

$$E^i = \sum_{n=1}^{\infty} \sum_{m=-n}^n \left\{ \tilde{\alpha}_{n,m} \tilde{M}_n^m(x) + \tilde{\beta}_{n,m} \tilde{N}_n^m(x) \right\},$$

where  $\alpha_{n,m}$  and  $\beta_{n,m}$  are coefficients to be determined, and  $\tilde{M}_n^m$  and  $\tilde{N}_n^m$  are given by (9.62). Using the definition of  $\tilde{M}_n^m$  and  $\tilde{N}_n^m$ , we may write this as (9.82)

$$\mathbf{E}^i = \nabla \times \left( \mathbf{x} I^{(1)}(\rho, \hat{\mathbf{x}}) \right) + \frac{1}{ik} \nabla \times \nabla \times \left( \mathbf{x} I^{(2)}(\rho, \hat{\mathbf{x}}) \right),$$

where  $\varrho = |\mathbf{x}|$  and

$$\begin{aligned} I^{(1)}(\rho, \hat{\mathbf{x}}) &= \sum_{n=1}^{\infty} \sum_{m=-n}^n \tilde{\alpha}_{n,m} j_n(k\rho) Y_n^m(\hat{\mathbf{x}}), \\ I^{(2)}(\rho, \hat{\mathbf{x}}) &= \sum_{n=1}^{\infty} \sum_{m=-n}^n \tilde{\beta}_{n,m} j_n(k\rho) Y_n^m(\hat{\mathbf{x}}). \end{aligned}$$

Using the definition of  $Y_n^m$  in (9.37), we obtain

$$I^{(2)}(\rho, \hat{\mathbf{x}}) = \sum_{n=1}^{\infty} \sum_{m=-n}^n j_n(k\rho) P_n^m(\cos \theta) \left( \tilde{\gamma}_{n,m}^{(1)} \cos(m\varphi) + \tilde{\gamma}_{n,m}^{(2)} \sin(m\varphi) \right),$$

where  $\tilde{\gamma}_{n,m}^{(j)}$ ,  $j = 1, 2$ , are coefficients related to the  $\beta_{n,m}$ . A similar expansion can also be obtained for  $I^{(1)}$ .

A direct computation of the curl in spherical coordinates given by (9.31) shows that the radial component of  $E^i$ , denoted  $E_\rho^i$ , is given by

$$E_\rho^i = \frac{1}{ik} \left( \frac{\partial^2}{\partial \rho^2} + \kappa^2 \right) \left( \rho I^{(2)}(\rho, \hat{\mathbf{x}}) \right),$$

where we have used the fact that  $I^{(2)}$  is a solution of the Helmholtz equation. Using equation (9.38) we see that

$$\left( \frac{d^2}{d\rho^2} + \kappa^2 - \frac{n(n+1)}{\rho^2} \right) (\rho j_n(k\rho)) = 0,$$

so (9.83)

$$E_\rho^i = \sum_{n=1}^{\infty} \sum_{m=0}^n \frac{n(n+1)}{ik\rho} j_n(k\rho) P_n^m(\cos \theta) \left( \tilde{\gamma}_{n,m}^{(1)} \cos(m\varphi) + \tilde{\gamma}_{n,m}^{(2)} \sin(m\varphi) \right).$$

Now we turn to the given incident field  $E^i = e_1 \exp(i\kappa x_3)$ . Note that

$$e_1 = \sin \theta \cos \varphi e_\rho + \cos \theta \cos \varphi e_\theta - \sin \varphi e_\varphi.$$

Thus, using the Jacobi–Anger expansion (9.53), we obtain

$$\begin{aligned} E_\rho^i &= \cos \varphi \sin \theta \exp(i\kappa \rho \cos \theta) = -\frac{\cos \varphi}{ik\rho} \exp(i\kappa \rho \cos \theta) \\ &= -\frac{\cos \varphi}{ik\rho} \sum_{n=0}^{\infty} i^n (2n+1) j_n(k\rho) \frac{\partial}{\partial \theta} P_n(\cos \theta). \end{aligned}$$

Noting that  $(d/d\theta)(P_0(\cos\theta)) = 0$  and  $(d/d\theta)(P_n(\cos\theta)) = -P_n^1(\cos\theta)$ , we obtain

$$E_\rho^i = -\frac{\cos\varphi}{\kappa\rho} \sum_{n=0}^{\infty} i^{n-1} (2n+1) j_n(\kappa\rho) P_n^1(\cos\theta).$$

Comparing this equation with (9.83), we see that  $\tilde{Y}_{n,m}^{(j)} = 0$  for  $j = 1, 2$  provided  $m \neq 1$  and  $\tilde{Y}_{n,1}^{(2)} = 0$  for all  $n$ . Finally,

$$\tilde{Y}_{n,1}^{(1)} = \frac{i^n(2n+1)}{n(n+1)}, \quad n = 1, 2, \dots$$

This provides the expansion for  $I^{(2)}$ . The expansion for  $I^{(1)}$  can be derived similarly. We thus obtain the following lemma.

**Lemma 9.31** *If  $E^i = e_i \exp(i\kappa x_3)$  then  $E^i$  has the series expansion (9.82), with (9.84)*

$$I^{(1)}(\rho, \hat{x}) = \sum_{n=1}^{\infty} j_n(\kappa\rho) P_n^1(\cos\theta) \frac{i^n(2n+1)}{n(n+1)} \cos\varphi, \quad (9.85)$$

$$I^{(2)}(\rho, \hat{x}) = \sum_{n=1}^{\infty} j_n(\kappa\rho) P_n^1(\cos\theta) \frac{i^n(2n+1)}{n(n+1)} \sin\varphi.$$

Now that we have a suitable expansion for the incident field, we can match it to the scattered field. Using the expression for  $E^s$  given in (9.64) outside  $B_R$ , we write

$$E^s = \sum_{n=1}^{\infty} \sum_{m=-n}^n \{ \alpha_{n,m} M_n^m(x) + \beta_{n,m} N_n^m(x) \},$$

where  $\alpha_{n,m}$  and  $\beta_{n,m}$  are constants depending on  $n$  and  $m$ . As before, using the definition of  $M_n^m$  and  $N_n^m$ , we may write this as (9.86)

$$E^s = \nabla \times (x I^{(1),s}(\rho, \hat{x})) + \frac{1}{ik} \nabla \times \nabla \times (x I^{(2),s}(\rho, \hat{x})),$$

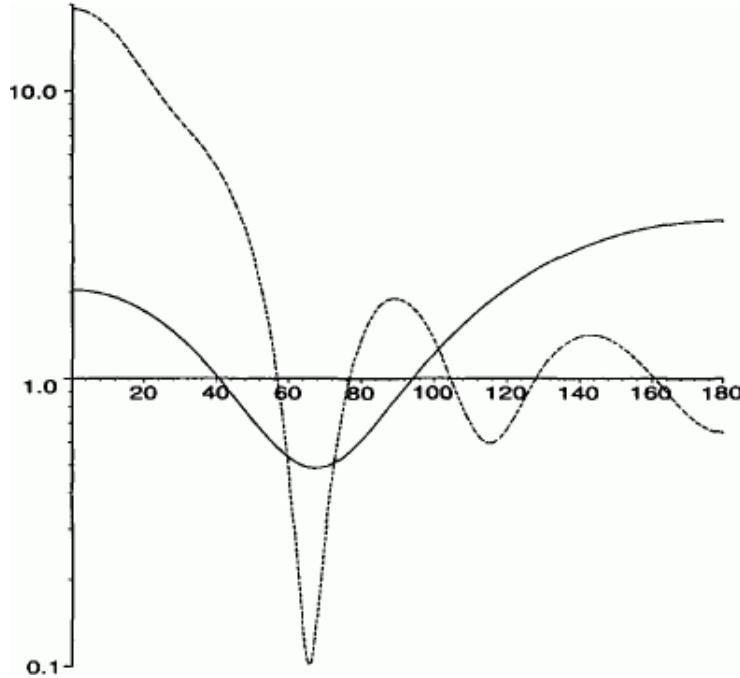
where (9.87)

$$I^{(1),s}(\rho, \hat{x}) = \sum_{n=1}^{\infty} \sum_{m=0}^n h_n^{(1)}(\kappa\rho) P_n^m(\cos\theta) \left( \delta_{n,m}^{(1)} \cos(m\varphi) + \delta_{n,m}^{(2)} \sin(m\varphi) \right)$$

and  $\delta_{n,m}^{(j)}, j = 1, 2$ , are coefficients to be determined (they are related to  $\alpha_{n,m}$ ). A similar expansion can also be obtained for  $I^{(2),s}$ . The boundary condition (9.27c) can be written as  $(E^s + E^i) \times \hat{x} = 0$  on  $\partial B_R$  and this reduces to

$$I^{(1),s} + I^{(1)} = 0,$$

Fig. 9.1. A plot of  $|E_\infty|$  against  $\theta$  for  $\varphi = 0$  and  $0 \leq \theta \leq \pi$  (converted to degrees) when  $R = 1$ . The dashed line shows the result for  $\kappa = 4.1$  and the solid line shows the result when  $\kappa = 1.1$ . This should be compared to Fig. 11.11 of [284] allowing for differences in the angles.



$$\frac{\partial}{\partial \rho} \left( \rho \left( I^{(2),s} + I^{(2)} \right) \right) = 0,$$

for  $\rho = R$  and all  $\hat{x}$ . Comparing (9.84) and (9.87) we thus obtain

$$I^{(1),s}(\rho, \hat{x}) = - \sum_{n=1}^{\infty} \frac{j_n(\kappa R)}{h_n^{(1)}(\kappa R)} \frac{i^n(2n+1)}{n(n+1)} h_n^{(1)}(\kappa \rho) P_n^1(\cos \theta) \cos \varphi,$$

and a similar calculation shows that

$$I^{(2),s}(\rho, \hat{x}) = - \sum_{n=1}^{\infty} \frac{j_n(\kappa R) + \kappa R j'_n(\kappa R)}{h_n^{(1)}(\kappa R) + \kappa R h_n^{(1)'}(\kappa R)} \frac{i^n(2n+1)}{n(n+1)} h_n^{(1)}(\kappa \rho) P_n^1(\cos \theta) \sin \varphi.$$

These formulae, when used in (9.86), give the famous *Mei series* for the scattered field. Although it may appear complicated, the series can be implemented in a few lines of MAPLE (a symbolic mathematics program) and so, by truncating the series, an approximation to  $E^s$  can be computed at any desired point in space.

The Mei series can also be used to approximate the far field pattern of the scattered field. Using the asymptotic estimates (9.44) and (9.45), it can be shown that (see [284])

$$E_{\infty} = \frac{i}{\kappa} \sum_{n=1}^{\infty} \frac{2n+1}{n(n+1)} \left\{ \frac{j_n(\kappa R)}{h_n^{(1)}(\kappa R)} \nabla_{\theta, \varphi} \left[ \cos \varphi P_n^1(\cos \theta) \right] \right. \\ \left. - \frac{j_n(\kappa R) + \kappa R j'_n(\kappa R)}{h_n^{(1)}(\kappa R) + \kappa R h_n^{(1)'}(\kappa R)} e_{\rho} \times \nabla_{\theta, \varphi} \left[ \cos \varphi P_n^1(\cos \theta) \right] \right\},$$

where  $\nabla_{\theta, \varphi}$  is the surface gradient on the surface of the unit sphere in spherical coordinates. As an example, we show a graph of  $|E_{\infty}|$  for  $\kappa = 1.1$  and  $\nu = 4.1$  when  $R = 1$  for  $0 \leq \theta \leq \pi$  and  $\varphi = 0$ . This should be compared with Fig. 11.10 of [284] allowing for the difference in the choice of angles. In this case we used a sum up to  $n = 6$ .

# 10 THE SCATTERING PROBLEM USING CALDERON MAPS

## 10.1 Introduction

In this chapter we shall present a variational formulation for the full scattering problem summarized in Section 1.4.3. In particular we shall allow both an inhomogeneous medium and a perfect conducting scatterer to be present. The formulation uses the electromagnetic analogue of the DtN map called the electric-to-magnetic Calderon operator described in Section 9.4.1. The method we shall describe is a generalization to Maxwell's equations of the variational formulation for the Helmholtz equation underlying the DtN absorbing boundary conditions in [182, 153, 154]. Understanding this method will provide a general existence result for weak solutions of the scattering problem which will be used later in Section 14. In addition, the variational formulation is suitable for applying finite elements to discretize Maxwell's equations, as well as being a starting point for understanding infinite elements [121] and absorbing boundary conditions [291, 154, 152]. Concerning the finite element approximation of this problem, we shall only give a partial convergence result assuming the use of an exact Calderon operator. For a discussion of the discretization of the Calderon operator via series truncation, see [121], where an analysis of this aspect of the discretization is carried out. We shall consider an alternative approach in the next chapter. The contents of this chapter are based on [190] (and the corrigendum [191]) together with [121].

The general idea of using boundary operators to truncate infinite domain problems has been used for sometime. My work with Andreas Kirsch on a similar method for the Helmholtz equation [188, 189] was motivated by the paper of Masmoudi [213], who carried out a program of analysis for the Helmholtz equation in  $\mathbb{R}^2$ . In the case of Maxwell's equations, Abboud and Nédélec [1] proved the existence and uniqueness for the continuous problem discussed here. They used the standard Sobolev space  $H^1(\Omega)$  as a basis for their variational problem, and so had to use a considerably more complex variational formulation than we shall need. For reasons discussed in Section 3.8, it seems desirable to avoid  $(H^1(\Omega))^3$ . Levillain [209] has analyzed the coupling of integral equations and variational methods for the problem we study. He shows that his formulation is equivalent to the formulation of Abboud and Nédélec.

Once we have analyzed the continuous problem to show that the variational formulation provides a solution of the scattering problem we shall prove the convergence of a finite element discretization. At the risk of annoying the reader, we do not use either of the techniques detailed in Chapter 7 (duality and collective

compactness). Instead, we shall verify the discrete version of the Babuška–Brezzi inf-sup condition of Theorem 2.22. Recently, Hiptmair [164] has used a similar approach to analyze edge element approximation of the perfectly conducting cavity problem. He uses a more sophisticated choice of the discrete projection operator. Hiptmair's proof and the proof here rely upon many of the same theoretical tools as the other proofs presented in Chapter 7. In common with the duality proof, the proof in this chapter results in a clean quasi-optimal error estimate, but also requires a restriction on  $\epsilon_r$  that we would prefer to avoid. Once we have discussed this proof, we will have seen three well-known proofs of convergence of edge elements in electromagnetism. After this chapter we shall use the collective compactness proof in the remainder of the book.

## 10.2 Reduction to a bounded domain

We wish to solve the time-harmonic scattering problem of computing a total field  $E$  that satisfies (1.26)–(1.29). As usual,  $D$  is assumed to be a bounded Lipschitz polyhedron with connected complement. For simplicity, we shall assume that  $\Gamma = \partial D$  is connected, and that the complement of  $D$  is simply connected, but these assumptions could be weakened at the expense of considerable technical details. The coefficients  $\epsilon_r$  and  $\mu_r$  in Maxwell's equations are assumed to satisfy the conditions of Section 4.2 and, in addition, we assume that there is a radius  $R_0$  such that  $\mu_r(x) = \epsilon_r(x) = 1$  when  $|x| > R_0$ , and such that  $D \subset B_{R_0}$ . Thus, the scatterer is assumed to be bounded.

Now we introduce a ball  $B_R$  with  $R > R_0$  and let  $\Omega = (\mathbb{R}^3 \setminus D) \cap B_R$ . This will be the computational domain. The auxiliary boundary is the boundary of  $B_R$  denoted  $\Sigma$ . We need to assume that the incident field  $E^i$  satisfies the homogeneous isotropic Maxwell system (i.e. with  $\epsilon_r = \mu_r = 1$ ) in  $B_{R_1}$  for some  $R_1 > R$ . This assumption can be removed by using a formulation based on the scattered field alone, so it is not essential. However, it is usually thought to be preferable to compute with the total field since this avoids potential difficulties with subtractive cancellation error in the shadow region. In that region, the total field has small magnitude, which implies that the scattered and incident field must almost cancel. Subtractive cancellation can lead to large relative errors.

The method we shall analyze is to use an exact non-local boundary conditions on the artificial boundary  $\Sigma$ . By virtue of our assumption on the incoming wave, we can assume that  $R$  is chosen so that  $E^i$  is analytic in a neighborhood of  $\Omega$ . In  $\Omega$  we shall solve for the total field  $E$ , while exterior to this domain we shall only solve for the scattered field  $E^s$ . Using  $\Omega$ , and matching the electromagnetic field across  $\Sigma$ , we obtain the problem of solving (10.1a)

$$\nabla \times \mu_r^{-1} \nabla \times E - \kappa^2 \epsilon_r E = 0 \text{ in } \Omega, \quad (10.1b)$$

$$\nabla \times \nabla \times E^s - \kappa^2 E^s = 0 \text{ in } \mathbb{R}^3 \setminus \bar{B}_R, \quad (10.1c)$$

$$\nu \times E = 0 \text{ on } \Gamma, \quad (10.1d)$$

$$E \times \hat{x} = E^s \times \hat{x} + E^i \times \hat{x} \text{ on } \Sigma,$$

$$\frac{1}{ik}(\nabla \times E) \times \hat{x} = \frac{1}{ik} \nabla \times (E^s + E^i) \times \hat{x} \text{ on } \Sigma, \quad (10.1e)$$

$$\lim_{x \rightarrow \infty} \rho((\nabla \times E^s) \times \hat{x} - i \kappa E^s) = 0. \quad (10.1f)$$

Here we have used the facts that  $\mu_r = \epsilon_r = 1$  for  $|x| > R_0$  and that  $\hat{x} = x/|x|$ .

We derive the Galerkin formulation of this problem as follows. If we multiply (10.1a) by a smooth test function  $\varphi$ , integrate over  $\Omega$  and formally use integration by parts, (3.51), we obtain

$$\begin{aligned} (\mu_r^{-1} \nabla \times E, \nabla \times \varphi) - \kappa^2 (\in_r E, \varphi) \\ + \int_{\partial\Omega} \hat{x} \times (\mu_r^{-1} \nabla \times E) \overline{(\hat{x} \times \varphi) \times \hat{x}} dA = 0, \end{aligned}$$

where  $(\cdot, \cdot)$  is the usual  $(L^2(\Omega))^3$  inner product.

Now using (10.1e) on  $\Sigma$  and requiring that  $v \times \varphi = 0$  on  $\Gamma$ , we may write

$$\begin{aligned} (\mu_r^{-1} \nabla \times E, \nabla \times \varphi) - \kappa^2 (\in_r E, \varphi) \\ + ik \left\langle \hat{x} \times \left( \frac{1}{ik} \nabla \times (E^s + E^i) \right), (\hat{x} \times \varphi) \times \hat{x} \right\rangle = 0, \end{aligned}$$

where, as usual,  $\langle \cdot, \cdot \rangle$  is the  $(L^2(\Sigma))^3$  inner product.

To complete the derivation of the variational problem, we need to make precise how  $\hat{x} \times (1/ik) v \times (E^s + E^i)$  depends on  $\hat{x} \times E$ . We use the Calderon operator  $G_e$  defined in Section 9.4.1. Choosing  $\lambda = \hat{x} \times (E - E^i)$  we see that the definition of  $G_e$  implies that on  $\Sigma$  we have  $\hat{x} \times (1/ik)(v \times E^s) = G_e(\hat{x} \times (E - E^i))$ . Our variational problem now becomes the problem of finding  $E$  such that  $v \times E = 0$  on  $\Gamma$  and

$$\begin{aligned} (\mu_r^{-1} \nabla \times E, \nabla \times \varphi) - \kappa^2 (\in_r E, \varphi) \\ + ik \left\langle G_e \left( \hat{x} \times \left( E - E^i \right) \right), \hat{x} \frac{1}{ik} \nabla \times E^i, \varphi_T \right\rangle = 0 \end{aligned}$$

for all smooth test functions  $\varphi$  such that  $v \times \varphi = 0$  on  $\Gamma$ .

From this discussion, we can see that the appropriate space for the solution  $E$  on  $\Omega$  is

$$\tilde{X} = \{u \in H(\text{curl}; \Omega) \mid v \times u = 0 \text{ on } \Gamma\}$$

equipped with the usual  $H(\text{curl}; \Omega)$  norm. Note that this space contains the space  $X$  defined in (4.3). Now we can state precisely the variational problem we shall analyze in this paper: find the vector field  $E \in X$  such that (10.2)

$$\begin{aligned} (\mu_r^{-1} \nabla \times E, \nabla \times \varphi) - \kappa^2 (\in_r E, \varphi) + ik \left\langle G_e(\hat{x} \times E^i), \varphi_T \right\rangle \\ = \left\langle i \kappa G_e(\hat{x} \times E^i) - \hat{x} \times \nabla \times E^i, \varphi_T \right\rangle \end{aligned}$$

for all  $\varphi \in X$ .

We shall show that this problem has a unique solution by applying an argument that generalizes our analysis of cavity

problems in Chapter 4 . This provides a simple proof of the existence of solutions of the scattering problem. The variational formulation is suitable for discretization using edge elements. In the last section we perform an error analysis by verifying a discrete inf-sup condition.

## 10.3 Analysis of the reduced problem

In order to provide a rigorously justified variational formulation for the scattering problem, we must analyze in some detail the variational problem (10.2). In particular, we shall prove the existence of a unique solution to this problem. First we show that the scattering problem has at most one solution.

**Theorem 10.1** *Under the standing assumptions for this chapter, there is at most one solution*

$$E \in H_{\text{loc}}(\text{curl}; \mathbb{R}^3 \setminus \bar{D})$$

*of the scattering problem (1.26)–(1.29).*

**Proof** By linearity, we need only consider the case  $E^i = 0$  whence  $E = E^s$  and thus  $E$  is a radiating solution of Maxwell's equations in the exterior of  $D$ . By taking the dot product of Maxwell's equations with  $E$  and integrating over  $\Omega$  (defined in the previous section), and then integrating by parts using (3.51) we obtain

$$\int_{\partial\Omega} v \times \mu_r^{-1} \nabla \times E \cdot \bar{E} dA + (\mu_r^{-1} \nabla \times E, \nabla \times E) - k^2 (\in_r E, E) = 0.$$

Using the fact that  $v \times E = 0$  on  $\Gamma$  and recalling that  $H = (1/i\kappa)\mu_r^{-1} \nabla \times E$ , we obtain

$$i\kappa \int_{\partial\Omega} v \times H \cdot \bar{E} dA + (\mu_r^{-1} \nabla \times E, \nabla \times E) - \kappa^2 (\in_r E, E) = 0.$$

Taking the complex conjugate of both sides, we obtain that

$$\Re \left( \int_{\partial\Omega} v \times E \cdot \bar{H} dA \right) = -\kappa \int_{\Omega} \Im(\in_r E) \cdot \bar{E} dV \leq 0.$$

Hence, by the Rellich Lemma 9.28, we conclude  $E = 0$  on  $\mathbb{R}^3 \setminus \mathcal{B}_R$ . We can now apply the unique continuation principle in Lemma 4.13 as in the proof of Theorem 4.12 to show that  $E = 0$  on  $\Omega$ . This completes the proof.  $\square$

Now suppose that the reduced problem (10.2) has a solution. The series solution constructed in Section 9.3.3 then provides an extension of this solution from the bounded domain  $\Omega$  to  $\mathbb{R}^3 \setminus \mathcal{B}_R$ . Due to the use of the Calderon map  $G_c$ , this extended solution satisfies the Maxwell system (10.1a) in the weak sense in  $H_{\text{loc}}(\text{curl}; \mathbb{R}^3 \setminus \bar{D})$  together with the radiation condition. The general uniqueness result in Theorem 10.1 then implies that this extension is the only solution of (10.2). Hence, once we have proved the existence of a solution of (10.2), we shall also have verified that (1.26)–(1.29) has a unique weak solution in  $H_{\text{loc}}(\text{curl}; \mathbb{R}^3 \setminus \bar{D})$ .

Now we shall start the analysis of the truncated problem. First, we write (10.2) as the problem of finding  $E \in X$  such that(10.3)

$$A(E, \varphi) = B(\varphi) \text{ for all } \varphi \in X,$$

where(10.4)

$$\begin{aligned} A(E, \varphi) &= \left( \mu_r^{-1} \nabla \times E, \nabla \times \varphi \right) - k^2 (\epsilon_r E, \varphi) \\ &\quad + i k \langle G_e(\hat{x} \times E), (\hat{x} \times \varphi) \times \hat{x} \rangle, \\ B(\varphi) &= \left\{ ik G_e \left( \hat{x} \times E^i \right) - \hat{x} \times \nabla \times E^i, (\hat{x} \times \varphi) \times \hat{x} \right\}. \end{aligned} \quad (10.5)$$

The sesquilinear form  $A(\cdot, \cdot)$  is a generalization of the form  $a(\cdot, \cdot)$  encountered in (4.5).

We now follow a path similar to that in Section 4.5, where we analyzed the cavity problem. We use a Helmholtz decomposition to factor out the null-space of the curl operator. This is given by the gradient of the following scalar space:

$$\tilde{S} = \left\{ q \in H^1(\Omega) \mid p = 0 \text{ on } \Gamma \right\}.$$

If there is no perfect conductor present, so  $D = \emptyset$ , we use

$$\tilde{S} = \left\{ q \in H^1(\Omega) \mid \int_{|x|=R} q dA = 0 \right\}.$$

Then we seek  $p \in \tilde{S}$  such that(10.6)

$$A(\nabla p, \nabla \xi) = B(\nabla \xi) \text{ for all } \xi \in \tilde{S}.$$

This equation can be rewritten as(10.7)

$$a_1(p, \xi) + b_1(p, \xi) = B(\nabla \xi) \text{ for all } \xi \in \tilde{S},$$

where we define (using  $G_e$  defined in (9.74))

$$\begin{aligned} a_1(p, \xi) &= -k^2 (\epsilon_r \nabla p, \nabla \xi) + k^2 \langle \tilde{G}_e(\hat{x} \times \nabla p), \nabla_\Sigma \xi \rangle, \\ b_1(p, \xi) &= ik \langle (G_e + ik \tilde{G}_e)(\hat{x} \times \nabla p), \nabla_\Sigma \xi \rangle, \quad p, \xi \in \tilde{S}. \end{aligned}$$

Here we have used (3.14) to write the tangential component of the gradient of  $\xi$  in terms of the tangential gradient on  $\Sigma$ . From Lemma 9.23 we know that  $G_e$  is negative definite and thus  $a_1(p, \xi)$  is a coercive sesquilinear form on  $\tilde{S} \times \tilde{S}$ .

To analyze (10.7), we introduce the operator  $K_1 : \tilde{S} \rightarrow \tilde{S}$  defined by

$$a_1(K_1 p, \xi) = b_1(p, \xi) \text{ for all } p, \xi \in \tilde{S}.$$

and the function  $b \in \tilde{S}$  defined by

$$\alpha_1(b, \xi) = B(\nabla \xi) \text{ for all } \xi \in \tilde{S}.$$

Given the previous observation concerning the positive definiteness of  $\alpha_1$  and the fact that  $G_\epsilon$  is bounded from  $H^{1/2}(\text{Div}; \Sigma)$  to  $H^{-1/2}(\text{Div}; \Sigma)$ , we can show that  $\alpha_1$  is continuous on  $S \times S$  and apply the Lax–Milgram Lemma 2.21. We conclude that  $K_1$  and  $b$  are well defined. Furthermore,  $p \in S$  satisfies(10.8)

$$(I + K_1)p = b.$$

By Lemma 9.24,  $G_\epsilon + i\kappa \tilde{G}_\epsilon : H_{\text{Div}}^{-1/2}(\text{Div}; \Sigma) \rightarrow H^{-1/2}(\text{Div}; \Sigma)$  is compact, and since  $\hat{x} \times \nabla p \in H_{\text{Div}}^{-1/2}(\text{Div}; \Sigma)$  for all  $p \in S$ , so we see that  $K_1$  is compact. Thus, we have a Fredholm equation.

We now need to prove uniqueness of any solution of (10.8). By linearity it suffices to consider the case when  $b = 0$ . Then  $p$  satisfies

$$-k^2(\epsilon_r \nabla p, \nabla \xi) + i\kappa \langle G_\epsilon(\hat{x} \times \nabla p), \nabla_\Sigma \xi \rangle = 0 \text{ for all } \xi \in \tilde{S}.$$

Choosing  $\xi = p$ , we have

$$i\kappa \langle G_\epsilon(\hat{x} \times \nabla p), \nabla_\Sigma p \rangle = \kappa^2 (\epsilon_r \nabla p, \nabla p).$$

But by the definition of  $G_\epsilon$ , if  $u \in H_{\text{loc}}(\text{curl}; \mathbb{R}^3 \setminus \bar{B}_R)$  is the weak solution (whose existence is verified by Theorem 9.17) of(10.9)

$$\begin{aligned} \nabla \times \nabla \times u - \kappa^2 u &= 0 \text{ in } \mathbb{R}^3 \setminus \bar{B}_R, \\ \hat{x} \times u &= \hat{x} \times \nabla p \text{ on } \Sigma, \end{aligned} \tag{10.10}$$

together with the Silver–Müller radiation condition, then

$$G_\epsilon(\hat{x} \times \nabla p) = \hat{x} \times v \text{ on } \Sigma,$$

where  $v = (1/i\kappa)^\vee \times u$  (i.e. the magnetic field corresponding to the electric field  $u$ ). The integral appearing in the Rellich uniqueness Lemma 9.28 is given by

$$\begin{aligned} \int_{\Sigma} \hat{x} \times u \cdot \bar{v} dA &= - \langle \nabla_\Sigma p, \hat{x} \times v \rangle = - \langle \nabla_\Sigma p, G_\epsilon(\hat{x} \times \nabla p) \rangle \\ &= - \overline{\langle G_\epsilon(\hat{x} \times \nabla p), \nabla_\Sigma p \rangle} = - \overline{\frac{\kappa}{i} (\epsilon_r \nabla p, \nabla p)}. \end{aligned}$$

Thus,

$$\Re \left( \int_{\Sigma} \hat{x} \times u \cdot \bar{v} dA \right) = -\kappa (J(\epsilon_r) \nabla p, \nabla p) \leq 0,$$

and by the Rellich lemma,  $u = 0$  in  $\mathbb{R}^3 \setminus \bar{B}_R$ . From (10.10), this implies that  $\nabla_\Sigma p = 0$  on  $\Sigma$  and thus  $(\epsilon_r \nabla p, \nabla p) = 0$ . We conclude via the Poincaré inequality in Lemma 3.13 that  $p = 0$ .

Since (10.8) has at most one solution and is a Fredholm equation, the Fredholm alternative in Theorem 2.33 completes the proof of the following result.

**Theorem 10.2** Under the conditions on the coefficients and data outlined in Section 10.2 and assuming that  $\epsilon_r = \mu_r = 1$  in a neighborhood of  $\Sigma$ , the following hold:

- The sesquilinear form  $a_1$  is bounded and coercive on  $\mathcal{S} \times \mathcal{S}$ . There exists a compact operator  $K_1$  from  $\mathcal{S}$  into itself with  $b_1(p, \xi) = a_1(K_1 p, \xi)$  for all  $p, \xi \in \mathcal{S}$ .
- The operator  $I + K_1$  is an isomorphism from  $\mathcal{S}$  onto itself. The variational problem of finding  $p \in \mathcal{S}$  such that  $A(\nabla p, \nabla \xi) = B(\nabla \xi)$  for all  $\xi \in \mathcal{S}$  is uniquely solvable in  $\mathcal{S}$  and the solution is given by  $p = (I + K_1)^{-1} b$ , where  $b \in \mathcal{S}$  satisfies  $a_1(b, \xi) = B(\nabla \xi)$  for all  $\xi \in \mathcal{S}$ .

We are now in a position to factor out  $\nabla \mathcal{S}$  from  $\tilde{X}$  by using an extension of the Helmholtz decomposition. This is done in the next section.

### 10.3.1 Extended Helmholtz decomposition

We use an extended Helmholtz decomposition (described in Lemma 10.3) using the space  $X_0$  defined by(10.11)

$$\begin{aligned} \tilde{X}_0 &= \left\{ u \in \tilde{X} \mid -k^2 (\epsilon_r u, \nabla \xi) + ik \langle G_e(\hat{x} \times u), \nabla_\Sigma \xi \rangle = 0 \text{ for all } \xi \in \tilde{\mathcal{S}} \right\} \\ &= \left\{ u \in \tilde{X} \mid \nabla \cdot (\epsilon_r u) = 0 \text{ in } \Omega, \text{ and } \hat{x} \cdot u = -\frac{i}{\kappa} \nabla_\Sigma \cdot G_e(\hat{x} \times u) \text{ on } \Sigma \right\}. \end{aligned}$$

This space plays the role that  $X_0$  played in the variational theory of the cavity problem in Section 4. We then have the following Helmholtz decomposition for these spaces.

**Lemma 10.3** *The spaces  $\nabla \mathcal{S}$  and  $X_0$  are closed subspaces of  $\tilde{X}$ . The space  $\tilde{X}$  is the direct sum of the spaces  $\nabla \mathcal{S}$  and  $X_0$ , that is,*

$$\tilde{X} = \tilde{X}_0 \oplus \nabla \tilde{\mathcal{S}}.$$

Furthermore, the projections onto the subspaces are bounded, that is there exist constants  $C_1, C_2 > 0$  with

$$\begin{aligned} C_1 \| w + \nabla p \|_{H(\text{curl}; \Omega)}^2 &\leq \| w \|_{H(\text{curl}; \Omega)}^2 + \| \nabla p \|_{H(\text{curl}; \Omega)}^2 \\ &\leq C_2 \| w + \nabla p \|_{H(\text{curl}; \Omega)}^2 \end{aligned}$$

for all  $w \in X_0$  and  $p \in \mathcal{S}$ .

**Proof** The closedness of  $\nabla \mathcal{S}$  is obvious. The subspace  $X_0$  is closed since, for fixed  $\xi \in \mathcal{S}$ , the linear functionals  $u \mapsto (\epsilon_r u, \nabla \xi)$  and  $u \mapsto \langle G_e(\hat{x} \times u), \nabla_\Sigma \xi \rangle$  are bounded on  $H(\text{curl}; \Omega)$  (the latter by the boundedness of the trace operator  $\gamma_\tau$  from  $H(\text{curl}; \Omega)$  into  $H^{1/2}(\text{Div}; \Sigma)$ ).

To show that  $\tilde{X} = X_0 \oplus \nabla \mathcal{S}$ , let  $u \in \tilde{X}$  be fixed. Define  $p \in \mathcal{S}$  to be the solution of

$$A(\nabla p, \nabla \xi) = A(u, \nabla \xi) \text{ for all } \xi \in \tilde{\mathcal{S}}.$$

Theorem 10.2 shows that this problem is wellposed and that there exists  $C > 0$  (independent of  $u$ ) with

$$\|\nabla p\|_{(L^2(\Omega))^3} \leq C \|u\|_{H(\text{curl};\Omega)}.$$

Furthermore, if  $w = u - \nabla p$  then  $w \in X_0$  as is seen directly from the variational equation. Finally, we have to show that  $\nabla S \cap X_0$  consists of the trivial function only. Suppose  $u = \nabla p \in \nabla S \cap X_0$ . Then

$$0 = A(u, \nabla \xi) = A(\nabla p, \nabla \xi) \quad \text{for all } \xi \in \tilde{S},$$

which implies (again by Theorem 10.2) that  $p = 0$ . This completes the proof.  $\square$

As a consequence of the compact imbedding theorem (Theorem 4.7), we have:

**Lemma 10.4** *The space  $X_0$  is compactly imbedded in  $(L^2(\Omega))^3$ .*

**Proof** Consider a bounded set of functions  $\{u_j\}_{j=1}^\infty \subset \tilde{X}_0$ . Each function  $u_j \in X_0$  can be extended to all of  $\mathbb{R}^3$  by solving the exterior Maxwell problem

$$\begin{aligned} \nabla \times (\nabla \times v_j) - \kappa^2 v_j &= 0 && \text{in } \mathbb{R}^3 \setminus \overline{B_R}, \\ \hat{x} \times v_j &= \hat{x} \times u_j && \text{on } \Sigma, \end{aligned}$$

together with the Silver–Müller radiation condition at infinity. The function  $u_j^e$  defined by

$$u_j^e = \begin{cases} u_j & \text{on } \Omega, \\ v_j & \text{on } \mathbb{R}^3 \setminus \overline{B_R}, \end{cases}$$

is in  $H_{\text{loc}}(\text{curl}; \mathbb{R}^3 \setminus D^-)$  since the tangential components are continuous across  $\Sigma$  (see Theorem 5.3). Furthermore, since  $u_j \in X_0$ , we have the constraint that  $\kappa^2 \hat{x} \cdot u_j = -i\kappa \nabla_\Sigma \cdot G_e(\hat{x} \times u_j)$ . But

$$G_e(\hat{x} \times u_j) = \frac{1}{i\kappa} \hat{x} \times \nabla \times u_j$$

and so using Maxwell's equations and (3.52) we have that  $\hat{x} \cdot u_j = \hat{x} \cdot u_j$  on  $\Sigma$ . Thus, the normal component of  $u_j^e$  is continuous and this extended function has a well-defined divergence. The divergence free conditions inside  $\Omega$  and in the complement of  $B_R$  then show that  $\nabla \cdot (u_j^e) = 0$  in  $\mathbb{R}^3 \setminus D^-$ .

Now we choose a cutoff function  $\chi \in C_0^\infty(\mathbb{R}^3)$  such that  $\chi = 1$  in  $\Omega^-$ . We can apply the general compactness Theorem 4.7 to the sequence  $\{\chi u_j^e\}$  and extract a subsequence converging strongly in  $(L^2(\Omega))^3$ . This completes the proof.  $\square$

### 10.3.2 An operator equation on $X_0$

In order to complete our proof of the existence of a solution of (10.2) we now use the Helmholtz decomposition from the previous section to decompose  $E = w + \nabla p$  for uniquely determined  $w \in X_0$  and  $p \in S$ . We observe that  $A(w, \nabla \xi) = 0$  for all  $\xi \in S$  by the definition of  $X_0$ . Therefore, we can decompose the variational equation (10.3) as (10.12)

$$A(\nabla p, \nabla \xi) + A(\nabla p, \psi) + A(w, \psi) = B(\nabla \xi) + B(\psi)$$

for all  $\psi \in X_0$  and  $\xi \in S$ .

Choosing  $\psi = 0$  we see that  $p$  satisfies (10.6), and so Theorem 10.2 shows that  $p$  is well defined and continuously dependent on the data. It thus remains to show that  $w \in X_0$ , which satisfies the equation (10.13)

$$A(w, \psi) = B(\psi) - A(\nabla p, \psi) \text{ for all } \psi \in \tilde{X}_0,$$

exists and is continuously dependent on the data. This will be done by decomposing the sesquilinear form into a coercive and compact part. Define

$$B(\psi) = B(\psi) - A(\nabla p, \psi) \text{ for all } \psi \in \tilde{X}_0,$$

At this point, we need to examine  $G_e$  in more detail. Suppose that  $\lambda \in H^{1/2}(\text{Div}; \Sigma)$  has the expansion

$$\lambda = \sum_{n=1}^{\infty} \sum_{m=-n}^n a_{n,m} U_n^m + b_{n,m} V_n^m.$$

Using the sequence  $\{\delta_n\}$  defined in Lemma 9.22 and the expansion for  $G_e$  in (9.71) we can write

$$\begin{aligned} G_e \lambda &= \sum_{n=1}^{\infty} \sum_{m=-n}^n \left[ -i\kappa R \frac{b_{nm}}{\delta_n} U_n^m + \frac{a_{nm}(\delta_n - \tilde{\delta}_n)}{i\kappa R} V_n^m \right] \\ &\quad + \frac{1}{i\kappa R} \sum_{n=1}^{\infty} \sum_{m=-n}^n a_{nm} \tilde{\delta}_n V_n^m \\ &= G_e^1 \lambda + G_e^2 \lambda, \end{aligned}$$

where the two sums in the above expression define  $G_e^1$  and  $G_e^2$  respectively. Next we show that the first series on the right-hand side above defines a compact operator from  $X_0$  into  $H^{1/2}(\text{Div}; \Sigma)$ .

**Lemma 10.5** *Let  $\gamma_t : H(\text{curl}; \Omega) \rightarrow H^{1/2}(\text{Div}; \Sigma)$  be the trace operator defined in (3.45). The operator  $G_e^1 \circ \gamma_t$ , that is the mapping  $G_e^1(\hat{x} \times u)$ , is compact from  $X_0$  into  $H^{1/2}(\text{Div}; \Sigma)$ .*

**Proof** We split  $G_e^1$  into two parts:

$$\begin{aligned} G_e^{1,U} \lambda &= \sum_{n=1}^{\infty} \sum_{m=-n}^n -i\kappa R \frac{b_{nm}}{\delta_n} U_n^m, \\ G_e^{1,V} \lambda &= \sum_{n=1}^{\infty} \sum_{m=-n}^n \frac{a_{nm}(\delta_n - \tilde{\delta}_n)}{i\kappa R} V_n^m. \end{aligned}$$

Using the expansion for  $\delta_n$  in the proof of Lemma 9.20 and for  $\tilde{\delta}_n$  in the proof of Lemma 9.22, the asymptotic behavior of  $\delta_n - \tilde{\delta}_n$  is given by

$$\delta_n - \tilde{\delta}_n = O\left(\frac{1}{n}\right), \quad n \geq 1.$$

Thus  $G_e^{1,V}$  is certainly compact from  $H^{1/2}(\text{Div}; \Sigma)$  into itself. Therefore, from the boundedness of the trace operator  $\gamma_t : H(\text{curl}; \Omega) \rightarrow H^{1/2}(\text{Div}; \Sigma)$ , we conclude that  $G_e^{1,V} \circ \gamma_t$  is compact from  $X_0$  into  $H^{1/2}(\text{Div}; \Sigma)$ .

From the definitions of the norm on  $H^{1/2}(\text{Div}; \Sigma)$  in (9.58) and the definition of  $G_e^{1,U}$ , we see that, for  $u \in X_0$ ,

$$\begin{aligned} \| (G_e^{1,U} \circ \gamma_t) u \|_{H^{-1/2}(\text{Div}, \Sigma)} &= \| G_e^{1,U}(\hat{x} \times u) \|_{H^{-1/2}(\text{Div}, \Sigma)} \\ &\leq C \| \nabla_{\Gamma} \cdot G_e(\hat{x} \times u) \|_{H^{-1/2}(\Sigma)} \end{aligned}$$

Now using the boundary condition satisfied by functions in  $X_0$  on  $\Sigma$  and the trace theorem in  $H(\text{div}; \Omega)$  (Theorem 3.24) we have that

$$\begin{aligned} \| (G_e^{1,U} \circ \gamma_t) u \|_{H^{-1/2}(\text{Div}, \Sigma)} &= C \| \hat{x} \cdot u \|_{H^{-1/2}(\Sigma)} \\ &\leq C \sqrt{\| u \|_{(L^2(\Omega))^3}^2 + \| \nabla \cdot (\in_r u) \|_{(L^2(\Omega))}^2} \\ &= C \| u \|_{(L^2(\Omega))^3} \end{aligned}$$

The compactness of  $X_0$  in  $(L^2(\Omega))^3$  (Lemma 10.4) yields the assertion.  $\square$

This lemma suggests to split the sesquilinear form  $A(\cdot, \cdot)$  into  $A = a_2 + b_2$  where

$$\begin{aligned} a_2(u, \psi) &= \left( u_{\Gamma}^{-1} \nabla \times u, \nabla \times \psi \right) + \kappa^2 (\in_r u, \psi) + ik \left\langle G_e^2(\hat{x} \times u), \psi_T \right\rangle, \\ b_2(u, \psi) &= -2k^2 (\in_r u, \psi) + ik \left\langle G_e^1(\hat{x} \times u), \psi_T \right\rangle. \end{aligned}$$

From the expansion for  $G_e^2$  we obtain

$$ik \left\langle G_e^2(\hat{x} \times \lambda), \lambda_T \right\rangle = -\frac{1}{R} \sum_{n=0}^{\infty} \sum_{m=-n}^n |b_{nm}|^2 \tilde{\delta}_n \geq 0.$$

Hence we conclude that  $a_2$  is coercive (see also Lemma 4.10); in addition, we shall show that  $b_2$  is compact. This gives rise to the compact operator  $K_2 : X_0 \rightarrow X_0$  defined by

$$b_2(u, \psi) = a_2(\mathcal{K}_2 u, \psi) \quad \text{for all } u, \psi \in \tilde{X}_0.$$

We have the following result.

**Theorem 10.6** Under the same hypotheses as Theorem 10.2 the following hold:

- The sesquilinear form  $a_2$  is bounded and coercive on  $\tilde{X}_0 \times \tilde{X}_0$ . There exists a compact operator  $K_2$  from  $\tilde{X}_0$  into itself with  $b_2(F, \varphi) = a_2(K_2 F, \varphi)$  for all  $F, \varphi \in \tilde{X}_0$ .
- The operator  $I + K_2$  is an isomorphism from  $\tilde{X}_0$  onto itself. The variational equation  $A(w, \varphi) = \tilde{B}(\varphi)$  for all  $\varphi \in \tilde{X}_0$  is uniquely solvable in  $\tilde{X}_0$  and the solution is given by  $w = (I + K_2)^{-1}B$ , where  $B$  satisfies  $a_2(B, \varphi) = \tilde{B}(\varphi)$  for all  $\varphi \in \tilde{X}_0$ .

**Proof** The sesquilinear form  $a_2$  is obviously bounded. By Lemma 9.23 and using the same argument as in the proof of Lemma 4.10 we can verify that

$$|a_2(u, u)| \geq C \|u\|_{H(\text{curl}; \Omega)}^2 \quad \text{for all } u \in \tilde{X}.$$

The Lax–Milgram Lemma 2.21 yields the existence of the operator  $K_2$  and the element  $B$ . Now we have to show that  $K_2$  is compact and that  $I + K_2$  is one-to-one.

To prove compactness, let  $\{F_n\} \subset X_0$  be a sequence converging to zero weakly in  $X$ . Then the trace  $\hat{x} \times F_n$  converges to zero weakly in  $H^{1/2}(\text{Div}; \Sigma)$  and, by Lemma 10.5,  $\|i\kappa G_e^1(\hat{x} \times F_n)\|_{H^{-1/2}(\text{Div}; \Sigma)}$  converges to zero. Also, by Lemma 10.4,  $\|F_n\|_{(L^2(\Omega))^3}^2 \rightarrow 0$ . Altogether we estimate

$$\begin{aligned} |b_2(F_n, \varphi)| &\leq \|i\kappa G_e^1(\hat{x} \times F_n)\|_{H^{1/2}(\text{Div}; \Sigma)} \|\varphi_T\|_{H^{1/2}(\text{curl}; \Sigma)} \\ &\quad + C \|F_n\|_{(L^2(\Omega))^3} \|\varphi\|_{(L^2(\Omega))^3} \\ &\leq C_n \|\varphi\|_{H(\text{curl}; \Omega)}, \end{aligned}$$

with  $C_n \rightarrow 0$  as  $n \rightarrow \infty$ . Then

$$\begin{aligned} C \|K_2 F_n\|_{H(\text{curl}; \Omega)}^2 &\leq a_2(K_2 F_n, K_2 F_n) \\ &= b_2(F_n, K_2 F_n) \leq C_n \|K_2 F_n\|_{H(\text{curl}; \Omega)}. \end{aligned}$$

which shows that  $\|K_2 F_n\|_{H(\text{curl}; \Omega)} \rightarrow 0$ . This proves that  $K_2$  is compact.

It remains to prove that  $(I + K_2)w = 0$  has only the trivial solution  $w = 0$ . If  $(I + K_2)w = 0$  then  $w$  satisfies

$$A(w, \varphi) = a_2(w + K_2 w, \varphi) = 0 \quad \text{for all } \varphi \in \tilde{X}_0.$$

But since  $w \in X_0$ , we have, for any  $p \in S$ ,

$$A(w, \varphi + \nabla p) = A(w, \nabla p) + A(w, \varphi) = 0,$$

so that  $w$  (extended to  $\mathbb{R}^3 \setminus \bar{B_R}$  as a solution of Maxwell's equations) is a weak solution of the exterior scattering problem with vanishing incoming wave and hence  $w = 0$  (see Theorem 10.1).

The Fredholm alternative now shows the existence of  $w$  for general data and completes the proof of the theorem.  $\square$

We now combine these results in the following main theorem of this section:

**Theorem 10.7** Under the same hypotheses as Theorem 10.2 the variational equation (10.2) is uniquely solvable in  $\tilde{X}$  for every incident field  $E^i$  that is a regular solution of the background Maxwell system in  $B_R$ .

It will be useful later in this chapter, and in Chapter 14, to have a generalization of this theorem. Let  $\lambda \in Y(\Gamma)$  ( $Y(\Gamma)$  is defined in (3.50)) and suppose that we wish to find  $E_\lambda \in H_{\text{loc}}(\text{curl}; \mathbb{R}^3 \setminus \bar{D})$  such that (10.14)

$$\nabla \times \mu_r^{-1} \nabla \times E_\lambda - k^2 E_\lambda = 0 \quad \text{in } \mathbb{R}^3 \setminus \bar{D}, \quad (10.15)$$

$$\nu \times E_\lambda = \lambda \quad \text{on } \Gamma,$$

and  $E_\lambda$  satisfies the Silver–Müller radiation condition. We have the following theorem.

**Theorem 10.8** Problem (10.14)–(10.15), together with the Silver–Müller radiation condition, has a unique solution  $E_\lambda \in H_{\text{loc}}(\text{curl}; \mathbb{R}^3 \setminus \bar{D})$  for any  $\lambda \in Y(\Gamma)$ . For any  $R > 0$  sufficiently large there is a constant  $C$  depending on  $R$  such that

$$\|E_\lambda\|_{H(\text{curl}; (\mathbb{R}^3 \setminus \bar{D}) \cap B_R)} \leq C \|\lambda\|_{Y(\Gamma)}.$$

**Proof** Let  $R$  be large enough that the scatterer is contained in  $B_R$  (i.e.  $R > R_0$  where  $R_0$  is defined in Section 10.2) and let  $\Omega = (\mathbb{R}^3 \setminus \bar{D}) \cap B_R$ . By the definition of  $\lambda \in Y(\Gamma)$  there is an  $F \in H(\text{curl}; \Omega)$  such that  $\nu \times F = \lambda$  on  $\Gamma$ . Now let  $\tilde{E} = E_\lambda - F \in \tilde{X}$ , which satisfies

$$A(\tilde{E}, \varphi) = -A(F, \varphi) \quad \text{for all } \varphi \in \tilde{X}.$$

The argument proving the previous theorem now yields the existence of  $\tilde{E} \in \tilde{X}$  and

$$\|\tilde{E}\|_{H(\text{curl}; \Omega)} \leq C \|F\|_{H(\text{curl}; \Omega)}.$$

Hence  $E_\lambda$  exists and  $\|E_\lambda\|_{H(\text{curl}; \Omega)} \leq C \|F\|_{H(\text{curl}; \Omega)}$  with  $C$  independent of  $F$ . Taking the infimum over all  $F \in H(\text{curl}; \Omega)$  such that  $\nu \times F = \lambda$  proves the desired result (taking into account the definition of the norm on  $Y(\Gamma)$ ).  $\square$

To motivate our proof of convergence of the finite element approximation to this problem, we next note that the results we have obtained so far enable us to verify the Babuška–Brezzi condition of Theorem 2.22. It will be useful here and in the next section to define (10.16)

$$A_1 = I + \mathcal{K}_1: \tilde{S} \rightarrow \tilde{S} \quad \text{and} \quad A_2 = I + \mathcal{K}_2: \tilde{X}_0 \rightarrow \tilde{X}_0.$$

**Lemma 10.9** Under the assumptions of Theorem 10.2, there exists  $\alpha > 0$  such that

$$\sup_{\varphi \in \tilde{X}} \frac{|A(u, \varphi)|}{\|\varphi\|_{H(\text{curl}; \Omega)}} \geq \alpha \|u\|_{H(\text{curl}; \Omega)} \quad \text{for all } u \in \tilde{X}.$$

The constant  $\alpha$  is independent of  $u$ .

**Proof** Let  $u \in X$  and use the generalized Helmholtz decomposition of Lemma 10.3 to write  $u = w + \nabla p$  for unique  $w \in X_0$  and  $p \in S$ . Now we take, for arbitrary  $\beta > 0$  independent of  $u$ ,

$$\varphi = A_2 w - \beta \nabla A_1 p,$$

where  $A_j$ ,  $j = 1, 2$ , are defined in (10.16). We shall shortly make a particular choice for  $\beta$ .

Using the fact that  $\mathcal{A}(w, \nabla \xi) = 0$  for all  $\xi \in S$ , we obtain (10.17)

$$\mathcal{A}(u, \varphi) = \mathcal{A}(w, A_2 w) - \beta \mathcal{A}(\nabla p, \nabla A_1 p) + \mathcal{A}(\nabla p, A_2 w).$$

For the first term on the right-hand side of (10.17) we use the definition of  $A_2$  and the results of Theorem 10.6 to obtain

$$\mathcal{A}(w, A_2 w) = \alpha_2(\mathcal{A}_2 w, \mathcal{A}_2 w) \geq C_1 \| \mathcal{A}_2 w \|_{H(\text{curl}; \Omega)}^2 \geq C_2 \| w \|_{H(\text{curl}; \Omega)}^2$$

for some positive constants  $C_1$  and  $C_2$ . For the second term on the right-hand side of (10.17) we use the definition of  $A_1$  and recall that  $\alpha_1(p, p)$  is non-positive for any  $p \in S$  to arrive at

$$\begin{aligned} -\mathcal{A}(\nabla p, \nabla A_1 p) &= -\alpha_1(\mathcal{A}_1 p, \mathcal{A}_1 p) \geq C_3 \| \mathcal{A}_1 p \|_{H^1(\Omega)}^2 \geq C_4 \| p \|_{H^1(\Omega)}^2 \\ &\geq C_4 \| \nabla p \|_{(L^2(\Omega))^3}^2 \end{aligned}$$

for some positive constants  $C_3$  and  $C_4$ . The last term on the right hand side of (10.17) is estimated by using the continuity of  $\mathcal{A}$  on  $X \times X$ , and the continuity of  $A_2$  as a map from  $X_0$  to  $X$ :

$$\begin{aligned} |\mathcal{A}(\nabla p, A_2 w)| &\leq C_5 \| \nabla p \|_{H(\text{curl}; \Omega)} \| \mathcal{A}_2 w \|_{H(\text{curl}; \Omega)} \\ &\leq C_6 \| \nabla p \|_{(L^2(\Omega))^3} \| w \|_{H(\text{curl}; \Omega)}. \end{aligned}$$

Putting these estimates in (10.17) gives

$$\mathcal{A}(u, \varphi) \geq C_2 \| w \|_{H(\text{curl}; \Omega)}^2 + \beta C_4 \| \nabla p \|_{(L^2(\Omega))^3}^2 - C_6 \| w \|_{H(\text{curl}; \Omega)} \| \nabla p \|_{(L^2(\Omega))^3}.$$

Using the arithmetic geometric mean inequality, we arrive at

$$\mathcal{A}(u, \varphi) \geq \left( C_2 - \frac{C_6}{2\gamma} \right) \| w \|_{H(\text{curl}; \Omega)}^2 + \left( \beta C_4 - C_6 \frac{\gamma}{2} \right) \| \nabla p \|_{(L^2(\Omega))^3}^2,$$

where  $\gamma$  is another arbitrary positive parameter. We now choose  $\gamma = C_6/C_2$  and  $\beta = C_6^2/(2C_2C_4) + \frac{1}{2}$  and conclude that (10.18)

$$\mathcal{A}(u, \varphi) \geq \frac{1}{2} \min(C_2, C_4) (\| w \|_{H(\text{curl}; \Omega)}^2 + \| \nabla p \|_{(L^2(\Omega))^3}^2).$$

Using Lemma 10.3 to write  $(\| w \|_{H(\text{curl}; \Omega)} + \| \nabla p \|_{(L^2(\Omega))^3}) \geq C \| u \|_{H(\text{curl}; \Omega)}$  and using, in addition, the fact that  $A_j$ ,  $j = 1, 2$ , are invertible we obtain

$$\begin{aligned} & \left( \|w\|_{H(\text{curl}; \Omega)} + \|\nabla p\|_{(L^2(\Omega))^3} \right) \\ & \geq C \left( \|A_2 w\|_{H(\text{curl}; \Omega)} + \|A_1 \nabla p\|_{(L^2(\Omega))^3} \right) \geq C \|v\|_{H(\text{curl}; \Omega)} \end{aligned}$$

and hence from (10.18) we have  $A(u, v) \geq \alpha \|u\|_{H(\text{curl}; \Omega)} \|v\|_{H(\text{curl}; \Omega)}$ . This ends the proof.

## 10.4 The discrete problem

To prove convergence by the method used in this chapter, we need now to make a severe restriction on the function  $\epsilon_r$ . We assume  $\epsilon_r \in C^1(\mathbb{R}^3 \setminus D)$ . This will be a standing assumption for the remainder of this chapter. A slight generalization of this assumption is used by Hiptmair [164] (in particular, he assumes that  $\epsilon_r$  is uniformly Lipschitz continuous) in his analysis of cavity problems. It would be desirable to show that the necessary properties of the finite element spaces can be obtained for less smooth refractive index  $\epsilon_r$ .

We shall discretize (10.3) using the usual family of edge finite element subspaces  $X_h \subset X$ , where the parameter  $h$  measures the maximum diameter of the elements in the associated finite element mesh. We assume that the mesh  $\tau_h$  is regular. We can use edge elements on tetrahedra (see Chapters 5 and 8) or on hexahedra (see Chapter 6). However, for definiteness, we shall assume the use of tetrahedra. Since the outer boundary of  $\Omega$  is not a polyhedron, we assume that the method of Section 8.3.2 has been used to obtain an exact curvilinear finite element covering of  $\Omega$ . As usual, associated with  $X_h$  is a scalar space  $S_h \subset S$  such that  $\nabla S_h \subset X_h$ . The most important implication of these assumptions is that the commuting diagram (5.59) holds. Note that apart from the mapping used to fit the curved boundary  $\Sigma$ , the spaces  $X_h$  and  $S_h$  are exactly the same as those used in Chapter 7.

As we have seen in Chapter 7, it is important to understand the Helmholtz decomposition of discrete functions. We need to discuss this further since  $X_h$  is not the same space as  $X_0$  used in Chapter 7. By virtue of Lemma 10.3, any vector  $u_h \in X_h$  can be written as (10.19)

$$u_h = \omega^h + \nabla p^h$$

for unique  $w^h \in X_0$  and  $p^h \in S$ .

Since  $\nabla \times u_h = \nabla \times w^h$  and  $A(w^h, \nabla \xi) = 0$  for all  $\xi \in S$ , the function  $w^h \in X_0$  is a weak solution of (10.20)

$$\begin{aligned} \nabla \times \omega^h &= \nabla \times u_h \quad \text{in } \Omega, \\ \nabla \cdot (\epsilon_r \omega^h) &= 0 \quad \text{in } \Omega, \end{aligned} \tag{10.21}$$

$$\nabla \times \omega^h = 0 \quad \text{on } \Gamma, \tag{10.22}$$

$$\hat{x} \cdot \omega^h = -\frac{i}{k} \nabla_\Sigma \cdot G_e(\hat{x} \times \omega^h) \quad \text{on } \Sigma. \tag{10.23}$$

In Lemma 10.14 below, we shall show that  $w^h \in (H^{1/2+\delta}(\Omega))^3$  for some  $\delta > 0$  provided  $\epsilon_r \in C^1(\mathbb{R}^3 \setminus D)$ . With this regularity, Lemma 5.38 implies that the edge finite element interpolant  $r_h w^h$  is well defined and Lemma 5.41 (or the equivalent lemma for other element types) shows that (10.24)

$$\|\omega^h - r_h \omega^h\|_{(L^2(\Omega))^3} \leq Ch^{1/2+\delta} \|u_h\|_{H(\text{curl}; \Omega)}$$

for some  $\delta > 0$ .

We shall need one more operator. We recall the projection operator  $P_b$  defined in (7.10). From the density of  $X_b$  in  $\tilde{X}$  we obtain the following result.

**Lemma 10.10** *For any  $\varphi \in \tilde{X}$ , (10.25)*

$$\lim_{h \rightarrow 0} \|P_h \varphi - \varphi\|_{H(\text{curl}; \Omega)} = 0.$$

**Proof** By Cea's lemma applied to (7.10), if  $\varphi \in (H^2(\Omega))^3 \cap \tilde{X}$  then (10.26)

$$\begin{aligned} \|P_h \varphi - \varphi\|_{H(\text{curl}; \Omega)} &\leq C_1 \|r_h \varphi - \varphi\|_{H(\text{curl}; \Omega)} \\ &\leq C_2 h \|\varphi\|_{(H^2(\Omega))^3} \xrightarrow{h \rightarrow 0} 0 \end{aligned}$$

where  $r_b$  is the edge element interpolation operator.

Furthermore, since  $P_b$  is an orthogonal projection, it is uniformly bounded in  $b$ . From the definition of  $P_b$  it follows that

$$(P_h \varphi, P_h \varphi)_{H(\text{curl}; \Omega)} = (\varphi, P_h \varphi)_{H(\text{curl}; \Omega)} \quad \text{for all } \varphi \in \tilde{X}$$

and thus  $\|P_b \varphi\|_{H(\text{curl}; \Omega)} \leq C \|\varphi\|_{H(\text{curl}; \Omega)}$  for all  $\varphi \in H(\text{curl}; \Omega)$  where  $C$  does not depend on  $b$ . The density of  $(C_0^\infty(\mathbb{R}^3 \setminus \bar{D}))^3$  in  $\tilde{X}$ , which follows from the extension Theorem 3.34 and the use of a cutoff function, implies that any given  $\varphi \in \tilde{X}$  can be approximated to arbitrary accuracy by a smooth function, which can in turn be approximated to arbitrary accuracy by a finite element function. This completes the proof.  $\square$

The finite element approximation of the solution  $E \in \tilde{X}$  of (10.2) is  $E_b \in X_b$ , which satisfies (10.27)

$$A(E_b, \varphi_h) = B(\varphi_h) \quad \text{for all } \varphi_h \in X_h.$$

We need to show that  $E_b$  is well defined and has good approximation properties. In order to prove this result, we now prove the following basic Babuška–Brezzi condition:

**Lemma 10.11** *Under the assumption that  $\varepsilon_r \in C^1(\mathbb{R}^3 \setminus D)$  and the assumptions of Theorem 10.2, there exists  $\alpha^* > 0$  such that, for all  $b$  sufficiently small independent of  $u_b$ ,*

$$\sup_{\varphi_h \in X_h} \frac{A(u_h, \varphi_h)}{\|\varphi_h\|_{H(\text{curl}; \Omega)}} \geq \tilde{\alpha} \|u_h\|_{H(\text{curl}; \Omega)} \quad \text{for all } u_h \in X_h.$$

**Proof** Given  $u_b \in X_b$  we first decompose  $u_b$  using the continuous Helmholtz decomposition in Lemma 10.3. Thus we write  $u_b = u^b + {}^\top p^b$  with  $p^b \in S$  and  $u^b \in X_0$ . Let  $P_b : \tilde{X} \rightarrow X_b$  be the orthogonal projection analyzed in Lemma 10.10. We set

$$\varphi_h = P_h \left( A_2 \omega^h - \beta \nabla A_1 p^h \right)$$

with  $\beta > 0$  independent of  $u_b$  taking the value chosen in the proof of Lemma 10.9.

Let  $\varphi = A_2 \mathbf{v}^h - \beta \nabla A_1 \mathbf{p}^h$ , which would be the appropriate choice for proving the continuous Babuška–Brezzi condition. From the definition of  $\varphi$ , (10.28)

$$A(u_h, \varphi_h) = A(u_h, \varphi) + A\left(u_h, (P_h - I)\left(A_2 \omega^h - \beta \nabla A_1 p^h\right)\right).$$

From the proof of Lemma 10.9 and the fact that  $P_h$  is uniformly bounded (see the proof of Lemma 10.10)

$$A(u_h, \varphi) \geq \alpha \|u_h\|_{H(\text{curl}; \Omega)} \|\varphi\|_{H(\text{curl}; \Omega)} \geq \alpha_1 \|u_h\|_{H(\text{curl}; \Omega)} \|\varphi\|_{H(\text{curl}; \Omega)}.$$

Now we turn to the second term on the right-hand side of (10.28). We start by using the definition of  $A_1$  and  $A_2$  in (10.16). Then using the fact that

$$\omega^h - \beta \nabla p^h = (1 + \beta) \omega^h - \beta u_h$$

and that  $(P_h - I)v_h = 0$  for any  $v_h \in X_h$ , we can expand  $A_2$  and  $A_1$  to obtain (10.29)

$$\begin{aligned} A\left(u_h, (P_h - I)\left(A_2 \omega^h - \beta \nabla A_1 p^h\right)\right) &= A\left(u_h, (P_h - I)\left(\omega^h - \beta \nabla p^h\right)\right) \\ &\quad + A\left(u_h, (P_h - I)\left(K_2 \omega^h - \beta \nabla K_1 p^h\right)\right) \\ &= (1 + \beta) A\left(u_h, (P_h - I)\left(\omega^h - r_h \omega^h\right)\right) \\ &\quad + A\left(u_h, (P_h - I)\left(K_2 \omega^h - \beta \nabla K_1 p^h\right)\right). \end{aligned}$$

The first term on the right-hand side of this equation is estimated by continuity and the boundedness of  $P_h$

$$\left| A\left(u_h, (P_h - I)\left(\omega^h - r_h \omega^h\right)\right) \right| \leq C \|u_h\|_{H(\text{curl}; \Omega)} \|\omega^h - r_h \omega^h\|_{H(\text{curl}; \Omega)}.$$

But, using the same argument as in the proof of equation (7.16) of Lemma 7.6, we have  $\nabla \times r_h \mathbf{p}^h = \nabla \times \mathbf{p}^h$  and hence  $\|\mathbf{p}^h - r_h \mathbf{p}^h\|_{H(\text{curl}; \Omega)} = \|\mathbf{p}^h - r_h \mathbf{p}^h\|_{L^2(\Omega)}$ . Thus, using (10.24), we complete the estimation of the first term on the right-hand side of (10.29): (10.30)

$$\left| A\left(u_h, \left(\omega^h - r_h \omega^h\right)\right) \right| \leq Ch^{1/2+\delta} \|u_h\|_{H(\text{curl}; \Omega)}^2.$$

To estimate the second term on the right-hand side of (10.29), we note that the operators  $K_2$  and  $K_1$  are compact in  $X_h$  and  $S$ , respectively. Therefore,  $\nabla K_1$  is compact from  $S$  into  $H(\text{curl}; \Omega)$ . Since  $\|P_h \varphi - \varphi\|_{H(\text{curl}; \Omega)}$  converges to zero for every  $\varphi \in H(\text{curl}; \Omega)$  (see (10.25)), the convergence of  $(P_h - I)K_2 \rightarrow 0$  in the operator norm of maps from  $X_h$  into  $H(\text{curl}; \Omega)$  and  $(P_h - I)\nabla K_1 \rightarrow 0$  in the operator norm of maps from  $S$  to  $H(\text{curl}; \Omega)$  follows (see Lemma 2.50).

Estimate (10.30) together with the above considerations shows that from (10.29) we have

$$\begin{aligned} \left| A\left(u_h, (P_h - I)\left(A_2 \omega^h - \beta \nabla A_1 p^h\right)\right) \right| &\leq C_h \|u_h\|_{H(\text{curl}; \Omega)} \left( \|\omega^h\|_{H(\text{curl}; \Omega)} + \|\nabla p^h\|_{H(\text{curl}; \Omega)} \right) \\ &\leq C_h \|u_h\|_{H(\text{curl}; \Omega)}^2 \xrightarrow{\longrightarrow} 0 \text{ as } h \longrightarrow 0 \end{aligned}$$

Altogether we have proved an estimate of the following form (where  $C_b \rightarrow 0$  as  $b \rightarrow 0$ )

$$|A(E_h, \varphi_h)| \geq C \|u_h\|_{H(\text{curl}; \Omega)}^2 - C_h \|u_h\|_{H(\text{curl}; \Omega)}^2 \geq C_h \|u_h\|_{H(\text{curl}; \Omega)}^2,$$

where the last inequality follows if  $b$  is small enough.

From the definition of  $\varphi_b$  and the boundedness of  $A_j$  ( $j = 1, 2$ ) and  $P_b$ , we conclude that

$$\|\varphi_h\|_{H(\text{curl}; \Omega)} \leq C \left( \|\omega^h\|_{H(\text{curl}; \Omega)} + \|p^h\|_{H^1(\Omega)} \right) \leq C \|u_h\|_{H(\text{curl}; \Omega)}.$$

This finally proves that

$$|A(u_h, \varphi_h)| \geq C \|u_h\|_{H(\text{curl}; \Omega)} \|\varphi_h\|_{H(\text{curl}; \Omega)},$$

from which the lemma follows.  $\square$

Using the above Babuška–Brezzi condition, we can prove the following error estimate

**Theorem 10.12** *The discrete solution  $E_b \in X_b$  is well defined provided  $b$  is small enough and satisfies the error estimate*

$$\|E - E_b\|_{H(\text{curl}; \Omega)} \leq C \inf_{u_h \in X_h} \|E - u_h\|_{H(\text{curl}; \Omega)}.$$

**Remark 10.13** *If  $E$  is smooth enough, this theorem gives an optimal error estimate. For example, if Nédélec's first-type or second-type finite elements of lowest order are used [233, 234], and provided the solution is smooth enough, we can prove that*

$$\|E - E_b\|_{H(\text{curl}; \Omega)} \leq Ch \left( \|E\|_{(H^1(\Omega))^3} + \|\nabla \times E\|_{(H^1(\Omega))^3} \right).$$

*Higher-order elements will give rise to higher-order convergence rates in the obvious way, provided the solution  $E$  is smooth.*

*In practice, the infinite sum in (9.71) would need to be truncated. This would produce an extra error in the computed solution. The error analysis for the corresponding problem for the Helmholtz equation may be found in [188], and an error analysis for truncation in  $G_e$  but not using finite elements is in [121]. In the next chapter, we shall analyze a similar method with discretization using finite elements and a truncation of  $G_e$ .*

*This theorem is just the beginning for a practical use of finite elements to approximate the scattering problem. The theorem guarantees that, up to the constant  $C$ , the computed solution  $E_b$  will optimally approximate the true solution  $E$ . But, of course, the actual error will depend critically on the mesh. Thus, the mesh will need to be refined towards edges and vertices of the domain  $D$ . For re-entrant edges, the design of such meshes, and the necessary estimates to accompany the analysis of refined meshes, can be found in [238]. As far as I am aware, there are no corresponding results for re-entrant vertices.*

**Proof of Theorem 10.12** Suppose that (10.27) has a solution  $E_h \in X_h$ . Let  $u_h \in X_h$  and  $\varphi_h \in X_h$ . From  $A(E_h, \varphi_h) = B(\varphi_h) = A(E, \varphi_h)$  and Lemma 10.11, we conclude that

$$\begin{aligned} \tilde{\alpha} \|E_h - u_h\|_{H(\text{curl}; \Omega)} &\leq \sup_{\varphi_h \in X_h \setminus \{0\}} \frac{|A(E_h - u_h, \varphi_h)|}{\|\varphi_h\|_{H(\text{curl}; \Omega)}} \\ &= \sup_{\varphi_h \in X_h \setminus \{0\}} \frac{|A(E - u_h, \varphi_h)|}{\|\varphi_h\|_{H(\text{curl}; \Omega)}} \leq c \|E - u_h\|_{H(\text{curl}; \Omega)}. \end{aligned}$$

The triangle inequality yields

$$\begin{aligned} \|E - E_h\|_{H(\text{curl}; \Omega)} &\leq \|E - u_h\|_{H(\text{curl}; \Omega)} + \|u_h - E_h\|_{H(\text{curl}; \Omega)} \\ &\leq \left(1 + \frac{c}{\tilde{\alpha}}\right) \|E - u_h\|_{H(\text{curl}; \Omega)}. \end{aligned}$$

Now we prove that there is at most one solution to (10.27). It suffices to show that  $E_h = 0$  is the only solution when  $B(\cdot) = 0$ . But, when  $B(\cdot) = 0$ , the only solution of the continuous problem (10.3) is  $E = 0$ . Using the previously established *a priori* estimate, for any  $u_h \in X_h$ ,

$$\|E_h\|_{H(\text{curl}; \Omega)} \leq \left(1 + \frac{c}{\tilde{\alpha}}\right) \|u_h\|_{H(\text{curl}; \Omega)},$$

which implies  $E_h = 0$ . Hence, uniqueness is verified. Since (10.27) is a linear system with as many equations as unknowns, this also implies the existence of  $E_h$  for a general  $B(\cdot, \cdot)$  and completes the proof.  $\square$

It remains to prove the auxiliary lemma concerning the smoothness of  $\omega^h$ .

**Lemma 10.14** Under the assumptions of this section,  $\omega^h \in (H^{1/2+\delta}(\Omega))^3$  for some  $\delta > 0$ , and

$$\|\omega^h\|_{(H^{1/2+\delta}(\Omega))^3} \leq C \|\nabla \times u_h\|_{(L^2(\Omega))^3}.$$

**Proof** Let  $W \in H_{\text{loc}}(\text{curl}; \mathbb{R}^3 \setminus \bar{B}_R)$  satisfy

$$\begin{aligned} \nabla \times \nabla \times W - \kappa^2 W &= 0 \quad \text{on } \mathbb{R}^3 \setminus \bar{B}_R, \\ \nu \times W &= \nu \times \omega^h \quad \text{on } \Sigma, \end{aligned}$$

together with the Silver–Müller radiation condition. Now let  $\omega^e$  be defined by

$$\omega^e = \begin{cases} \omega^h & \text{on } \Omega, \\ W & \text{on } \mathbb{R}^3 \setminus \bar{B}_R. \end{cases}$$

Then, because of the continuity of the tangential component across  $\Sigma$ , we have  $\omega^e \in H_{\text{loc}}(\text{curl}; \mathbb{R}^3 \setminus D^-)$ . In addition, using exactly the same argument as in the proof of Lemma 10.4, we see that the normal component of  $\omega^e$  is continuous

across  $\sum$ . Since  $\varepsilon_r w^h$  is divergence free in  $D$  and  $W$  is divergence free in  $\mathbb{R}^3 \setminus D^-$ , we can be sure that  $\nabla \cdot (\varepsilon_r w^h) = 0$  in  $\mathbb{R}^3 \setminus D^-$ . Now take a cutoff function  $\chi \in C_0^\infty(\mathbb{R}^3)$  that is unity on  $\Omega$ . Let  $B_{R_1}$  be a ball of radius  $R_1$  large enough to contain the support of  $\chi$ . The function  $\chi w^h \in H_0(\text{curl}; B_{R_1}) \cap H(\text{div}; B_{R_1})$  and so, by the regularity result in Theorem 3.50, we know that  $\chi w^h \in (H^{1/2+\delta}(B_{R_1}))^3$  for some  $\delta > 0$ . Using the fact that  $\varepsilon_r w^h + (\nabla \varepsilon_r) \cdot w^h = \nabla \cdot (\varepsilon_r w^h) = 0$ , and the *a priori* estimate for solutions of the scattering problem in Theorem 10.8 we have using also Theorem 3.50, (10.31)

$$\|\omega^h\|_{(H^{1/2+\delta}(\Omega))^3} \leq C \left( \|\nabla \times \omega^h\|_{(L^2(\Omega))^3} + \|\omega^h\|_{(L^2(\Omega))^3} \right).$$

But using Theorem 10.2, we see that the problem (10.20)–(10.23) has at most one solution and since, by Lemma 10.4,  $X_0$  is compactly imbedded in  $(L^2(\Omega))^3$ , the proof of Corollary 3.51 shows that there is a constant  $C$  such that for all  $v \in X_0$

$$\|v\|_{(L^2(\Omega))^3} \leq C \|\nabla \times v\|_{(L^2(\Omega))^3}.$$

Use of this estimate in (10.31) proves the desired *a priori* estimate and completes the proof of the lemma.  $\square$

# 11SCATTERING BY A BOUNDED IN HOMOGENEITY

## 11.1 Introduction

In the previous chapter we reduced the scattering problem to a problem posed on a bounded domain using an appropriate Calderon operator. The resulting finite element problem is not completely discrete, since we assume that the exterior Calderon operator is computed exactly. Here we shall show how to avoid this by decoupling the problem into discrete interior and exterior problems using a Lagrange multiplier to enforce the desired continuity of the solution on the artificial boundary. The problem we shall solve is that given in (1.26)–(1.29), with the important modification that there is no perfect conductor present (so  $D = \emptyset$ ). This very much simplifies the presentation of the analysis. Of course, in practice the numerical method we shall describe can be used when a perfect conductor is present. This chapter is mainly derived from [90].

The motivation for this problem is to compute the interaction of microwave radiation with a dispersive medium (such as biological tissue). In this case we can assume that the scatterer is a bounded inhomogeneous conductor with a potentially complicated distribution of permittivity and conductivity (the permeability is usually constant in this application, but we shall allow it to vary). The scatterer is assumed to be illuminated by a time-harmonic microwave source (e.g. from a cell-phone antenna). The microwave source produces an incident electromagnetic field that interacts with the scatterer and produces a scattered time-harmonic electromagnetic field.

Using a domain decomposition approach, we employ finite elements to discretize in the vicinity of the scatterer and an approximate Calderon operator, expressed as a truncated series, to discretize the exterior domain. The idea of using a truncated special function expansion to approximately model the exterior domain has been used effectively for the Helmholtz equation (see, e.g. [182, 144, 168, 153]). A similar approach to obtaining artificial boundary conditions for Maxwell's equations has been proposed by Grote and Keller [154] using the special function expansion of the solution that we use in this chapter. They show how to implement the method efficiently for the time-dependent problem. For more recent results in this direction (again for the time domain problem), see [152]. Our results show that, for the time-harmonic problem, the use of truncated spherical harmonic expansions on the artificial boundary produces a well-posed discrete problem (under the conditions of Theorem 11.17) that converges to the exact solution.

Of course, we already know, from the previous chapter, that the problem

(1.26)–(1.29) has a unique variational solution provided the coefficients satisfy the conditions in Section 4.2. The theory presented here will produce an alternative proof of this fact.

The layout of the chapter is as follows. In Section 11.2, we formulate the domain decomposition scheme for the continuous problem. Then we show that the domain-decomposed problem possesses a unique solution. To do this, we view the scattering problem as a compact perturbation of the free-space problem in a suitable sense. We also show that, despite the fact that we do not explicitly handle the divergence condition explicitly, the solution is unique.

Section 11.3 is devoted to describing the finite-dimensional discrete problem based on using the edge finite elements from Chapter 5 in the interior and a truncated Fourier series on the surface of the sphere. In Section 11.4, we analyze the interior finite element problem and derive an error estimate for the interior scheme. Finally, in Section 11.5, we analyze the overall discrete problem, prove that it possesses a unique solution and derive an error estimate. The analysis of this problem is complicated by the fact that we have been unable to write the boundary Calderon operator as a compact perturbation of a coercive operator (see also [116]). Thus, we have to adopt a more general splitting, writing the operator as an invertible operator plus a compact perturbation. This is possible because of the very special boundary space that we use on the artificial boundary.

## 11.2 Derivation of the domain-decomposed problem

Before we show how to reduce the scattering problem to a problem posed on a bounded domain, we shall make explicit the assumptions on the coefficients  $\epsilon_r$  and  $\mu_r$ . Later, in the section on numerical analysis, we shall further restrict the class of coefficients to enable us to prove error estimates. The basic assumptions are given in Section 4.2. Here we make one more assumption that is reasonable in the biological context. We assume that  $\Im(\epsilon_r) \neq 0$  on some subdomain in the scatterer. This is used to guarantee that a suitable interior problem has a unique solution.

Let  $B_R$  be a ball of radius  $R$  containing the scatterer in its interior (i.e. there exists  $a < R$  such that  $\epsilon_r(x) = \mu_r(x) = 1$  for  $|x| > a$ ). The computational domain  $\Omega = B_R$ . The artificial boundary is  $\Sigma = \partial B_R$  and  $\Gamma = \emptyset$ . Inside  $\Omega$ , the electric field satisfies the Maxwell system

$$\nabla \times \mu_r^{-1} \nabla \times E - k^2 \in_r E = 0.$$

Outside  $\Omega$ , in  $\mathbb{R}^3 \setminus \overline{\Omega}$ , the scattered field  $E^s$  satisfies the following constant coefficient Maxwell system together with the Silver–Müller radiation condition:

$$\begin{aligned} \nabla \times \nabla \times E^s - k^2 E^s &= 0 \quad \text{in } \mathbb{R}^3 \setminus \overline{\Omega}, \\ \lim_{\rho \rightarrow \infty} \rho (\nabla \times E^s) \times \hat{x} - ik E^s &= 0. \end{aligned}$$

As in the previous chapter, across the artificial boundary, these problems are linked by enforcing the continuity of the tangential components of the electric and magnetic fields:

(11.1)

$$\hat{x} \times \frac{1}{ik} \nabla \times E = \hat{x} \times \frac{1}{ik} \nabla \times E^s + \hat{x} \times \frac{1}{ik} \nabla \times E^i \text{ on } \Sigma,$$

(11.2)

$$\hat{x} \times E = \hat{x} \times E^s + \hat{x} \times E^i \text{ on } \Sigma.$$

Here we view  $E$  as defined on  $\Omega$ , and  $E^s$  and  $E^i$  as defined on  $\mathbb{R}^3 \setminus \overline{\Omega}$ .

Next, we want to explicitly decouple the two fields and pose the problem as an operator equation on  $\Sigma$ . We use the interior and exterior Calderon operators denoted  $\mathcal{G}_i$  and  $\mathcal{G}_e$  defined, respectively, by (9.77) and (9.75). Using the two Calderon operators, we see that if  $\lambda = \mathcal{O} \times (1/ix)^\nu \times E$  on  $\Sigma$  then, using the boundary relations (11.1) and (11.2), we have

$$\mathcal{G}_i \lambda - \mathcal{G}_e \left( \lambda - \hat{x} \times \frac{1}{ik} \nabla \times E^i \right) = \hat{x} \times E^i|_\Sigma.$$

Now we can state the problem we wish to solve precisely. Given a function  $f \in H^{-1/2}(\text{Div}; \Sigma)$ , we wish to find  $\lambda \in H^{-1/2}(\text{Div}; \Sigma)$  such that (11.3)

$$(\mathcal{G}_i - \mathcal{G}_e)\lambda = f.$$

As we have seen, in applications to the scattering problem, (11.4)

$$f = \hat{x} \times E^i - \mathcal{G}_e \left( \hat{x} \times \frac{1}{ik} \nabla \times E^i \right).$$

Once we have computed  $f$  via (11.3), we can compute  $E$  on  $\Omega$  by solving (9.78a) and (9.78b). Similarly, we can compute  $E^i$  in  $\mathbb{R}^3 \setminus \overline{\Omega}$  by solving (9.76a)–(9.76c), with  $\lambda$  replaced by  $\lambda - \hat{x} \times (1/ix)^\nu \times E^i$ .

Next, we shall establish the following theorem concerning the continuous Calderon operators for the coupled problem. It is the cornerstone of our later analysis of the numerical method.

**Theorem 11.1** Suppose the coefficients  $\varepsilon_r$  and  $\mu_r$  satisfy the conditions outlined at the beginning of this section and that  $\Omega$  is chosen such that  $\nu$  is not an eigenvalue for the interior magnetic Maxwell eigenvalue problem when  $\varepsilon_r = \mu_r = 1$ . Then,

$$\mathcal{G}_i - \mathcal{G}_e = T + K,$$

where  $T$  is a bounded invertible operator from  $H^s(\text{Div}; \Sigma)$  onto  $H^s(\text{Div}; \Sigma)$ , for any  $s$ , and  $K$  is a compact perturbation.

**Remark 11.2** The interior magnetic Maxwell eigenvalue problem is to find  $E \neq 0$  and  $\nu$  such that

$$\begin{aligned} \nabla \times \nabla \times E &= k^2 E \quad \text{in } \Omega \\ v \times (\nabla \times E) &= 0 \quad \text{on } \Gamma. \end{aligned}$$

It seems odd that, when  $\mu_r = \varepsilon_r = 1$ , the interior Maxwell eigenvalues enter the picture. They do so because of the use of an interior problem as a stepping stone in the analysis. By perturbing this problem, we could avoid this restriction so in fact this theorem holds for all  $\nu > 0$ .

The outline of the proof is as follows. First, we establish the theorem in the case when  $\varepsilon_r = \mu_r = 1$ . Then we show how the result can be extended to the general case. We denote by  $\tilde{\mathcal{G}}_i$  the interior Calderon operator when  $\varepsilon_r = \mu_r = 1$  and use a suitable series solution to establish the result. Some preliminary properties for  $\tilde{\mathcal{G}}_i$ , including its series representation, are given in Lemma 9.26.

**Theorem 11.3** *Under the conditions on  $\sum$  in Lemma 9.26, if  $\varepsilon_r = \mu_r = 1$ , then*

$$\mathcal{G}_e - \tilde{\mathcal{G}}_i = T + K_1$$

where  $T : H^s(\text{Div}; \Sigma) \rightarrow H^s(\text{Div}; \Sigma)$  is bounded and invertible and  $K_1$  is compact for any  $s$ .

**Proof** By Lemmas 9.25 and 9.26, if  $\lambda$  is given by (9.57) then

$$(\mathcal{G}_e - \tilde{\mathcal{G}}_i)\lambda = \sum_{n=1}^{\infty} \sum_{m=-n}^n \left[ b_n^m \left( \frac{1}{\delta_n} - \frac{1}{\hat{\delta}_n} \right) U_n^m - a_n^m (\delta_n - \hat{\delta}_n) V_n^m \right].$$

Using the asymptotic relations for spherical Hankel and Bessel functions (9.46)–(9.50) and the Wronskian identity (9.51), we can derive the following estimates

$$\begin{aligned} \frac{1}{\delta_n} - \frac{1}{\hat{\delta}_n} &= -\frac{2ikR}{n} \left( 1 + O\left(\frac{1}{n}\right) \right), \\ \delta_n - \hat{\delta}_n &= \frac{2i}{kR} n \left( 1 + O\left(\frac{1}{n}\right) \right). \end{aligned}$$

Hence, if we define the operator  $T$  by (11.5)

$$T\lambda = - \sum_{n=1}^{\infty} \sum_{m=-n}^n \left[ \frac{2ikR}{n} b_n^m U_n^m + \frac{2i}{kR} n a_n^m V_n^m \right],$$

we have derived the desired decomposition.  $\square$

The next step is to extend the above result to the case of a general medium. To do this, we state the following regularity result. In this result we choose  $\varrho$  so that  $a < \varrho < R$ . Then the scatterer is contained in the interior of the ball of radius  $\varrho$ . Let  $\Omega_{\varrho,R}$  denote the annulus  $\{x \mid \varrho < |x| < R\}$  having boundaries  $\Sigma$  and  $\Sigma_\varrho$  where  $\Sigma_\varrho$  is the surface of the sphere of radius  $\varrho$  centered at the origin. We use the space of functions on  $\Sigma$  defined in (9.59).

**Theorem 11.4** *Assume that  $\varrho$  is chosen so that  $\varkappa$  is not a Maxwell eigenvalue for the annulus  $\Omega_{\varrho,R}$  (i.e. the following interior problem possesses a unique solution). Let the operator  $L : \gamma \rightarrow \hat{x} \times E|_{\Sigma}$  be defined by*

*Then  $L$  is bounded from  $H^{-1/2}(\text{Div}; \Sigma)$  into  $H^s(\text{Div}; \Sigma)$  for any  $s$ .*

**Proof** Recall the tangential fields  $U_n^m$  and  $V_n^m$  defined in (9.56). Using the vector basis functions defined in (9.62) and (9.60), we know that

$$E = \sum_{n=1}^{\infty} \sum_{m=-n}^n \left[ (a_n^m M_m^n + \beta_n^m N_m^n) + (\hat{a}_n^m \hat{M}_m^n + \hat{\beta}_n^m \hat{N}_m^n) \right]$$

for suitable constants  $\{a_n^m, \hat{a}_n^m, \beta_n^m, \hat{\beta}_n^m\}$ . For convenience we define

$$\tilde{h}_n(z) = h_n^{(1)}(z) + zh_n^{(1)'}(z), \quad n \geq 0,$$

with a similar expression for  $\tilde{j}_n$ . Then, using the relationships between the boundary and volume basis in (9.66)–(9.67) with similar relationships for the interior fields we obtain that

$$\begin{aligned} \hat{x} \times E &= \sum_{n=1}^{\infty} \sum_{m=-n}^n \left[ a_n^m h_n^{(1)}(kr) U_n^m + \frac{\tilde{h}_n(kr)}{ikr} \beta_n^m V_n^m \right] \\ &\quad + \left[ \hat{a}_n^m j_n(kr) \hat{U}_n^m + \frac{\tilde{j}_n(kr)}{ikr} \hat{\beta}_n^m \hat{V}_n^m \right], \end{aligned}$$

where  $r = \rho$  or  $r = R$ , depending upon which boundary is under consideration. Furthermore, since  $H = (1/ik) \nabla \times E$ , we have

$$H = \sum_{n=1}^{\infty} \sum_{m=-n}^n \left[ (a_n^m N_m^n - \beta_n^m M_m^n) + (\hat{a}_n^m \hat{N}_m^n - \hat{\beta}_n^m \hat{M}_m^n) \right]$$

and hence

$$\begin{aligned} \hat{x} \times H &= \sum_{n=1}^{\infty} \sum_{m=-n}^n \left[ a_n^m h_n^{(1)}(kr) U_n^m + \frac{\tilde{h}_n(kr)}{ikr} \beta_n^m V_n^m \right] \\ &\quad + \left[ \hat{a}_n^m j_n(kr) \hat{U}_n^m + \frac{\tilde{j}_n(kr)}{ikr} \hat{\beta}_n^m \hat{V}_n^m \right] \end{aligned}$$

where  $r = R$  or  $r = \rho$ , depending upon if we are at the inner or outer boundary of the annular region.

We determine the coefficients  $a_n^m, \beta_n^m, \hat{a}_n^m$  and  $\hat{\beta}_n^m$ ,  $m = -n, \dots, n$  and  $n = 1, 2, \dots$  from the boundary conditions. Suppose

$$\lambda = \sum_{n=1}^{\infty} \sum_{m=-n}^n [a_n^m U_n^m + b_n^m V_n^m].$$

Using the boundary condition on  $r = \rho$ , we have

$$\begin{aligned} a_n^m h_n^{(1)}(k\rho) + \hat{a}_n^m j_n(k\rho) &= a_n^m / \sqrt{n(n+1)}, \\ \beta_n^m \tilde{h}_n(k\rho) + \hat{\beta}_n^m \tilde{j}_n(k\rho) &= ik\rho b_n^m / \sqrt{n(n+1)}. \end{aligned}$$

On the boundary  $r = R$ , we have the vanishing tangential component of the magnetic field. Hence

$$a_n^m \tilde{h}_n(kR) + \hat{a}_n^m \tilde{j}_n(kR) = 0,$$

$$\beta_n^m h_n^{(1)}(kR) + \hat{\beta}_n^m j_n(k\rho) = 0.$$

These two systems can be solved for the unknown coefficients to yield

$$\begin{aligned} a_n^m &= \frac{1}{D_1} \frac{a_n^m \tilde{j}_n(kR)}{\sqrt{n(n+1)}}, \quad \hat{a}_n^m = -\frac{1}{D_1} \frac{a_n^m \tilde{h}_n(kR)}{\sqrt{n(n+1)}}, \\ \beta_n^m &= \frac{1}{D_2} \frac{i k \rho b_n^m j_n(kR)}{\sqrt{n(n+1)}}, \quad \hat{\beta}_n^m = -\frac{1}{D_2} \frac{i k \rho b_n^m h_n^{(1)}(kR)}{\sqrt{n(n+1)}}, \end{aligned}$$

where

$$D_1 = h_n^{(1)}(k\rho) \tilde{j}_n(kR) - \tilde{h}_n(kR) j_n(k\rho),$$

$$D_2 = \tilde{h}_n(k\rho) j_n(kR) - \tilde{j}_n(k\rho) h_n^{(1)}(kR).$$

Now using the asymptotic estimates (9.46)–(9.50), we can easily show that

$$\begin{aligned} L\lambda &= \sum_{n=1}^{\infty} \sum_{m=n}^n [a_n^m h_n(kR) + \hat{a}_n^m j_n(kR)] U_n^m \\ &\quad + \sum_{n=1}^{\infty} \sum_{m=n}^n [b_n^m \tilde{h}_n(kR) + \hat{b}_n^m \tilde{j}_n(kR)] V_n^m \\ &= \sum_{n=1}^{\infty} \sum_{m=-n}^n \left[ \left(\frac{\rho}{R}\right)^n (1 + \lambda_n) a_n^m U_n^m + \left(\frac{\rho}{R}\right)^{n+2} (1 + \mu_n) b_n^m V_n^m \right], \end{aligned}$$

where  $\lambda_n, \mu_n = O(1/n)$ . The fact that  $\rho < R$  then implies the necessary decay in the coefficients in this expansion.  $\square$

Now we consider the interior problem with general coefficients. It might appear that this case is covered by the theory of Chapter 4, but this is ruled out by our assumption that  $\lambda > 0$ . Our results here extend the previous theory to the case where  $\lambda$  is identically zero (at least in the case  $D = \emptyset$ , but this assumption can be removed). The approach to proving existence is very like the one used in Chapter 4 and we only sketch it here.

Given  $\lambda \in H^{-1/2}(\text{Div}; \Sigma)$ , we recall that we can define the operator

$$\mathcal{G}_{\mathbf{i}} : H^{-1/2}(\text{Div}; \Sigma) \rightarrow H^{-1/2}(\text{Div}; \Sigma)$$

by  $\mathbf{g}\lambda = \mathcal{O} \times w|_{\Sigma}$ , where  $w \in H(\text{curl}; \Omega)$  satisfies (9.78a) and (9.78b). To obtain a variational formulation of this problem suitable for later finite element

discretization, we can multiply (9.78a) by a test function  $\varphi \in H(\text{curl}; \Omega)$  and integrate by parts (using (9.78b) for the boundary term) to obtain(11.6)

$$\left( \mu_r^{-1} \nabla \times u, \nabla \times \varphi \right) - k^2 (\epsilon_r u, \varphi) + i k \langle \lambda, \varphi \rangle = 0$$

for all  $\varphi \in H(\text{curl}; \Omega)$ . In order to show that  $\mu_r \lambda$  is well defined and maps  $H^{-1/2}(\text{Div}; \Sigma)$  into  $H^{-1/2}(\text{Div}; \Sigma)$  it suffices to show that a unique solution of (11.6) in  $H(\text{curl}; \Omega)$  exists. As usual for problems of this type we do this in two steps. First we show uniqueness and then use the Fredholm alternative to obtain existence.

**Lemma 11.5** *Assume that  $\epsilon_r$  and  $\mu_r$  satisfy the conditions of Section 4.2 and, in addition, that  $\Im(\epsilon_r) \neq 0$  in some subdomain of  $\Omega$ . Then problem(11.6) has at most one solution.*

**Proof** By linearity, it suffices to consider the case when  $\lambda = 0$  and choose  $\varphi = u$  in (11.6). Then

$$\left( \mu_r^{-1} \nabla \times u, \nabla \times u \right) - k^2 (\epsilon_r u, u) = 0$$

and hence (since  $\mu_r$  is real symmetric)  $\Im(\epsilon_r u, u) = 0$ . This implies that  $u = 0$  in every subdomain in which  $\Im(\epsilon_r) \neq 0$  (at least one such subdomain exists by assumption). Now using the unique continuation result of Theorem 4.13 as in the proof of Theorem 4.12, we conclude that  $u = 0$  in  $\Omega$ .  $\square$

Next we prove that  $u$  exists, using the Fredholm alternative. To do this, we first have to prove a compact imbedding result. We know that  $\nabla H^1(\Omega)/\mathbb{R}$  is a closed subspace of  $H(\text{curl}; \Omega)$ . Hence, if we define  $\hat{S} = H^1(\Omega)/\mathbb{R}$ , we have the Helmholtz decomposition(11.7)

$$H(\text{curl}; \Omega) = \hat{X}_0 \oplus \nabla \hat{S},$$

where(11.8)

$$\hat{X}_0 = \left\{ v \in H(\text{curl}; \Omega) \mid (\epsilon_r v, \nabla q) = 0 \text{ for all } q \in \hat{S} \right\}.$$

This space plays the role that  $X_0$  played in Chapter 4 . The following result is well known (see, e.g. [207]).

**Theorem 11.6** *The divergence free space  $\hat{X}_0$ defined in(11.8)is compactly imbedded in  $(L^2(\Omega))^3$ .*

**Proof** When  $\epsilon_r = \mu_r = 1$ , this result follows from the second part of Costabel's regularity result in Theorem 3.47. For general  $\epsilon_r$  and  $\mu_r$  satisfying the conditions of Section 4.2, a slight modification of the proof of Theorem 4.7 proves the result. We do not provide the details (for another proof of this result, see [159]).  $\square$

Now suppose that  $u \in \hat{X}_0$  and  $\nabla \times u = 0$  in  $\Omega$ . Since  $u$  is curl free, there is a function  $p \in \hat{S}$  such that  $u = \nabla p$  and, since  $u \in \hat{X}_0$ , we have  $(\epsilon_r p, \nabla p) = 0$ . The positive definiteness of the real part of  $\epsilon_r$  now implies  $p = 0$ . Combining this uniqueness result with the above compactness result implies the following corollary:

**Corollary 11.7** There exists a constant  $C > 0$  such that, for all  $u \in X_0$ ,

$$\|u\|_{(L^2(\Omega))^3} \leq C \|\nabla \times u\|_{(L^2(\Omega))^3}.$$

Using these results, we can prove the promised existence result.

**Theorem 11.8** There exists a unique solution  $u \in H(\text{curl}; \Omega)$  to the interior boundary value problem (11.6) and hence  $\mathbf{g}_i : H^{-1/2}(\text{Div}; \Sigma) \rightarrow H^{1/2}(\text{Div}; \Sigma)$  is well defined and bounded.

**Proof** For given  $\lambda \in H^{-1/2}(\text{Div}; \Sigma)$ , we define  $p \in \hat{S}$  by (11.9)

$$k^2(\epsilon_r \nabla p, \nabla q) = ik\langle \lambda, \nabla q \rangle \quad \text{forall } q \in \hat{S}.$$

For this problem, existence and uniqueness follow from the Lax–Milgram lemma, since  $\Re(\epsilon_r)$  is uniformly positive definite. Now we make the ansatz (11.10)

$$u = z + \nabla p,$$

where  $z \in H(\text{curl}; \Omega)$  satisfies (11.11)

$$(\mu_r^{-1} \nabla \times z, \nabla \times \varphi) - k^2(\epsilon_r z, \varphi) = -ik\langle \lambda, \varphi \rangle + k^2(\epsilon_r \nabla p, \varphi)$$

for all  $\varphi \in H(\text{curl}; \Omega)$ . By choosing  $\varphi = \nabla q$  for an arbitrary  $q \in \hat{S}$ , we see that  $(\epsilon_r z, \nabla q) = 0$  and thus  $z \in X_0$ . Since  $H(\text{curl}; \Omega)$  is the direct sum of  $X_0$  and  $\nabla \hat{S}$ , we can rewrite (11.11) as the problem of finding  $z \in X_0$  such that (11.12)

$$(\mu_r^{-1} \nabla \times z, \nabla \times \varphi) - k^2(\epsilon_r z, \varphi) = -ik\langle \lambda, \varphi \rangle + k^2(\epsilon_r \nabla p, \varphi)$$

for all  $\varphi \in X_0$ . By Corollary 11.7, the first term on the left-hand side of (11.12) defines a bounded and coercive sesquilinear form on  $X_0$ . Hence we can define the operator  $\mathcal{B} : (L^2(\Omega))^3 \rightarrow (L^2(\Omega))^3$  such that for  $f \in (L^2(\Omega))^3$  we require  $\mathcal{B}f \in X_0 \subset (L^2(\Omega))^3$  to be the solution of (11.13)

$$(\mu_r^{-1} \nabla \times Bf, \nabla \times \varphi) = -k^2(\epsilon_r f, \varphi)$$

for all  $\varphi \in X_0$ . The operator  $\mathcal{B}$  is compact since it is continuous from  $(L^2(\Omega))^3$  into  $X_0$  and  $X_0$  is compactly embedded in  $(L^2(\Omega))^3$  (see Theorem 11.6).

We now define  $F \in X_0 \subset (L^2(\Omega))^3$  to be the solution of (11.14)

$$(\mu_r^{-1} \nabla \times F, \nabla \times \varphi) = -ik\langle \lambda, \varphi \rangle + k^2(\epsilon_r \nabla p, \varphi)$$

for all  $\varphi \in X_0$ . Then the original problem is equivalent to finding  $z \in (L^2(\Omega))^3$  such that

$$(i + \mathcal{B})z = F$$

and the existence of a solution to this problem (and hence to the original problem) follows from the Fredholm alternative (Theorem 2.33) and the uniqueness result proved in Lemma 11.5. Using the same argument as in the discussion following the proof of Theorem 4.11, we can then conclude that  $z \in X_0$  with an appropriate *a priori* estimate.

Now that we have verified the existence of the operator  $\mathbf{g}_i$ , we can prove Theorem 11.1.

**Proof of Theorem 11.1** Using Theorem 11.3 and the definition of  $\tilde{\mathbf{g}}_i$  and  $\mathbf{g}_i$ , we can write

$$\mathcal{G}_e \lambda - \mathcal{G}_i \lambda = (\mathcal{G}_e - \tilde{\mathcal{G}}_i) \lambda + (\tilde{\mathcal{G}}_i - \mathcal{G}_i) \lambda = T \lambda + K_1 \lambda + \hat{x} \times (\tilde{u} - u),$$

where  $u$  solves (11.6) and  $\tilde{u}$  solves (11.6) with  $\epsilon_r = \mu_r = 1$ . Now if we define  $w = \tilde{u} - u$  then  $w \in H(\text{curl}; \Omega)$  satisfies

$$\left( \mu_r^{-1} \nabla \times \omega, \nabla \times \varphi \right) - k^2 (\epsilon_r \omega, \varphi) = \left( (\mu_r^{-1} - 1) \nabla \times \tilde{u}, \nabla \times \varphi \right) - k^2 ((\epsilon_r - 1) \tilde{u}, \varphi)$$

for all  $\varphi \in H(\text{curl}; \Omega)$ . Using exactly the same argument as in the previous theorem (but with a different right-hand side), we can see that  $w$  is the unique solution of the above variational problem. Now let us choose  $Q < R$  such that  $B_Q$  contains the support of  $(\mu_r - 1)$  and  $(\epsilon_r - 1)$  in its interior. Then, by the trace theorem for  $H(\text{curl}; \Omega)$ , the function  $\mathcal{O} \times w|_{\Sigma_Q} \in H^{-1/2}(\text{Div}; \Sigma_Q)$  is bounded in terms of the curl norm of  $\tilde{u}$  and hence, by Theorem 11.8, in terms of the  $H^{-1/2}(\text{Div}; \Sigma)$  norm of  $\lambda$ . Then, using Theorem 11.4, we conclude that  $\mathcal{O} \times w \in H(\text{Div}; \Sigma)$  for any  $s$ . Hence  $\tilde{\mathbf{g}}_i - \mathbf{g}_i$  is a compact map from  $H^{-1/2}(\text{Div}; \Sigma)$  into  $H^{-1/2}(\text{Div}; \Sigma)$ . We have thus proved Theorem 11.1 since

$$\mathcal{G}_e - \mathcal{G}_i = T + (K_1 + \tilde{\mathcal{G}}_i - \mathcal{G}_i).$$

and  $K = K_1 + \tilde{\mathbf{g}}_i - \mathbf{g}_i$  is compact.  $\square$

Theorem 11.1 can be used to prove the existence of a weak solution of the original scattering problem (of course, this is already known from the previous chapter). To do this, it is necessary to prove that (11.3) has a unique solution.

**Lemma 11.9** Under the conditions of Theorem 11.1, problem (11.3) has at most one solution.

**Proof** By linearity it suffices to consider the case  $f = 0$ . For a given solution  $\lambda$  we define  $u_1$  to satisfy (9.76a)–(9.76c) and define  $u_2$  to satisfy (9.78a)–(9.78b). Existence and uniqueness of  $u_1$  is given by Lemma 9.25 and we have proved the existence and uniqueness of a weak solution of (9.78a)–(9.78b) in Theorem 11.8. Then eqn (11.3) ensures that if we define the function  $u$  by

$$u = \begin{cases} u_1 & \text{in } \mathbb{R}^3 \setminus \bar{\Omega}, \\ u_2 & \text{in } \Omega, \end{cases}$$

then  $u$  is a solution of the Maxwell system (1.26)–(1.29) with vanishing incident field (and  $D = 0$ ). The classical uniqueness result in Theorem 10.1 for the solution of the Maxwell system then shows that  $u = 0$ . Hence  $\lambda = 0$ .  $\square$

Using Theorem 11.1 and the above uniqueness lemma, by the application of the Fredholm alternative (Theorem 2.33), we can prove the following result:

**Theorem 11.10** For every  $f \in H^{-1/2}(\text{Div}; \Sigma)$  there exists a unique solution  $\lambda \in H^{-1/2}(\text{Div}; \Sigma)$  to (11.3) and

$$\|\lambda\|_{H^{-1/2}(\text{Div}, \Sigma)} \leq C \|f\|_{H^{-1/2}(\text{Div}, \Sigma)}.$$

### 11.3 The finite-dimensional problem

In this section we describe the discrete problem derived from (11.3). The idea is to seek an approximation of  $\lambda$  on  $\Sigma$  using the space  $S_N$  defined as follows

$$S_N = \left\{ u \in H^{-1/2}(\text{Div}, \Sigma) \middle| u = \sum_{n=1}^N \sum_{m \leq n} [a_{n,m} U_n^m + \beta_{n,m} V_n^m] \right. \\ \left. \text{with } a_{n,m}, \beta_{n,m} \in \mathbb{C} \right\}.$$

In other words, we seek to approximate  $\lambda$  by a finite Fourier series. We define  $P_N : H^{-1/2}(\text{Div}; \Sigma) \rightarrow S_N$  to be the orthogonal projection onto  $S_N$  in the  $H^{-1/2}(\text{Div}; \Sigma)$  inner product. Due to the orthogonality properties of the basis functions, this is nothing more than the truncation operator (see Section 9.3.3). Of course, for any  $\lambda \in H^{-1/2}(\text{Div}; \Sigma)$

$$P_N \lambda \rightarrow \lambda \text{ in } H^{-1/2}(\text{Div}, \Sigma) \text{ as } N \rightarrow \infty.$$

We also have the following error estimate (11.15)

$$\|(I - P_N)\lambda\|_{H^{-1/2}(\text{Div}, \Sigma)} \leq CN^{-\sigma-1/2} \|\lambda\|_{H^\sigma(\text{Div}, \Sigma)}$$

for any  $\sigma; \geq -1/2$ . This is seen by using the series definition of the norm and elementary manipulations as follows.

$$\begin{aligned} \|(I - P_N)\lambda\|_{H^{-1/2}(\text{Div}, \Sigma)}^2 &\leq C \sum_{n>N} \sum_{m=-n}^n [n|a_n^m|^2 + \frac{1}{n}|b_n^m|^2] \\ &= C \sum_{n>N} n^{-1-2\sigma} \sum_{m=-n}^n [n^{2+2\sigma}|a_n^m|^2 + n^{2\sigma}|b_n^m|^2] \\ &\leq \frac{C}{N^{1+2\sigma}} \sum_{n>N} \sum_{m=-n}^n [n^{2+2\sigma}|a_n^m|^2 + n^{2\sigma}|b_n^m|^2] \\ &\leq \frac{C}{N^{1+2\sigma}} \|\lambda\|_{H^\sigma(\text{Div}, \Sigma)}^2, \end{aligned}$$

where  $\lambda = \sum_{n=1}^{\infty} \sum_{m=-n}^n (a_n^m U_n^m + b_n^m V_n^m)$ . We also note that  $S_N$  satisfies the following inverse estimate, for any  $\lambda_N \in S_N$  (11.16)

$$\|\lambda_N\|_{H^{1/2}(\text{Div}, \Sigma)} \leq CN \|\lambda_N\|_{H^{-1/2}(\text{Div}, \Sigma)}.$$

This is again seen by using the series representation of  $\lambda_N$ :

$$\begin{aligned}\|\lambda\|_{H^{-1/2}(\text{Div}, \Sigma)}^2 &\leq C \sum_{n=1}^N \sum_{m=-n}^n \left[ n^3 |a_n^m|^2 + n |b_n^m|^2 \right] \\ &\leq CN^2 \sum_{n=1}^N \sum_{m=-n}^n \left[ n |a_n^m|^2 + \frac{1}{n} |b_n^m|^2 \right] \\ &= CN^2 \|\lambda_N\|_{H^{-1/2}(\text{Div}, \Sigma)}^2.\end{aligned}$$

Since  $T$ , defined by (11.5), is a diagonal operator when restricted to  $S_N$ , it is easy to see that  $T$  and  $P_N$  commute:

$$P_N T = T P_N.$$

For any function  $\lambda \in S_N$ , the function  $\mathbf{g}_e \lambda_N$  is easy to calculate using the truncation of (9.80).

The interior operator  $\mathbf{g}_i$  also needs to be discretized. For this, we apply the finite element method using the edge elements of Nédélec [233] from Chapter 5 as modified for a spherical domain in Section 8.3.2. We shall limit ourselves to the lowest-order edge space in the remainder of this chapter, so that the Dubois locally mapped elements (see Section 8.3.1) can also be used. If the construction in Section 8.3.2 is used, there is no reason in principle not to use higher-order elements.

Let  $\tau_b$  be a regular curvilinear mesh for  $\Omega$  and let  $V_b$  denote the corresponding space of degree  $k = 1$  edge elements (defined in Section 8.3.2). Of course,  $V_b \subset H(\text{curl}; \Omega)$ . We now wish to discretize  $\mathbf{g}_i$ . For any function  $\lambda \in H^{-1/2}(\text{Div}; \Sigma)$ , we define  $\mathbf{g}_{i,b} \lambda = \mathcal{O} \times u_b$ , where  $u_b \in V_b$  satisfies the discrete analogue of (11.6):(11.17)

$$\begin{aligned}\left( \mu_r^{-1} \nabla \times u_b, \nabla \times \varphi_h \right) - k^2 (\in_r u_b, \varphi_h) + i k \{\lambda, \varphi_h\} &= 0 \\ \text{for all } \varphi_h \in V_h.\end{aligned}$$

In Section 11.4 we shall show that this problem has a unique solution, and derive some error estimates.

Now that we have a discrete analogue of  $\mathbf{g}_i$ , we define the discrete analogue of (11.3) to be the problem of finding  $\lambda_{N,b} \in S_N$  such that (11.18)

$$(P_N \mathcal{G}_{i,h} - \mathcal{G}_e) \lambda_{N,h} = P_N f.$$

The remainder of the chapter is devoted to showing that this problem has a unique solution that converges at an optimal rate to the exact solution.

## 11.4 Analysis of the interior finite element problem

We wish to allow the coefficients  $\epsilon_r$  and  $\mu_r$  to be piecewise smooth. At the interfaces where  $\epsilon_r$  is discontinuous, we know that  $E$  is generally discontinuous.

Similarly, where  $\mu_r$  is discontinuous, we know that  $\nabla \times E$  will generally be discontinuous (irrespective of the smoothness of the interfaces). So we want to extend the function spaces to allow for such discontinuities. Let

$$\begin{aligned} PH^1(\text{curl}; \Omega) = & \left\{ u \in \left(L^2(\Omega)\right)^3 \mid u|_{\Omega_n \cap \Omega} \in \left(H^1(\Omega_n \cap \Omega)\right)^3 \text{ and} \right. \\ & \left. \nabla \times u \Big|_{\Omega_n \cap \Omega} \in \left(H^1(\Omega_n \cap \Omega)\right)^3, n = 0, 1, \dots, N \right\}, \end{aligned}$$

where the domains  $\Omega_n$ ,  $n = 0, 1, \dots, N$  were introduced in Section 4.2. The norm on this space is

$$\|u\|_{PH^1(\text{curl}; \Omega)}^2 = \|u\|_{\left(L^2(\Omega)\right)^3}^2 + \sum_{n=0}^N \left[ \|u\|_{\left(H^1(\Omega_n \cap \Omega)\right)^3}^2 + \|\nabla \times u\|_{\left(H^1(\Omega_n \cap \Omega)\right)^3}^2 \right].$$

In addition, we define, for  $s \in \mathbb{N}$ ,

$$PH^s(\Omega) = \left\{ p \in \left(L^2(\Omega)\right)^3 \mid p|_{\Omega_n \cap \Omega} \in \left(H^s(\Omega_n \cap \Omega)\right)^3 \right\}$$

with the norm

$$\|p\|_{PH^s(\Omega)}^2 = \|p\|_{\left(L^2(\Omega)\right)^3}^2 + \sum_{n=0}^N \left[ \|p\|_{H^s(\Omega_n \cap \Omega)}^2 \right].$$

We must assume that the interfaces where  $\epsilon_r$  or  $\mu_r$  are discontinuous (i.e. between the domains  $\Omega_n$ ,  $n = 0, \dots, N$ ) lie along the faces of the mesh. Of course, this means that either the interfaces are polyhedral, or we have used the methods of Section 8.3.2 to develop a curvilinear mesh in each  $\Omega_n$ .

Using Lemma 5.38, we know that the interpolation operator  $r_h$  corresponding to the  $k = 1$  edge space  $V_h$  is well defined for functions in  $PH^1(\text{curl}; \Omega)$  and the following estimate holds for the curvilinear interpolant:(11.19)

$$\|u - r_h u\|_{\left(L^2(\Omega)\right)^3} + \|\nabla \times (u - r_h u)\|_{\left(L^2(\Omega)\right)^3} \leq Ch \|u\|_{PH^1(\text{curl}; \Omega)}.$$

Of course, the interpolation operator is well defined for much less regular functions but we wish to prove optimal error estimates for which the above smoothness is sufficient.

Let  $\hat{S}_h = U_h / R$ , where  $U_h$  is the mapped curvilinear space of continuous piecewise-linear scalar functions defined in Section 8.3.2. We can then define the space of discrete divergence-free fields to be

$$\hat{X}_{0,h} = \left\{ u_h \in V_h \mid (\epsilon_r u_h, \nabla p_h) = 0 \text{ for all } p_h \in \hat{S}_h \right\}.$$

Now suppose we have a sequence of refinements of the mesh indexed by mesh sizes  $h_1 > h_2 > \dots$ . We assume  $h_n \rightarrow 0$  as  $n \rightarrow \infty$  and set(11.20)

$$\wedge = \{h_n \mid n = 1, 2, \dots\}.$$

We want to show convergence of  $\mathcal{G}_{i,h_n}$  to  $\mathcal{G}_i$  as  $n$  increases. In order to prove this, we proceed as in [119] using a discrete compactness argument. First we give the

discrete compactness result for this case (a generalization of the original result of Kikuchi [185] to variable  $\varepsilon_r$  [71]).

**Theorem 11.11** Suppose  $\{u_n\}_{n=1}^\infty \subset H(\text{curl}; \Omega)$  is a bounded sequence such that for each  $n$  there is an  $m = m(n)$  such that  $u_n \in X_{h_m}$  and  $b_m \rightarrow 0$  as  $n \rightarrow 0$ . Then there is a subsequence, also denoted by  $\{u_n\}_{n=1}^\infty$ , which converges weakly in  $H(\text{curl}; \Omega)$  to a function  $u \in X_0$ , that is,

$$(\mathbf{r}u, \nabla p) = 0 \text{ for all } p \in H^1(\Omega),$$

and  $u_n \rightarrow u$  strongly in  $(L^2(\Omega))^3$ .

**Proof** The proof follows the same outline as the proof of Theorem 7.18. First we prove discrete compactness when  $\varepsilon_r = 1$ . For each  $n$  we define  $u^n \in X_0$  by

$$\begin{aligned} \nabla \times u^n &= \nabla \times u_n \quad \text{in } \Omega, \\ \nabla \cdot u^n &= 0 \quad \text{in } \Omega, \\ u^n \cdot \hat{x} &= 0 \quad \text{on } \Sigma. \end{aligned}$$

Hence, by Theorem 3.47, we know that there is a subsequence of  $\{u^n\}$ , still denoted by  $\{u^n\}$ , and a function  $u \in X_0$  such that  $u^n \rightarrow u$  strongly in  $(L^2(\Omega))^3$  (and weakly in  $H(\text{curl}; \Omega)$ ). But since  $\Omega$  is smooth,  $u^n \in (H^{1/2+\delta}(\Omega))^3$  for some  $\delta > 0$  by Theorem 3.50. In this case  $\mathbf{r}_{h_m} u^n$  is defined (see Lemma 5.38) and in Theorem 8.20 we note that the second estimate of Theorem 5.41 holds for the curvilinear elements.

Now we may write, using the Helmholtz decomposition (11.7),

$$u_n = u^n + \nabla p^n$$

for some  $p^n \in \hat{S}$  and  $u^n \in X_0$ . Using the commuting diagram for the curvilinear space  $u_n = \mathbf{r}_{h_m} u^n + \nabla p_n$  for some  $p_n \in \hat{S}_b$ , and using the fact that  $u \in X_0$  and  $u_n \in X_{h_m}$ , we may write

$$\begin{aligned} (u - u_n, u - u_n) &= (u - u_n, u - \mathbf{r}_{h_m} u^n) \\ &= (u - u_n, u - u^n) + (u - u_n, u^n - \mathbf{r}_{h_m} u^n). \end{aligned}$$

Hence

$$\|u - u_n\|_{(L^2(\Omega))^3} \leq \|u - u^n\|_{(L^2(\Omega))^3} + \|u^n - \mathbf{r}_{h_m} u^n\|_{(L^2(\Omega))^3} \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Hence the discrete compactness result is proved when  $\varepsilon_r = 1$ .

Now we follow [71] as in the proof of Theorem 7.18, allowing for the fact that  $\varepsilon_r$  is complex to complete the proof.  $\square$

Using this result as in the proof of Lemma 7.20 we have the following result which generalizes Lemma 11.7 to the finite element context. We do not give the proof because it is so similar.

**Corollary 11.12** *Provided  $b$  is sufficiently small, there is a constant  $C$  independent of  $b$  and  $u_b$  such that*

$$\|u_h\|_{(L^2(\Omega))^3} \leq C \|\nabla \times u_h\|_{(L^2(\Omega))^3} \text{ for all } u_h \in \widehat{X}_{0,h}.$$

Now we show that  $\mathbf{g}_{i,b}$  is well defined by showing that (11.17) has a unique solution. We proceed along the same lines as in the proof of Theorem 7.25. Let  $p_b \in \widehat{S}_b$  satisfy the discrete analogue of (11.9), (11.21)

$$k^2(\in_r \nabla p_h, \nabla \xi_h) = ik\langle \lambda, \nabla \xi_h \rangle \text{ for all } \xi_h \in \widehat{S}_h.$$

Since the real part of  $\varepsilon_r$  is positive definite and the average value of  $p_b$  is zero, this problem has a unique solution by the Lax–Milgram lemma.

Next we make the ansatz (11.22)

$$u_h = z_h + \nabla p_h,$$

where  $z_b \in X_{0,b}$  satisfies (11.23)

$$\left( \mu_r^{-1} \nabla \times z_h, \nabla \times \varphi_h \right) - k^2(\in_r z_h, \varphi_h) = -ik\langle \lambda, \varphi \rangle + k^2(\in_r \nabla p_h, \varphi_h)$$

for all  $\varphi_b \in X_{0,b}$ . To convert this to an operator equation, we define  $\mathcal{B}_b : (L^2(\Omega))^3 \rightarrow (L^2(\Omega))^3$  by  $\mathcal{B}_b f = w_b \in X_{0,b}$ , where  $w_b$  satisfies

$$\left( \mu_r^{-1} \nabla \times \omega_h, \nabla \times \varphi_h \right) = -k^2(\in_r f, \varphi_h) \text{ for all } \varphi_h \in \widehat{X}_{0,h}.$$

Note that  $\mathcal{B}_b$  is actually a bounded map from  $(L^2(\Omega))^3$  into  $X_{0,b}$ . We also define  $F_b \in X_{0,b} \subset (L^2(\Omega))^3$  as the solution of (11.24)

$$\left( \mu_r^{-1} \nabla \times \mathfrak{F}_h, \nabla \times \varphi_h \right) = -ik\langle \lambda, \varphi \rangle + k^2(\in_r \nabla p_h, \varphi_h)$$

for all  $\varphi_b \in X_{0,b}$ . By Corollary 11.12 and the Lax–Milgram lemma, these problems have a unique solution. Thus, we consider the operator equation for finding  $v \in (L^2(\Omega))^3$  such that (11.25)

$$(i + \mathcal{B}_b)v = \mathfrak{F}_h.$$

Note that if we can uniquely solve this problem then  $v = -\mathcal{B}_b v + F_b \in X_{0,b}$ , so that  $v \in X_{0,b}$ . In addition,

$$\left( \mu_r^{-1} \nabla \times (v + \mathcal{B}_b v), \nabla \times \varphi_h \right) = \left( \mu_r^{-1} \nabla \times \mathfrak{F}_h, \nabla \times \varphi_h \right) \text{ for all } \varphi_h \in \widehat{X}_{0,h}.$$

Now, using the definition of  $\mathcal{B}_b$  and  $F_b$ , we see that  $v$  satisfies (11.23) and so, in fact,  $z_b = v$ . We have the following result:

**Theorem 11.13** *The collection of operators  $\{\mathcal{B}_b\}_{b \in \Lambda}$ , where the discrete set  $\Lambda$  is given by (11.20), converges pointwise to the operator  $\mathcal{B}$  defined in (11.13) in  $(L^2(\Omega))^3$ . In addition, the set of operators  $\{\mathcal{B}_b\}_{b \in \Lambda}$  is collectively compact when considered as maps from  $(L^2(\Omega))^3$  to  $(L^2(\Omega))^3$ .*

**Proof** This proof parallels the proof of similar results in Chapter 7, in particular Theorems 7.11 and 7.14. Thus, to prove pointwise convergence we rewrite the definition of  $\mathcal{B}f$  as a mixed problem. For  $f \in (L^2(\Omega))^3$  we see that  $\mathcal{B}f \in H(\text{curl}; \Omega)$  and  $q \in \hat{S}$  satisfy(11.26)

$$\begin{aligned} & \left( \mu_r^{-1} \nabla \times \mathcal{B}f, \nabla \times \varphi \right) + (\epsilon_r \varphi, \nabla q) = -k^2 (\epsilon_r f, \varphi) \\ & \text{forall } \varphi \in H(\text{curl}; \Omega), \\ & (\epsilon_r \mathcal{B}f, \nabla \xi) = 0 \text{ for all } \xi \in \hat{S}. \end{aligned} \quad (11.27)$$

The second equation ensures that  $\mathcal{B}f \in X_0$ . Similarly, we can see that  $\mathcal{B}f \in V_b$  and  $q_b \in \hat{S}_b$  satisfy(11.28)

$$\begin{aligned} & \left( \mu_r^{-1} \nabla \times \mathcal{B}_h f, \nabla \times \varphi_h \right) + (\epsilon_r \varphi_h, \nabla q_h) = -k^2 (\epsilon_r f, \varphi_h) \\ & \text{forall } \varphi_h \in V_h, \\ & (\epsilon_r \mathcal{B}_h f, \nabla \xi_h) = 0 \text{ for all } \xi_h \in \hat{S}_h. \end{aligned} \quad (11.29)$$

The last equation ensures  $\mathcal{B}f \in X_{0,b}$ .

Now Corollary 11.12 shows that the bilinear form  $(\mu_r^{-1} \nabla \times \cdot, \nabla \times \cdot)$  is coercive on  $X_{0,b}$  and the fact that  $\nabla \hat{S}_b \subset V_b$  can be used to verify the Babuška–Brezzi condition (as in the proof of Theorem 7.11). Thus, we know that(11.30)

$$\begin{aligned} \|\mathcal{B}f - \mathcal{B}_h f\|_{H(\text{curl}; \Omega)} & \leq C \left\{ \inf_{\mathcal{X}_h \in V_h} \|\mathcal{B}f - \mathcal{X}_h\|_{H(\text{curl}; \Omega)} \right. \\ & \quad \left. \inf_{\xi_h \in \hat{S}_h} \|\nabla(q - \xi_h)\|_{(L^2(\Omega))^3} \right\}. \end{aligned}$$

Then the density of  $V_b$  in  $H(\text{curl}; \Omega)$  and of  $\hat{S}_b$  in  $H^1(\Omega)/\mathbb{R}$  completes the proof (actually we have proved pointwise convergence in  $H(\text{curl}; \Omega)$ , which is more than sufficient).

Next we show that the set of operators is collectively compact. Let  $U \subset (L^2(\Omega))^3$  be a bounded set. Then, if  $u \in U$ , we know that  $\mathcal{B}_h u \in X_{0,b}$  satisfies

$$\left( \mu_r^{-1} \nabla \times \mathcal{B}_h u, \nabla \times \varphi_h \right) = -k^2 (\epsilon_r u, \varphi_h) \text{ forall } \varphi_h \in \hat{X}_{0,h}.$$

It follows from Corollary 11.12 that  $\|\nabla \times \mathcal{B}_h u\|_{(L^2(\Omega))^3} \leq C \|u\|_{(L^2(\Omega))^3}$ . But using the discrete Friedrichs inequality in Corollary 11.12, we have

$$\|\mathcal{B}_h u\|_{(L^2(\Omega))^3} + \|\nabla \times \mathcal{B}_h u\|_{(L^2(\Omega))^3} \leq C \|u\|_{(L^2(\Omega))^3}.$$

Then, by the discrete compactness property, we can extract a convergent subsequence from  $\mathcal{B}_h(U)$ . Thus,  $\mathcal{B}_h(U)$  is pre-compact in  $(L^2(\Omega))^3$ , as required.  $\square$

We have written the finite element problem as an operator equation (see (11.25)) so that we can now prove the basic existence and convergence theorem for  $\mathbf{g}_{i,b}$ :

**Theorem 11.14** For sufficiently small  $b$ , the operator  $\mathbf{g}_{i,b}$  is well defined and

$$\|(\mathcal{G}_i - \mathcal{G}_{i,h})\lambda\|_{H^{-1/2}(\text{Div}; \Sigma)} \rightarrow 0$$

as  $b \rightarrow 0$ .

**Proof** First we show that  $\zeta_b$  (see (11.22)) is well defined, using Theorem 2.51. Via the collective compactness and pointwise convergence of  $\mathcal{B}_b$ , we know that, provided  $b$  is small enough,  $(I + \mathcal{B}_b)$  is invertible with uniformly bounded inverse as a map from  $(L^2(\Omega))^3$  into itself. Hence  $\zeta_b$  and  $p_b$  in (11.22) are well defined. Furthermore, the following error estimate holds:

$$\|z - z_h\|_{(L^2(\Omega))^3} \leq C \left( \|(\mathcal{B}_h - \mathcal{B})z\|_{(L^2(\Omega))^3} + \|\mathcal{F} - \mathcal{F}_h\|_{(L^2(\Omega))^3} \right).$$

We estimate the first term on the right-hand side using (11.30) and the fact that  $q = 0$  because  $\zeta \in X_0$ . The same arguments as in the proof of Lemma 7.12 show that

$$\|\mathcal{F} - \mathcal{F}_h\|_{(L^2(\Omega))^3} \leq C \left( \|F - \eta_h\|_{H(\text{curl}; \Omega)} + \|p - \varphi_h\|_{H^1(\Omega)} \right)$$

for any  $\eta_b \in V_b$  and  $\varphi_b \in \hat{S}_b$ . We have thus proved that (11.31)

$$\begin{aligned} \|z - z_h\|_{(L^2(\Omega))^3} &\leq C \left\{ \|p - \psi_h\|_{H^1(\Omega)} + \|\mathcal{F} - \eta_h\|_{H(\text{curl}; \Omega)} \right. \\ &\quad \left. + \|\mathcal{B}_z - \psi_h\|_{H(\text{curl}; \Omega)} \right\}. \end{aligned}$$

It remains to derive an error estimate in the  $H(\text{curl}; \Omega)$  norm. This is done as in the proof of Theorem 7.25. We obtain (11.32)

$$\begin{aligned} \|z - z_h\|_{H(\text{curl}; \Omega)} &\leq C \left\{ \|p - \psi_h\|_{H^1(\Omega)} + \|\mathcal{F} - \eta_h\|_{H(\text{curl}; \Omega)} \right. \\ &\quad \left. + \|\mathcal{B}_z - \psi_h\|_{H(\text{curl}; \Omega)} \right\}. \end{aligned}$$

To obtain an estimate for  $u - u_h$ , we use (11.22) to write (11.33)

$$\|u - u_h\|_{H(\text{curl}; \Omega)} \leq C \left[ \|z - z_h\|_{H(\text{curl}; \Omega)} + \|\nabla(p - p_h)\|_{(L^2(\Omega))^3} \right].$$

But  $p_b$  converges to  $p$  in  $\hat{S}$  and  $\zeta_b$  converges to  $\zeta$  by the density of  $\hat{S}_b$  in  $\hat{S}$  and of  $X_b$  in  $H(\text{curl}; \Omega)$ , respectively. Using the trace theorem for  $H(\text{curl}; \Omega)$  on smooth surfaces, we have (11.34)

$$\|(\mathcal{G}_i - \mathcal{G}_{i,h})\lambda\|_{H^{-1/2}(\text{Div}; \Sigma)} \leq \|u - u_h\|_{H(\text{curl}; \Omega)} \rightarrow 0 \text{ as } h \rightarrow 0.$$

We have proved the desired result.  $\square$

The above estimate shows that the  $\mathbf{g}_{i,b}$  converges to  $\mathbf{g}_i$  with very general assumptions on the smoothness of the data ( $\epsilon_r$ ,  $\mu_r$  and  $\Omega$ ) to the problem but with no rate of convergence. To obtain optimal error estimates for the coupled problem, we need to establish a convergence rate. Hence, for the remainder of the Chapter, we shall assume that the coefficients  $\epsilon_r$  and  $\mu_r$  are sufficiently smooth that the following *a priori* estimates hold.

(1) For every  $p \in \hat{S}$  satisfying  $\nabla \cdot (\epsilon_r \nabla p) = 0$  in  $\Omega$ , we have

$$\|\nabla p\|_{PH^1(\text{curl}; \Omega)} \leq C \left\| \frac{\partial p}{\partial \hat{x}} \right\|_{H^{1/2}(\Sigma)}. \quad (11.35)$$

This is a typical elliptic regularity estimate for  $p$  for smooth data.

(2) Suppose  $f \in (L^2(\Omega))^3$  is such that  $\nabla \cdot (\mu_r f) = 0$  in  $\Omega$ . Let  $u \in H(\text{curl}; \Omega)$  satisfy

$$\begin{aligned} \nabla \times u &= \mu_r f && \text{in } \Omega, \\ \nabla \cdot (e_r u) &= 0 && \text{in } \Omega, \\ \hat{x} \cdot u &= 0 && \text{on } \Sigma. \end{aligned}$$

Then  $u \in PH^1(\Omega)$  and (11.36)

$$\|u\|_{PH^1(\Omega)} \leq C \|f\|_{(L^2(\Omega))^3}.$$

(3) Let  $f \in (L^2(\Omega))^3$  and  $g \in H^{1/2}(\text{Div}; \Sigma)$  satisfy the compatibility condition that

$$(e_r f, \nabla \xi) + (g, \nabla \xi) = 0 \quad \text{forall } \xi \in H^1(\Omega). \quad (11.37)$$

Let  $v \in H(\text{curl}; \Omega)$  satisfy

$$\begin{aligned} \nabla \times v &= e_r f && \text{in } \Omega, \\ \nabla \cdot (\mu_r v) &= 0 && \text{in } \Omega, \\ \hat{x} \cdot v &= g && \text{on } \Sigma. \end{aligned}$$

Then  $v \in PH^1(\Omega)$  and (11.38)

$$\|v\|_{PH^1(\Omega)} \leq C \left[ \|f\|_{(L^2(\Omega))^3} + \|g\|_{H^{1/2}(\text{Div}; \Sigma)} \right].$$

Obviously, these assumptions rule out rough boundaries between the domains  $\Omega_\epsilon$  where  $\epsilon_r$  and  $\mu_r$  are smooth (see Section 4.2 for details on the assumptions on the data). Note that all the above estimates hold if  $\epsilon_r$  and  $\mu_r$  are continuously differentiable in  $\Omega^-$ . For a discussion of regularity of Maxwell's equations in the presence of piecewise smooth functions with smooth interfaces, see [293], and for the case of piecewise constant coefficients with non-smooth interfaces see [107].

Using (11.36) and (11.38), if  $f$  and  $g$  satisfy the compatibility condition (11.37) and if  $u \in X_0$  satisfies

$$\begin{aligned} \nabla \times \mu_r^{-1} \nabla \times u &= e_r f && \text{in } \Omega, \\ \hat{x} \times \nabla \times u &= g && \text{on } \Sigma, \end{aligned}$$

then (11.39)

$$\|u\|_{PH^1(\text{curl}; \Omega)} \leq C \left[ \|f\|_{(L^2(\Omega))^3} + \|g\|_{H^{1/2}(\text{Div}; \Sigma)} \right].$$

The goal of the remainder of this section is to prove the following error estimate.

**Theorem 11.15** Assume that (11.35)–(11.38) hold. Then there exists a constant  $C$  such that

$$\|(g_i - g_{i,h})\lambda\|_{H^{-1/2}(\text{Div}, \Sigma)} \leq Ch\|\lambda\|_{H^{1/2}(\text{Div}, \Sigma)}, \quad h \in \Lambda.$$

for all  $\lambda \in H^{1/2}(\text{Div}; \Sigma)$

Before starting the proof of this theorem, note that if the functions  $p$  and  $Bf$  are smooth, we also have an error estimate for  $\mathcal{B}_h$  as the next theorem shows (this follows from (11.30) and the approximation properties of  $V_h$  and  $\hat{S}_h$ ).

**Theorem 11.16** If  $Bf \in PH^1(\text{curl}; \Omega)$  and  $p \in PH^2(\Omega)$  then

$$\|\mathcal{B}f - \mathcal{B}_h f\|_{H(\text{curl}; \Omega)} \leq Ch\left(\|\mathcal{B}f\|_{PH^1(\text{curl}; \Omega)} + \|p\|_{PH^2(\Omega)}\right).$$

**Proof of Theorem 11.15** We can simply use estimate (11.32) followed by (11.33) and (11.34). First we estimate  $p - \varphi_h$ . Note that  $p \in \hat{S}$  is defined by (11.9) and so satisfies

$$\begin{aligned} \nabla \cdot (\epsilon_r \nabla p) &= 0 \quad \text{in } \Omega, \\ \frac{\partial p}{\partial \vec{x}} &= \frac{i}{k} \nabla_\Sigma \cdot \lambda \text{ on } \Sigma. \end{aligned}$$

Using assumption (11.35) and choosing  $\varphi_h$  to be the  $\hat{S}$  projection of  $p$ , we have

$$\|p - \varphi_h\|_{H^1(\Omega)} \leq Ch\|p\|_{PH^2(\Omega)} \leq Ch\|\lambda\|_{H^{1/2}(\text{Div}, \Sigma)}.$$

Now we estimate  $\|F - \eta_h\|$  by choosing  $\eta_h = r_h F$ . Then  $\|F - \eta_h\|_{H(\text{curl}; \Omega)} \leq Ch\|F\|_{PH^1(\text{curl}; \Omega)}$ . Using (11.39), we have

$$\|\mathcal{F}\|_{PH^1(\text{curl}; \Omega)} \leq C\left(\|\nabla p\|_{(L^2(\Omega))^3} + \|\lambda\|_{H^{1/2}(\text{Div}, \Sigma)}\right)$$

and thus

$$\begin{aligned} \|\mathcal{F} - \eta_h\|_{(L^2(\Omega))^3} &\leq Ch\left(\|\nabla p\|_{(L^2(\Omega))^3} + \|\lambda\|_{H^{1/2}(\text{Div}, \Sigma)}\right) \\ &\leq Ch\|\lambda\|_{H^{1/2}(\text{Div}, \Sigma)}. \end{aligned}$$

It remains only to estimate  $z - \tau_h z$  and  $\mathcal{B}_h - \psi_h$ . We choose  $\tau_h = r_h z$  and  $\psi_h = r_h \mathcal{B}_h z$  and, proceeding as for the other estimates, we can show that

$$\begin{aligned} \|z - \pi_h z\|_{H(\text{curl}; \Omega)} &\leq Ch\|z\|_{PH^1(\text{curl}; \Omega)} \leq Ch\|\lambda\|_{H^{1/2}(\text{Div}, \Sigma)}, \\ \|\mathcal{B}z - \mathcal{B}_h z\|_{H(\text{curl}; \Omega)} &\leq Ch\|\mathcal{B}z\|_{PH^1(\text{curl}; \Omega)} \leq Ch\|\lambda\|_{H^{1/2}(\text{Div}, \Sigma)}. \end{aligned}$$

Combining all the estimates in (11.32), (11.33) and (11.34) proves the theorem.  $\square$

## 11.5 Error estimates for the fully discrete problem

In this section we shall analyze the fully discrete problem. In particular, we shall prove the following theorem.

**Theorem 11.17** *Assume that (11.35)–(11.38) hold. Then there is a  $\delta > 0$  such that, for  $N$  sufficiently large and  $bN < \delta$ , there is unique solution  $\lambda_{N,b} \in S_N$  of (11.18) satisfying*

$$\|\lambda_{N,h} - \lambda\|_{H^{-1/2}(\text{Div}; \Sigma)} \leq C \left( h \|f\|_{H^{1/2}(\text{Div}; \Sigma)} + (h + 1/N) \|\lambda\|_{H^{1/2}(\text{Div}; \Sigma)} \right).$$

**Remark 11.18** *We can obtain a higher power of  $N$  in this estimate (at the expense of a higher norm of  $\lambda$ ). For the Helmholtz equation in two dimensions, using a similar method, Andreas Kirsch and I were able to prove optimal estimates without a stability relation between  $b$  and  $N$  [188]. Grote and Keller [153] proved the same result for the Helmholtz equation in three dimensions. Unfortunately, we have been unable to prove this for Maxwell's equations, and instead must require that the mesh size be sufficiently small compared to the number of modes used on the boundary.*

**Proof of Theorem 11.17** Note first that by operating on (11.3) by  $P_N$  and using the fact that  $\mathbf{g}_e$  and  $P_N$  commute we have (11.40)

$$P_N \mathcal{G}_i \lambda - \mathcal{G}_e P_N \lambda = P_N f.$$

Let us define  $e_{N,b} = \lambda_{N,b} - P_N \lambda$ . Then using (11.40) and (11.18) we have

$$\begin{aligned} (P_N \mathcal{G}_{i,h} - P_N \mathcal{G}_e) e_{N,h} &= P_N (\mathcal{G}_i - \mathcal{G}_e) e_{N,h} + P_N (\mathcal{G}_{i,h} - \mathcal{G}_i) e_{N,h} \\ &= P_N (\mathcal{G}_i - \mathcal{G}_{i,h}) P_N \lambda - P_N \mathcal{G}_i (P_N \lambda - \lambda), \end{aligned}$$

where we have used the fact that  $P_{N,\mathbf{g}_e}(P_N \lambda - \lambda) = \mathbf{g}_e P_N(P_N \lambda - \lambda) = 0$ . Using Theorem 11.1 we have the decomposition  $\mathbf{g}_i - \mathbf{g}_e = T + K$ , where  $T, K : H^{1/2}(\text{Div}; \Sigma) \rightarrow H^{-1/2}(\text{Div}; \Sigma)$ ,  $T$  is an isomorphism and  $K$  is compact. Using the fact that  $P_N$  and  $T$  commute, we obtain our fundamental error equation:

$$\begin{aligned} T e_{N,h} + P_N K e_{N,h} + P_N (\mathcal{G}_{i,h} - \mathcal{G}_i) e_{N,h} \\ = P_N (\mathcal{G}_i - \mathcal{G}_{i,h}) P_N \lambda - P_N \mathcal{G}_i (P_N \lambda - \lambda). \end{aligned}$$

First we need to show that this equation has a solution. The operator on the left-hand side is

$$T + P_N K + P_N (\mathcal{G}_{i,h} - \mathcal{G}_i) = (T + K) + (P_N K - K) + P_N (\mathcal{G}_{i,h} - \mathcal{G}_i).$$

The operator  $T + K$  is invertible due to Theorem 11.10. We now apply Theorem 2.27 by showing that  $P_N K \rightarrow K$  and  $P_N (\mathbf{g}_{i,b} - \mathbf{g}_i) \rightarrow 0$  considered as operators from  $H^{-1/2}(\text{Div}; \Sigma)$  to  $H^{-1/2}(\text{Div}; \Sigma)$ .

Since  $P_N$  is the orthogonal projection for  $H^{-1/2}(\text{Div}; \Sigma)$  into  $S_N$  and  $K$  is compact in this space, we know that  $P_N K \rightarrow K$  in the operator norm of

$H^{-1/2}(\text{Div}; \Sigma)$  (see Lemma 2.50). For the other term on the left-hand side above, we can use the error estimate for the finite element solution in Theorem 11.15, together with the inverse estimate (11.16), to show that

$$\begin{aligned} \|P_N(g_{i,h} - g_i)e_{N,h}\|_{H^{-1/2}(\text{Div}; \Sigma)} &\leq \|(g_{i,h} - g_i)e_{N,h}\|_{H^{-1/2}(\text{Div}; \Sigma)} \\ &\leq Ch\|e_{N,h}\|_{H^{1/2}(\text{Div}; \Sigma)} \\ &\leq CNh\|e_{N,h}\|_{H^{-1/2}(\text{Div}; \Sigma)}. \end{aligned}$$

Thus, for sufficiently large  $N$  and small  $Nh$ , the operator  $T + P_N K + P_N(\mathbf{g}_{i,b} - \mathbf{g}_i)$  is invertible with uniformly bounded inverse. This implies that  $e_{N,b}$  and hence  $\lambda_{N,b}$  is well defined and we can obtain an error estimate simply by estimating the right-hand side using Theorem (11.15) and the estimate (11.15):

$$\begin{aligned} \|P_N(g_i - g_{i,h})P_N\lambda\|_{H^{-1/2}(\text{Div}; \Sigma)} &\leq Ch\|P_N\lambda\|_{H^{1/2}(\text{Div}; \Sigma)}, \\ \|P_N g_i(P_N\lambda - \lambda)\|_{H^{-1/2}(\text{Div}; \Sigma)} &\leq C\|P_N\lambda - \lambda\|_{H^{-1/2}(\text{Div}; \Sigma)} \\ &\leq \frac{C}{N}\|\lambda\|_{H^{-1/2}(\text{Div}; \Sigma)}. \end{aligned}$$

Putting these estimates together, we obtain the estimate

$$\|e_{N,h}\|_{H^{-1/2}(\text{Div}; \Sigma)} \leq C\left(h\|f\|_{H^{1/2}(\text{Div}; \Sigma)} + (h+1/N)\|\lambda\|_{H^{1/2}(\text{Div}; \Sigma)}\right).$$

The use of the triangle equality then proves the estimate of the theorem.  $\square$

Our final result gives an error estimate for the field in the scatterer and near it. It follows from the previous result.

**Corollary 11.19** *Assume that (11.35)–(11.38) hold. Let  $E \in H_{\text{loc}}(\text{curl}; \mathbb{R}^3)$  satisfy (1.26)–(1.29) with  $D = \emptyset$ . Define  $E_b \in V_b$  to satisfy*

$$\left(\mu_r^{-1}\nabla \times E_h, \nabla \times \varphi_h\right) - k^2(\in_r E_h, \varphi) + ik\langle\lambda_{h,N}, \varphi_h\rangle = 0 \quad \text{forall } \varphi_h \in V_h,$$

where  $\lambda_{N,b} \in S_N$  satisfies (11.18). Then, for  $b$  sufficiently small,  $E_b$  is well defined and

$$\|E - E_h\|_{H(\text{curl}; \Omega)} \leq C\left(h\|f\|_{H^{1/2}(\text{Div}; \Sigma)} + (h+1/N)\|\lambda\|_{H^{1/2}(\text{Div}; \Sigma)}\right).$$

Looking through the proofs we have given in this chapter, we see that we have made almost no use of the fact that we used the lowest-order edge element space. This choice was made to simplify the assumptions on regularity for proving convergence to obtain an optimal convergence rate. We could use higher-order curvilinear edge elements and gain higher-order convergence for smooth solutions.

We now need to comment on how to compute with this scheme. The simplest implementation (attractive in two dimensions) is to compute the matrix representing  $P_{N,\mathbf{g},b} - \mathbf{g}_e$  on  $S_N$ . To do this, we must solve the interior finite element problem (11.17) with right-hand side  $\lambda$  taken to be each basis function in  $S_N$ .

(i.e.  $u_n^m$  and  $v_n^m$ ,  $1 \leq n \leq N$  and  $-n \leq m \leq n$ ). For each basis function, after expressing the result as a series in  $S_N$  we obtain from this series the coefficients of one column of a matrix C called the capacitance matrix. Once C has been determined, it is only necessary to project f onto  $S_N$  and solve a matrix problem to solve (11.18). Despite the fact that C is a dense matrix, this can be an efficient strategy in two dimensions particularly if the problem is to be solved for many right hand sides (i.e. many incident waves). In three dimensions this becomes memory and time consuming.

Another approach is to write the equation (11.18) as (11.41)

$$\left( \mathcal{G}_e^{-1} P_N \mathcal{G}_{i,h} - i \right) \lambda_{N,h} = \mathcal{G}_e^{-1} P_N f .$$

Now we can apply an iterative method to this equation, for example the biconjugate gradient scheme (or the conjugate gradient scheme for the normal equations for (11.41)). In these iterative schemes, there is no need to compute and store the matrix representing  $\mathcal{G}_{i,h}$ . Instead, we must be able to compute the action of  $\mathcal{G}_e^{-1} P_N \mathcal{G}_{i,h}$  on a vector  $\mu \in S_N$  which can be done by solving just one interior Maxwell problem (11.17). We also need to compute the adjoint of this operator (i.e. the conjugate transpose of the corresponding matrix) applied to a general vector  $\mu \in S_N$ . We now show how to compute  $\mathcal{G}_{i,h}^* \mu : H^{-1/2}(\text{Curl}; \Sigma) \rightarrow H^{-1/2}(\text{Curl}; \Sigma)$ . By definition, for any  $\lambda \in H^{-1/2}(\text{Div}; \Sigma)$  and  $\mu \in H^{-1/2}(\text{Curl}; \Sigma)$ , we have

$$\langle \lambda, \mathcal{G}_{i,h}^* \mu \rangle = \langle \mathcal{G}_{i,h} \lambda, \mu \rangle .$$

But  $\mathcal{G}_{i,h} \lambda = \mathcal{O} \times w_b$ , where  $w_b \in V_b$  satisfies (11.17) and so, taking into account conjugation in the definition of  $\langle \cdot, \cdot \rangle$ ,

$$\langle \lambda, \mathcal{G}_{i,h}^* \mu \rangle = \langle \omega_h, \mu \times \hat{x} \rangle = \langle \bar{\mu} \times \hat{x}, \bar{\omega}_h \rangle .$$

Now let  $v_b \in V_b$  satisfy (11.42)

$$\left( \mu_r^{-1} \nabla \times u_h \nabla \times \varphi_h \right) - k^2 (\in_r u_h, \varphi_h) = -i k \langle \bar{\mu} \times \hat{x}, \varphi_h \rangle \quad \text{forall } \varphi_h \in V_h .$$

This has a unique solution (at least if  $b$  is small enough), by Theorem 11.14. Thus,

$$\begin{aligned} \langle \lambda, \mathcal{G}_{i,h}^* \mu \rangle &= (-1/i k) \left( \left( \mu_r^{-1} \nabla \times u_h, \nabla \times \bar{\omega}_h \right) - k^2 (\in_r u_h, \bar{\omega}_h) \right) \\ &= (-1/i k) \left( \left( \mu_r^{-1} \nabla \times \omega_h, \nabla \times \bar{v}_h \right) - k^2 (\in_r \omega_h, \bar{v}_h) \right) = \langle \lambda, \bar{v}_h \rangle . \end{aligned}$$

We see that  $\mathcal{G}_{i,h}^* \mu = \bar{v}_{h,T}$ . Hence we can compute  $\mathcal{G}_{i,h}^* \mu$  at the cost of solving a second interior Maxwell problem like (11.17). For a two dimensional example including numerical results, see [189].

Of course, the method we presented in this chapter is not the only possible way to solve the problem. Since there is no perfect conductor, one way is by using suitable volume integral equations (see, e.g. [94, 177]). This method handles the

infinite domain precisely, but requires the evaluation of singular integrals and the approximate inversion of a large dense system (of course, using a suitable iterative method).

There are many other possible methods for approximating this scattering problem. For example, the interior finite element method can be coupled to a boundary element method that effectively computes the Calderon operator (and allows a rather general artificial boundary) [177]. This method is very often used in practice and has been analyzed by Hiptmair [163]. Other methods include the use of a perfectly matched layer of (see [36] and Section 13.5.3) and infinite elements (see Section 13.5) to terminate the finite element region.

In the method presented here, the scattering problem is decomposed into two parts, one on the bounded domain inside the artificial boundary and the other on its infinite complement. Matching is done on the artificial boundary. As a result, the method is said to be a non-overlapping domain decomposition scheme. An alternative scheme proposed in [159] is to use an overlapping method. This introduces a coupling between the solution at some interior points and the solution at some points on the artificial boundary. The method allows a very general artificial boundary, whereas the method we describe here is restricted to a spherical outer boundary. More general boundaries are possible (e.g. ellipsoidal boundaries), at the expense of working with suitable basis functions in more general coordinate systems. We shall discuss the overlapping method in the next chapter.

# 12 SCATTERING BY A BURIED OBJECT

## 12.1 Introduction

In this chapter we describe a method for approximating the electromagnetic field scattered from objects embedded in a non-uniform background medium. We have in mind scattering from buried objects. In the simple model for this problem presented in Chapter 1 and considered later in this chapter, the space  $\mathbb{R}^3$  is divided into two half spaces by the plane  $x_3 = 0$ . In the lower half space is a uniform conducting medium modeling the earth. In the upper half space is a non-conducting medium modeling the air. This two-layered medium is the background medium. Buried objects are perturbations of the lower layer in this background (see Section 1.3).

The method we are going to describe uses an integral representation of the field away from the scatterer. In particular, this requires a knowledge of the Green's function for the background medium. Although complex, this is well known for the layered earth model just mentioned. Within the scatterers if necessary, and in a domain containing them, we represent the field by finite elements. Thus, there is an overlapping region in which both the finite element method and integral representation give an approximation to the electromagnetic field. The overlapping method is due to Hazard and Lenoir [159] for a homogeneous isotropic background and to Cutzach and Hazard [111] in the case of a layered background medium. The work has its roots in the paper of Jami and Lenoir [173].

In fact, this overlapping scheme is not the standard method for problems of this type. Instead it is usual to use an integral formulation on the boundary of the scatterer to take care of the infinite region. This requires the use of an integral equation with a singular kernel which complicates the implementation of the scheme, since a special quadrature scheme is needed to take care of the singular integral. In the overlapping method no singular integrals are encountered. Of course, the overlapping method requires to mesh a larger domain than for the non-overlapping scheme (although not much larger) and the resulting linear system is less structured. However, Liu and Jin [212] have shown that an overlapping scheme may have advantages from the point of view of implementation and we shall return to this point later in the chapter.

The plan for this chapter is as follows. In the first section, we apply the overlapping scheme to our model scattering problem of scattering by a bounded perfectly conducting object in a homogeneous uniform background medium. It turns out that a direct application of the Hazard and Lenoir approach leads to

unwieldy matrices. Thus, we apply flux-recovery procedures [294, 21–23] in the discretization of the method resulting in a fully discrete problem that is better suited to implementation. We prove convergence using the technique from [167] and comment on implementation and on the relationship to Liu and Jin's scheme.

In Section 12.3 we comment on how the method must be modified when there is an infinite perfectly conducting ground plane present. Finally in the last section, we formulate the problem for a buried perfectly conducting scatterer.

## 12.2 Homogeneous isotropic background

To describe the overlapping method as simply as possible, we first consider our standard model problem consisting of a perfectly conducting scatterer which occupies a bounded, Lipschitz, polyhedral region  $D$  in  $\mathbb{R}^3$  with connected complement. In this case the background medium is isotropic and homogeneous. Thus, we wish to approximate the total field  $E \in H_{\text{loc}}(\text{curl}; \mathbb{R}^3 \setminus D)$  such that (1.26)–(1.29) are satisfied in the special case  $\varepsilon_r = \mu_r = 1$ .

We shall now derive an integral representation of the scattered field away from the scatterer in free space. It turns out that the Stratton–Chu formula given in Theorem 9.4 is not suitable for our purposes, since we need a formula that we can relate to the finite element variational problem. As we shall see, by rewriting the Stratton–Chu formula, we can achieve a better formula. In preparation for this, we define the matrix function(12.1)

$$\mathbb{G}(x, y) = \Phi(x, y)\mathbb{I} + \frac{1}{\kappa^2} \nabla_y \nabla_y \Phi(x, y), \quad x \neq y,$$

where  $\mathbb{I}$  is the  $3 \times 3$  identity matrix and  $\nabla_y \nabla_y \Phi(x, y)$  is the Hessian matrix for  $\Phi$  defined by

$$(\nabla_y \nabla_y \Phi(x, y))_{l,m} = \frac{\partial^2 \Phi}{\partial y_l \partial y_m}, \quad 1 \leq l, m \leq 3.$$

**Definition 12.1** The matrix  $\mathbb{G}$  in (12.1) is called the *dyadic Green's function* for Maxwell's equations.

We shall shortly see that the dyadic Green's function arises naturally. We denote by  $g(x, y)$  the  $l$ th column of  $\mathbb{G}(x, y)$  and define  $\nabla \times \mathbb{G}$  to be the matrix with  $l$ th column  $\nabla \times g$ .

The dyadic Green's function is related to Maxwell's equations in the following way. A simple calculation shows that each column satisfies the homogeneous Maxwell system when  $x \neq y$  and we shall see that in fact each column satisfies

$$\nabla_y \times \nabla_y \times g_l - \kappa^2 g_l = e_l \delta_x \text{ in } \mathbb{R}^3$$

together with the Silver–Müller radiation condition (9.14) where  $e_l$  is the  $l$ th unit vector. Using the extension of curl to matrices defined above, we can write this as

$$\nabla_y \times \nabla_y \times \mathbb{G} - \kappa^2 \mathbb{G} = \mathbb{I} \delta_x \text{ in } \mathbb{R}^3.$$

The representation of  $\mathbb{G}$  in (12.1) can be derived directly from Maxwell's equations using properties of  $\Phi$  (see Theorem 5.2.1 of [236]), but we only use the result of the next corollary so we have adopted a more roundabout approach of using the Stratton–Chu formula.

**Theorem 12.2** Under the conditions of Theorem 9.4, for any  $x \in \mathbb{R}^3 \setminus \bar{D}$ ,

$$\begin{aligned} E(x) &= \int_{\Gamma} \left\{ \mathbb{G}^T(x, y)(v \times (\nabla \times E))(y) \right. \\ &\quad \left. + (\nabla_y \times \mathbb{G})^T(x, y)(v \times E)(y) \right\} dA(y), \end{aligned}$$

where  $v$  is the unit outward normal to  $D$ ,  $\Gamma = \partial D$ , and  $\mathbb{G}^T(x, y)(v \times (\nabla \times E))(y)$  and  $(\nabla_y \times \mathbb{G})^T(x, y)(v \times E)(y)$  are understood as matrix–vector multiplications.

**Remark 12.3** Note that we assume that  $\varepsilon_r = \mu_r = 1$  in  $\mathbb{R}^3 \setminus \bar{D}$ . If either  $\varepsilon_r \neq 1$  or  $\mu_r \neq 1$ , we would need to replace  $\Gamma$  in this theorem with another Lipschitz smooth surface  $S$  containing the scatterer (i.e. both  $D$  and the region where  $\varepsilon_r$  or  $\mu_r$  are not unity) in its interior.

**Proof of Theorem 12.2** Using Theorem 9.4, we need to rewrite the expressions on the right-hand side of (9.15). First, using the fact that  $\Phi$  is the fundamental solution of the Helmholtz equation and  $x \neq y$ , we have (12.3)

$$\begin{aligned} -\frac{1}{i\kappa} \nabla \times \nabla \times \int_{\Gamma} (v \times H)(y) \Phi(x, y) dA(y) \\ = \frac{1}{i\kappa} (\Delta - \nabla \nabla \cdot) \int_{\partial D} (v \times H)(y) \Phi(x, y) dA(y) \\ = -\frac{1}{i\kappa} \int_{\Gamma} \left\{ \kappa^2 (v \times H)(y) \Phi(x, y) \right. \\ \left. + \nabla_x [(v \times H)(y) \cdot \nabla_x \Phi(x, y)] \right\} dA(y). \end{aligned}$$

But the  $k$ th entry of the gradient term is

$$\begin{aligned} (\nabla_x [(v \times H)(y) \cdot \nabla_x \Phi(x, y)])_k &= \frac{\partial}{\partial x_k} \sum_{m=1}^3 (v \times H)_m(y) \frac{\partial \Phi}{\partial x_m}(x, y) \\ &= \sum_{m=1}^3 (v \times H)_m(y) \frac{\partial^2 \Phi}{\partial x_k \partial x_m}(x, y). \end{aligned}$$

Using this fact in (12.3) and the fact that  $H = (1/i\kappa) \nabla \times E$  shows that

$$\begin{aligned} -\frac{1}{i\kappa} \nabla \times \nabla \times \int_{\partial D} (v \times H)(y) \Phi(x, y) dA(y) \\ = \int_{\Gamma} \left( \Phi \mathbb{I} + \frac{1}{\kappa^2} \nabla_y \nabla_y \Phi \right)^T(x, y) (v \times (\nabla \times E))(y) dA(y) \end{aligned}$$

as required. For the other term, using the fact that  $\nabla_x \Phi = -\nabla_y \Phi$ ,

$$\begin{aligned} \nabla \times \int_{\Gamma} (\mathbf{v} \times \mathbf{E})(y) \Phi(x, y) dV(y) &= \int_{\Gamma} (\mathbf{v} \times \mathbf{E})(y) \times \nabla_y \Phi(x, y) dV(y) \\ &= \int_{\Gamma} (\nabla_y \times (\mathbb{I}\Phi))^T(x, y) (\mathbf{v} \times \mathbf{E})(y) dV(y) \\ &= \int_{\Gamma} (\nabla_y \times \mathbb{G})^T(x, y) (\mathbf{v} \times \mathbf{E})(y) dV(y), \end{aligned}$$

and we are done.  $\square$

The scattering problem we wish to solve is the standard model problem of finding  $\mathbf{E}$  and  $\mathbf{E}^s$  such that (12.4a)

$$\begin{aligned} \nabla \times (\nabla \times \mathbf{E}) - \kappa^2 \mathbf{E} &= \mathbf{F} \text{ in } \mathbb{R}^3 \setminus D, \\ \mathbf{v} \times \mathbf{E} &= 0 \text{ on } \Gamma, \end{aligned} \tag{12.4b}$$

$$\mathbf{E} = \mathbf{E}^i + \mathbf{E}^s \text{ in } \mathbb{R}^3 \setminus D, \tag{12.4c}$$

$$\lim_{x \rightarrow \infty} \rho((\nabla \times \mathbf{E}^s) \times \hat{\mathbf{x}} - i\kappa \mathbf{E}^s) = 0. \tag{12.4d}$$

Note that  $\mathbf{E}^s$  can then be represented using the results of Theorem 12.2 and this was our reason for deriving the result. As usual for the model scattering problem, we suppose there is a known incident field  $\mathbf{E}^i$  that satisfies the homogeneous, isotropic Maxwell's equations in the neighborhood of  $D$  and in  $D$ . It is thus an analytic function of  $x$  in a neighborhood of  $D$ . In particular, we have in mind two standard cases.

- (1) *Point source* We suppose the incident field is due to a point dipole source located at  $x_p \in \Omega$  with polarization  $p$ ,  $|p| = 1$ . In this case we take

$$\mathbf{E}^i(x) = \mathbb{G}(x_p, x)p.$$

Clearly

$$\nabla \times (\nabla \times \mathbf{E}^i(x)) - \kappa^2 \mathbf{E}^i(x) = p \delta_{x_p} \text{ in } \mathbb{R}^3,$$

so that  $\mathbf{F} = p \delta_{x_p}$  in (12.4a).

- (2) *Plane Wave* An incident plane wave with polarization  $p$  and direction of propagation  $d$  is given by (1.20). In this case, (12.4a) is satisfied with  $F = 0$ .

The scattering problem is posed on an infinite region  $\mathbb{R}^3 \setminus D$ . In order to apply a finite element method, we truncate the domain. Following Hazard and Lenoir, we introduce a connected, Lipschitz, polyhedral surface  $\Sigma$ , with interior  $D_\Sigma$ , such that  $D \subset D_\Sigma$ . The outward unit normal on  $\Sigma$  is again denoted by  $\mathbf{n}$ . We define the truncated computational domain  $\Omega = D_\Sigma \setminus D$ , and we assume that  $D_\Sigma$  and  $D$  are such that  $\Omega$  is simply connected and that the boundary of  $\Omega$  consists of two disjoint, connected components  $\Sigma$  and  $\Gamma$ .

The goal is to use finite elements in  $\Omega$  to approximate  $E$ , but we need a boundary condition on  $\Sigma$ . This is provided by using the version of the Stratton–Chu formula given in Theorem 12.2. For  $x \in \mathbb{R}^3 \setminus D$ , we define(12.5)

$$\begin{aligned}\mathcal{G}(E^s) &= \int_{\Gamma} \left\{ \mathbb{G}^T(x, y) (\nu \times (\nabla \times E^s))(y) \right. \\ &\quad \left. + (\nabla_y \times \mathbb{G})^T(x, y) (\nu \times E^s)(y) \right\} dA(y).\end{aligned}$$

Using the fact that  $E^i$  and the columns of  $\mathbb{G}$  are regular solutions of Maxwell's equations inside  $D$  (since  $x$  is outside  $D$ ), we have  $I(E^i) = 0$  in  $\Omega$  and thus

$$E = E^i + \mathcal{G}(E) \text{ in } \Omega,$$

provided  $E$  is regular enough for  $I(E)$ , defined in (12.5), to be well defined.

Unfortunately, the regularity requirement implicit in (12.5) is not met by functions in  $H(\text{curl}; \Omega)$ , since the term  $\nu \times \nabla \times u$  is not defined if  $u$  is a general function in this space. We therefore need to extend the definition of  $I$  to allow for less regular arguments.

Let  $\chi \in C_0^\infty(D_\Sigma)$  denote a cutoff function such that  $\chi = 1$  on  $\Gamma$  and define  $\widetilde{\mathbb{G}}(x, \cdot) \in H(\text{curl}; \Omega)$  by(12.6)

$$\widetilde{\mathbb{G}}(x, y) = X(y) \mathbb{G}(x, y).$$

We can now define the regularized integral operator(12.7)

$$\begin{aligned}\mathcal{G}^R(E^s) &= \int_{\Omega} \left( (\nabla_y \times \widetilde{\mathbb{G}})^T(x, y) \nabla \times E^s(y) \right. \\ &\quad \left. - \kappa^2 \widetilde{\mathbb{G}}^T(x, y) E^s(y) \right) dV(y) \\ &\quad + \int_{\Gamma} (\nabla_y \times \mathbb{G})^T(x, y) (\nu \times E^s)(y) dA(y),\end{aligned}$$

where the curl is again with respect to  $y$  and the integral is evaluated for  $x$  outside the support of the cutoff function  $\chi$  (in particular for  $x$  in a neighborhood of  $\Sigma$ ). Using integration by parts, we can verify that for a smooth solution  $E^s$  of (12.4) we have  $\mathcal{J}(E^s) = \mathcal{J}(E)$ , and thus(12.8)

$$E = E^i + \mathcal{G}^R(E).$$

Note also that, since  $\mathcal{J}(E)$  is evaluated outside the support of  $\chi$ , a further integration by parts, and the use of the perfectly conducting boundary condition on  $\Gamma$ , shows that(12.9)

$$\mathcal{G}^R(E) = \int_{\Omega} \left( \nabla_y \times (\nabla_y \times \widetilde{\mathbb{G}}) - \kappa^2 \widetilde{\mathbb{G}} \right)^T(x, y) E(y) dV(y).$$

This is the form of  $\mathcal{J}$  we shall use for the first part of the upcoming analysis.

Before stating the variational problem for Maxwell's equations, we define one further operator. For a sufficiently smooth field  $u$ , we can define a tangential impedance boundary condition operator on  $\Sigma$  as follows:(12.10)

$$T(u) = (\nabla \times u)|_{\Sigma} \times v - ik u_T \text{ on } \Sigma.$$

Now we can apply the Galerkin method to obtain a variational formulation of (12.4) as we did in the introduction to Chapter 4 . In particular we need to recall the subspace of  $H(\text{curl}; \Omega)$  denoted by  $X$  and defined in (4.3). Then the truncated version of problem (12.4) is to find  $E \in X$  such that(12.11)

$$\begin{aligned} (\nabla \times E, \nabla \times \varphi) - \kappa^2(E, \varphi) - ik \langle E_T, \varphi_T \rangle - \left\langle T\left(\mathcal{J}^R(E)\right), \varphi_T \right\rangle \\ = \left\langle T\left(E^i\right), \varphi_T \right\rangle \quad \text{forall } \varphi \in X. \end{aligned}$$

Hazard and Lenoir [159] show that problem (12.11) has a unique solution for every  $\kappa > 0$ , and given incident field  $E^i$ . We shall shortly give a modified version of this proof suitable for our later numerical analysis. First we define the finite element approximation of the above equation.

We suppose that  $\Omega$  has been covered by a regular mesh  $\tau_b$  consisting of tetrahedra of maximum diameter  $b$ . In addition, as in Section 7.3, we need to assume that  $\tau_b$  is quasi-uniform on  $\Sigma$ . On this mesh we have the standard space of  $k$ th-order edge elements denoted by  $X_b$  and defined in (7.1) that is derived from the space  $V_b$  of edge elements defined in (5.40).

For later use we need to discretize the operator  $\mathcal{J}^R$  defined in (12.7). Recall that we will only evaluate  $\mathcal{J}^R(E)$  in a neighborhood of  $\Sigma$ .

**Definition 12.4** Let  $G_b(x, \cdot)$  denote the matrix function for which  $\tilde{g}_{b,l}(x, \cdot)$  is the  $l$ th column of  $G_b(x, \cdot)$ . Then  $G_b$  is an *admissible discrete dyadic Green's function* if the following hold (where  $g(x, \cdot)$  is the  $m$ th column of  $G(x, \cdot)$ ) for  $1 \leq l \leq 3$ :

- (1)  $\tilde{g}_{b,l}(x, \cdot) \in V_b$ ;
- (2)  $(\tilde{g}_{b,l}(x, \cdot))_T$  interpolates  $(g(x, \cdot))_T$  on  $\Gamma$  (using edge and face degrees of freedom (5.36) and (5.37));
- (3)  $\tilde{g}_{b,l}(x, \cdot) = 0$  on all tetrahedra having a vertex, face or edge on  $\Sigma$ .

Obviously, this discretization of  $G(x, \cdot)$  is not uniquely determined by the above requirements. For computational convenience, we use  $\tilde{g}_{b,l}$ ,  $l = 1, 2, 3$ , that decay to zero rapidly away from  $\Gamma$ . This minimizes the support of  $G_b$  and is the reason for discretizing  $G$ .

We can now define the discretized version of the integral operator defined in (12.7) for  $u \in H(\text{curl } \Omega)$  and  $x$  outside the support of  $G_b$  (in particular, for  $x \in \Sigma$ ) by

$$\begin{aligned} \mathcal{J}_h(u)(x) = & \int_{\Omega} \left( (\nabla \times \widetilde{\mathbb{G}}_h)^h(x, y) \nabla \times u(y) \right. \\ & \left. - \kappa^2 \widetilde{\mathbb{G}}_h^T(x, y) u(y) \right) dV(y). \end{aligned} \quad (12.12)$$

As long as  $x$  is on  $\sum$ ,  $\mathcal{J}_h(u)$  is a smooth function of  $x$ . Hence,  $T(\mathcal{J}_h(u))$  is a well defined and smooth (tangential) vector field on each face on  $\sum$ .

The fully discrete finite element analogue of (12.11) is to find  $E_h \in X_h$  such that (12.13)

$$\begin{aligned} (\nabla \times E_h, \nabla \times \varphi_h) - \kappa^2(E_h, \varphi_h) - \langle i\kappa E_{h,T} + T(\mathcal{J}_h(E_h)), \varphi_{h,T} \rangle \\ = \langle T(E^i), \varphi_{h,T} \rangle \text{ for all } \varphi_h \in X_h. \end{aligned}$$

Unfortunately, we have been unable to prove directly that  $E_h$  converges to  $E$ . Instead, we first analyze the convergence of the solution of the following intermediate problem of finding  $\tilde{E}_h \in X_h$  such that (12.14)

$$\begin{aligned} (\nabla \times \tilde{E}_h, \nabla \times \varphi_h) - \kappa^2(\tilde{E}_h, \varphi_h) - \langle i\kappa \tilde{E}_{h,T} + T(\mathcal{J}^R(\tilde{E}_h)), \varphi_{h,T} \rangle \\ = \langle T(E^i), \varphi_{h,T} \rangle \text{ for all } \varphi_h \in X_h. \end{aligned}$$

Here the operator  $\mathcal{J}^R$  is not discretized.

In the next section we shall show that  $\tilde{E}_h$  is well defined and converges to the true solution  $E$ . In principle, we could implement (12.14) but the integral operator  $\mathcal{J}^R$  would become increasingly more expensive to evaluate as the mesh size decreases since a volume integral over a fixed volume must be evaluated. Hence we prefer to compute with (12.13), since  $\mathcal{J}_h$  can be constructed to only involve a skin of tetrahedra that share an edge with  $\Gamma$ .

### 12.2.1 Analysis of the scheme

We will prove that as the mesh size  $h$  decreases, the solutions of the discrete problem (12.14) approach the exact solution of (12.11). The approach follows very closely that of Section 7.3 and is from [167]. In order to use the Fredholm alternative in the analysis of the finite element formulation, we rewrite the continuous variational problem (12.11) and the discrete finite element problem (12.13) as operator equations. We recall the bilinear form  $a_+$  defined in (4.14) with  $\varepsilon_r = \mu_r = 1$  and  $\lambda = 1$ . Thus for  $u, v \in X$  we have (12.15)

$$a_+(u, v) = (\nabla \times u, \nabla \times v) + \kappa^2(u, v) - i\kappa \langle u_T, v_T \rangle.$$

Note that  $|a_+(u, u)|^{1/2}$  is a norm on  $X$  equivalent to  $\|u\|_X$ .

Now recall the space  $X_0$  of divergence-free fields in  $X$  defined by (4.8). Define the operator  $A : (L^2(\Omega))^3 \rightarrow (L^2(\Omega))^3$  such that for all  $f \in (L^2(\Omega))^3$ ,  $Af \in X_0 \subset (L^2(\Omega))^3$  satisfies (12.16)

$$a_+(Af, \varphi) = -2\kappa^2(f, \varphi) - \langle T(\mathcal{J}^R(f)), \varphi_T \rangle \text{ for all } \varphi \in X_0.$$

By the Lax–Milgram lemma, this problem is well posed. In particular, using the expression for  $\mathcal{J}^R$  in (12.9) shows that  $\|T(\mathcal{J}^R(u))\|_{(L^2(\Omega))^3}^2 \leq C \|u\|_{(L^2(\Omega))^3}^2$  which

allows us to prove the continuity of  $a_+(\cdot, \cdot)$ . The operator  $\mathcal{A}$  plays the part of the operator  $K$  in (4.15) for the theory here.

Similarly, we define  $F \in X_0$  by(12.17)

$$a + (\mathcal{F}, \varphi) = \left\langle T(E^i), \varphi \right\rangle \text{ for all } \varphi \in X_0 .$$

We proceed to show that the operator problem of finding  $E \in (L^2(\Omega))^3$  such that(12.18)

$$E + AE = \mathcal{F}$$

is exactly equivalent to solving the Hazard–Lenoir equation (12.11). Any solution of (12.11) is divergence-free and thus if we pick a test function  $\varphi \in X_0$ , we can recast (12.11) as the problem of finding  $E \in X_0$  such that

$$a + (E + AE = \mathcal{F}, \varphi) = 0 \text{ for all } \varphi \in X_0 .$$

Hence, in  $X$ ,  $E + AE - F = 0$  and this certainly implies equality in  $(L^2(\Omega))^3$ . Conversely, if we have a solution  $E \in (L^2(\Omega))^3$  of

$$E + AE = \mathcal{F},$$

then, since  $E = F - AE$ , we know that  $E \in X_0$ . Therefore,  $E$  satisfies

$$a + (E + AE = \mathcal{F}, \xi) = 0 \text{ for all } \xi \in X,$$

which is the Hazard–Lenoir equation (12.11). This shows the equivalence of the operator equation (12.18) and the Hazard–Lenoir equation (12.11).

Hazard and Lenoir prove the compactness of  $\mathcal{A}$  as an operator from  $X_0$  to  $X_0$ . We need to perform the analysis in  $(L^2(\Omega))^3$ , since  $X_{0,b} \not\subset X_0$ . In fact,  $\mathcal{A}$  is compact as a map from  $(L^2(\Omega))^3$  to  $(L^2(\Omega))^3$  as the next lemma shows.

**Lemma 12.5** *The map  $\mathcal{A} : (L^2(\Omega))^3 \rightarrow (L^2(\Omega))^3$  is compact.*

**Proof** By the Lax–Milgram lemma,  $\mathcal{A}$  is well defined and bounded as a map from  $(L^2(\Omega))^3$  into  $X_0$ . Theorem 4.7 shows that  $X_0$  is compactly embedded in  $(L^2(\Omega))^3$ . This proves the compactness of  $\mathcal{A}$ .  $\square$

Using this lemma we can see that (12.18) is a Fredholm equation on  $(L^2(\Omega))^3$ . Theorem 10.1 implies that there is at most one solution and hence (12.18) has a unique solution  $E$  in  $X$  (this is a third proof of the existence of a solution to the exterior scattering problem in this book!).

Now we write the discrete problem (12.14) as an operator equation. We define the operator  $\tilde{\mathcal{A}}_b : (L^2(\Omega))^3 \rightarrow (L^2(\Omega))^3$  as the straightforward discrete analogue of  $\mathcal{A}$ . By this we mean that for a given  $f \in (L^2(\Omega))^3$ , the function  $\tilde{\mathcal{A}}_b f \in X_{0,b}$  satisfies(12.19)

$$a + (\tilde{\mathcal{A}}_b f, \xi_h) = -2\kappa^2(f, \xi_h) - \left\langle T(\mathcal{G}^R(f)), \xi_{h,T} \right\rangle \text{ for all } \xi_h \in X_{0,h} .$$

We can also define  $F_b \in X_{0,b}$  by(12.20)

$$a + (\mathcal{F}_h, \xi_h) = \left\langle T(E^i), \xi_{h,T} \right\rangle \text{ for all } \xi_h \in X_{0,h} .$$

The operator  $\tilde{\mathcal{A}}_b$  and vector  $F_b$  are well defined by the Lax–Milgram lemma.

We can now pose the problem of finding  $\tilde{E}_b \in (L^2(\Omega))^3$  such that (12.21)

$$\tilde{E}_b + \tilde{A}_b \tilde{E}_b = \mathcal{F}_b.$$

Assuming that such a solution can be found,  $\tilde{E}_b = F_b - \tilde{A}_b \tilde{E}_b \in X_{0,b}$ . As a first step in our analysis of this problem, we need to demonstrate that as the mesh size  $b$  decreases, the discrete operator  $\tilde{A}_b$  converges pointwise to  $A$ . This is the content of the next lemma, the proof of which is rather classical (see [179]).

**Lemma 12.6** *For fixed  $f \in (L^2(\Omega))^3$ ,  $\tilde{A}_b f \rightarrow Af$  in  $X$  as  $b \rightarrow 0$ .*

**Proof** This follows by the same argument used to prove Theorem 7.11.  $\square$

As we saw in Section 7.3, the pointwise convergence of  $\tilde{A}_b$  to  $A$  is not sufficient to conclude that the operator  $(I + \tilde{A}_b)$  is invertible. We use collective compactness to provide the missing ingredient in the convergence proof (as in Section 7.3.2).

Let  $\Lambda$  be a countable set of positive real numbers whose only accumulation point is at zero. We assume that the mesh size  $b \in \Lambda$  and hence that there is a sequence of mesh sizes  $b_n \rightarrow 0$  as  $n \rightarrow \infty$ .

**Lemma 12.7** *Assuming that the mesh is regular and quasi-uniform on  $\Sigma$ , the set of operators  $\{\tilde{A}_b\}_{b \in \Lambda}$  is collectively compact considered as maps from  $(L^2(\Omega))^3$  to  $(L^2(\Omega))^3$ .*

**Proof** The proof is essentially that of Theorem 7.14 using Theorem 7.18.  $\square$

We can now analyze the operator-based problems (12.18) and (12.21) which are to find  $E \in (L^2(\Omega))^3$  and  $\tilde{E}_b \in (L^2(\Omega))^3$  such that

$$(I + A)E = \mathcal{F} \text{ and } (I + \tilde{A}_b)\tilde{E}_b = \mathcal{F}_b,$$

for  $b \in \Lambda$ . We have the following theorem.

**Theorem 12.8** *Let  $\tau_b$  be a regular triangulation of  $\Omega$  that is quasi-uniform on  $\Sigma$ . Under the conditions on the domain in Section 4.2, and assuming  $\varepsilon_r = \mu_r = 1$ , we have the following result. For  $b \in \Lambda$  sufficiently small,  $(I + \tilde{A}_b)^{-1}$  exists and is uniformly bounded as a map from  $(L^2(\Omega))^3$  to  $(L^2(\Omega))^3$ . The following error estimate*

$$\|\tilde{E}_b - E\|_{(L^2(\Omega))^3} \leq C \left( \|\mathcal{F} - \mathcal{F}_b\|_{(L^2(\Omega))^3} + \|(\tilde{A}_b - A)E\|_{(L^2(\Omega))^3} \right)$$

holds, with  $C$  independent of  $b$ ,  $E$  and  $F$ .

**Proof** This follows the proof of Theorem 7.24.  $\square$

**Theorem 12.9** *Under the conditions of Theorem 12.8 and provided  $b \in \Lambda$  is small enough, the discrete variational problem (12.14) has a unique solution  $\tilde{E}_b \in X_b$ . Furthermore,*

$$\|\tilde{E}_b - E\|_X \leq C \left( \inf_{\mathcal{X}_b \in X_b} \|\mathcal{F} - \mathcal{X}_b\|_X + \inf_{\psi_b \in X_b} \|AE - \psi_b\|_X \right).$$

In general,  $\tilde{E}_b \rightarrow E$  in  $X$  as  $b \rightarrow 0$ .

**Proof** From the previous theorem,  $\tilde{E}_b$  is proved to exist uniquely. It remains to estimate the error in  $X$ . The proof then follows that of Theorem 7.25. In this case, because of the special right-hand side, the function  $p$  appearing in Theorem 7.25 vanishes.  $\square$

This result can be made more specific provided the solution is regular enough.

Let

$$\begin{aligned} H^s(\text{curl}; \Omega) &= \left\{ u \in (H^s(\Omega))^3 \mid \nabla \times u \in (H^s(\Omega))^3, v \times u \in (H^s(f))^s \right. \\ &\quad \left. \text{for each face } f \text{ of } \Sigma \right\} \end{aligned}$$

for some  $s \geq 0$  with norm

$$\|u\|_{H^s(\text{curl}; \Omega)}^2 = \|u\|_{H^s(\Omega)}^2 + \|\nabla \times u\|_{H^s(\Omega)}^2 + \sum_{f \in \Sigma} \|v \times u\|_{H^s(\Omega)}^2.$$

Then the error estimate of Theorem (12.9) can be written as shown below.

**Corollary 12.10** *If  $F, AE \in H^s(\text{curl}; \Omega)$  for some  $s > \frac{1}{2}$ , then*

$$\|E - \tilde{E}_h\|_X \leq ch^{\min(s, \kappa)}.$$

**Remark 12.11** *For a Lipschitz polyhedral domain, the best we can generally expect is that the above regularity requirements hold for some  $s$  with  $\frac{1}{2} < s$  but possibly with  $s$  less than 1.*

## 12.2.2 The fully discrete problem

The discretization we have considered to this point is not optimal for implementation, since  $\mathcal{J}^R$  is expensive to compute. We prefer to use (12.13) in place of (12.14). Let us define  $A_b : (L^2(\Omega))^3 \rightarrow (L^2(\Omega))^3$  such that if  $f \in (L^2(\Omega))^3$ , then  $A_b f \in X_{0,b}$  satisfies (12.22)

$$a + (A_b f, \xi_h) = -2\kappa^2(f, \xi_h) - \langle T(\mathcal{J}_h(f)), \xi_h \rangle \quad \text{forall } \xi_h \in X_{0,h}.$$

Then (12.13) is equivalent to solving (12.23)

$$(I + A_h)E_h = \mathcal{F}_h.$$

In order to prove convergence, we make a specific choice of  $G_b$ . We choose  $G_b$  to interpolate  $G$  on  $\Omega$  (as a function of  $y$ ). Using this choice, we can prove the following lemma.

**Lemma 12.12** *There is a constant  $C$  such that for any  $u \in X$ ,*

$$\|(A_h - \tilde{A}_h)u\|_X \leq Ch^\kappa \|u\|_X,$$

where  $k$  is the order of the edge finite elements used to build  $X_b$ .

**Proof** By the definition of  $\mathcal{A}_b$  and  $\tilde{\mathcal{A}}_b$ ,

$$a + ((A_h - \tilde{A}_h)u, (A_h - \tilde{A}_h)u) = - \int_{\Omega} T(\mathcal{G}_h(u) - \mathcal{G}^R(u)) \cdot \overline{(A_h - \tilde{A}_h)u} dA.$$

Thus,  $\|(\mathcal{A}_b - \tilde{\mathcal{A}}_b)u\|_X \leq C \|T(\mathcal{J}_b(u) - \mathcal{J}^R(u))\|_{(L^2(\Omega))^3}$ . However, for any derivative  $D_x$  with respect to  $x$ , for any  $x \in \Sigma$ ,

$$\begin{aligned} |D_x(\mathcal{G}_h(u) - \mathcal{G}^R(u))| &= \left| \int_{\Omega} (\nabla \times D_x(\tilde{\mathbf{G}}_h - \tilde{\mathbf{G}}))^T \nabla \times u - \kappa^2 D_x(\tilde{\mathbf{G}}_h - \tilde{\mathbf{G}})^T u dV \right| \\ &\leq C \|D_x(\tilde{\mathbf{G}}_h - \tilde{\mathbf{G}})\|_X \|u\|_X. \end{aligned}$$

But since  $D_x \mathbf{G}$  is smooth when  $x \neq y$ , and  $\mathbf{G}_b$  interpolates  $\mathbf{G}$ , we may use the first interpolation estimate in Theorem 5.41 to show that  $\|D_x(\mathbf{G}_b - \mathbf{G})\|_X \leq Ch^k$  and we are done.  $\square$

Next we verify that  $(I + \tilde{\mathcal{A}}_b)$  is invertible as a map from  $X$  to  $X$ .

**Lemma 12.13** *For all  $b$  sufficiently small, the operator  $(I + \tilde{\mathcal{A}}_b)$  is invertible with a uniformly bounded inverse as a map from  $X$  to  $X$ .*

**Proof** We have already seen that this lemma holds with  $(L^2(\Omega))^3$  in place of  $X$ . Now let  $u \in (L^2(\Omega))^3$  solve  $u + \tilde{\mathcal{A}}_b u = F$  for some  $F \in X$ . Then, since  $u = F - \tilde{\mathcal{A}}_b u \in X$ , we may estimate

$$\|u\|_X \leq \|\mathcal{F}\|_X + \|\tilde{\mathcal{A}}_b u\|_X \leq \|\mathcal{F}\|_X + C \|u\|_{(L^2(\Omega))^3} \leq \|\mathcal{F}\|_X + C \|\mathcal{F}\|_{(L^2(\Omega))^3}.$$

Thus,  $\|(I + \tilde{\mathcal{A}}_b)^{-1} F\|_X \leq C \|F\|_\infty$  and we are done.  $\square$

Now we can prove that (12.23) has a unique solution that is close to the solution  $\tilde{E}_b$  of (12.14).

**Theorem 12.14** *Under the conditions on the domain, mesh and data in Theorem 12.9 and provided  $b$  is sufficiently small, eqn (12.23) (or, equivalently, (12.13)) has a unique solution  $E_b \in X_b$ , and if  $\tilde{E}_b$  is the solution of (12.14) with  $\mathbf{G}_b$  chosen to interpolate  $\mathbf{G}$ , then*

$$\|E_b - \tilde{E}_b\|_X \leq Ch^\kappa \|E_b\|_X.$$

**Remark 12.15** *As a result of this theorem, we can conclude that  $E_b$  satisfies the error estimates in Theorem 12.9 and Corollary 12.10.*

**Proof of Theorem 12.14** In Lemma 12.13 we have already verified that  $(I + \tilde{\mathcal{A}}_b)$  is invertible as a map from  $X$  to  $X$  and the inverse is uniformly bounded. Since

$$E_b + \tilde{\mathcal{A}}_b E_b + (A_h - \tilde{A}_h)E_b \simeq \mathcal{F}_h,$$

we have  $(I + C_b)E_b = (I + \tilde{\mathcal{A}}_b)^{-1} F_b$ , where  $C_b = (I + \tilde{\mathcal{A}}_b)^{-1} (A_b - \tilde{A}_b)$  and hence using Lemma 12.12  $\|C_b\|_{X \rightarrow X} \leq Ch^k < 1$  for  $b$  sufficiently small. This

implies, via Theorem 2.27, that  $(I + C_b)$  is invertible with bounded inverse in  $X$  and hence  $E_b$  exists.

We have

$$(I + \tilde{A}_h)(E_h - \tilde{E}_h) = (\tilde{A}_h - A_h)E_h$$

so that, using Lemma 12.12 and the boundedness of  $(I + \tilde{A}_h)^{-1}$ , we have

$$\|E_h - \tilde{E}_h\|_X \leq C \|(\tilde{A}_h - A_h)E_h\|_X \leq Ch^\kappa \|E_h\|_X.$$

Thus, we can conclude that Theorem 12.9 holds for  $E_b$ .  $\square$

We now show that  $E_b$  is the unique solution of the discrete problem regardless of how the discrete Green's dyadic is chosen, providing it is admissible according to Definition 12.4.

**Lemma 12.16** Suppose  $G_b$  is admissible according to Definition 12.4. Then, under the conditions of Theorem 12.9 and provided  $b$  is small enough, eqn (12.13) has a unique solution.

**Proof** Suppose  $E_b$  is the solution of (12.13) corresponding to the special choice of  $G_b$  that interpolates  $G$  (shown to be the unique solution of (12.13) with this discrete Green's dyadic by Theorem 12.14). We denote this choice of discrete Green's dyadic by  $\tilde{G}_h^{(1)}$ . Suppose that  $\tilde{G}_h^{(2)}$  is another admissible discrete Green's dyadic. Let  $\mathcal{G}_h^{(i)}$ ,  $i = 1, 2$  denote the operator in (12.12) using  $\tilde{G}_h^{(i)}$ . We first show that  $I_h^{(2)}(E_h) = I_h^{(1)}(E_h)$ . By definition

$$\begin{aligned} \left( \mathcal{G}_h^{(2)}(E_h) - \mathcal{G}_h^{(1)}(E_h) \right)^T &= \int_{\Omega} \left\{ (\nabla \times E_h)^T \nabla \times \left( \tilde{G}_h^{(2)} - \tilde{G}_h^{(1)} \right) \right. \\ &\quad \left. - \kappa^2 E_h^T \left( \tilde{G}_h^{(2)} - \tilde{G}_h^{(1)} \right) \right\} dV(y). \end{aligned}$$

Now the  $k$ th column of  $\tilde{G}_h^{(2)} - \tilde{G}_h^{(1)}$  is  $\left( \tilde{G}_h^{(2)} - \tilde{G}_h^{(1)} \right)_k = \tilde{g}_{h,l}^{(2)} - \tilde{g}_{h,l}^{(1)}$ , and since  $\tilde{g}_{h,l}^{(j)}$ ,  $j = 1, 2$  interpolates  $g$  on  $\Gamma$ , the tangential component of the difference vanishes there. Hence,  $\tilde{g}_{h,l}^{(2)} - \tilde{g}_{h,l}^{(1)} \in X_h$ , and so since  $\left( \tilde{g}_{h,l}^{(2)} - \tilde{g}_{h,l}^{(1)} \right)_l = 0$ ,  $l = 1, 2, 3$ , on  $\Sigma$  we have from the definition of  $E_b$  in (12.13) and using the test function  $\phi_h = (\tilde{g}_{h,l}^{(2)} - \tilde{g}_{h,l}^{(1)})$ ,

$$\int_{\Omega} \left\{ (\nabla \times E_h) \cdot \nabla \times \left( \tilde{g}_{h,l}^{(2)} - \tilde{g}_{h,l}^{(1)} \right) - \kappa^2 E_h \cdot \left( \tilde{g}_{h,l}^{(2)} - \tilde{g}_{h,l}^{(1)} \right) \right\} dV = 0,$$

and so  $\mathcal{G}_h^{(2)}(E_h) = \mathcal{G}_h^{(1)}(E_h)$ . Thus,  $E_b$  satisfies (12.13) with  $\mathcal{G}_h = \mathcal{G}_h^{(2)}$ . Reversing the argument, we see that if  $E_h^{(2)}$  satisfies (12.13) with  $\mathcal{G}_h = \mathcal{G}_h^{(2)}$ , then it also satisfies (12.13) with  $\mathcal{G}_h = \mathcal{G}_h^{(1)}$ . Hence by the uniqueness of the solution of (12.13) in this case,  $E_h^{(2)} = E_h$ , and we are done.  $\square$

### 12.2.3 Computational considerations

Now we shall show why (12.13) helps in the discretization of this problem. Let  $\{\xi_i\}_{i=1}^{N_h}$  be a basis for  $X_b$ . Usually this basis would be constructed using the degrees of freedom (5.33), but other choices are possible [234]. Then we can express  $E_b \in X_b$  as

$$E_b = \sum_{l=1}^{N_h} E_l \xi_l,$$

and we may write the variational equation (12.13) as a matrix equation. Let  $\vec{E} = (E_1, \dots, E_{N_h})^T$  and let  $S$  and  $L$  be  $N_b \times N_b$  matrices with

$$\begin{aligned} S_{l,m} &= (\nabla \times \xi_m, \nabla \times \xi_l) - \kappa^2(\xi_m, \xi_l) - ik\{\xi_m, T, \xi_l, T\}, \\ L_{l,m} &= -\langle T(\mathcal{J}_h(\xi_m)), \xi_l \rangle, \end{aligned}$$

for  $1 \leq l, m \leq N_b$ . Let  $\vec{F}$  be the vector with  $F_l = \langle T(E), \xi_l \rangle$  for  $1 \leq l \leq N_b$ . Then (12.24)

$$(S + L) \vec{E} = \vec{F}.$$

Our analysis guarantees that  $S + L$  is invertible for  $b$  sufficiently small, but  $S + L$  is not particularly well structured from the point of view of numerical linear algebra. It is non-definite and non-symmetric.

The matrix  $S$  is somewhat better behaved than  $L$ . It is sparse and symmetric (but not Hermitian). It corresponds to the standard discretization of the interior problem studied in Chapter 7 and is also invertible for  $b$  sufficiently small. In general,  $S$  has  $O(N_b)$  non-zero entries.

If we choose  $G_b$  to interpolate zero away from  $\Gamma$ , then  $\mathcal{J}_b(\xi)$  vanishes when  $\xi$  is zero on all tetrahedra sharing an edge with  $\Gamma$ . Thus,  $L_{l,m} \neq 0$  only if  $\xi_l$  is associated with an edge or face on  $\Sigma$  and  $\xi_m$  is associated with a tetrahedron touching  $\Gamma$ . For a quasi-uniform mesh, we expect  $O(N_b^{2/3})$  edges and faces on  $\Sigma$  and  $O(N_b^{2/3})$  tetrahedra to touch  $\Gamma$ . Hence,  $L$  has  $O(N_b^{4/3})$  non-zero entries which is far more than  $S$ . Thus,  $L$  is very expensive to compute and store. This suggests that (12.24) should be solved by an iterative technique (e.g. GMRES) and then only the action of  $L$  needs to be computed. We expect that this can be computed rapidly using the fast multipole method [263] to yield a fast overall solver. In fact, Liu and Jin [212] have done this using a method that is closely related to the one outlined in this section. Liu and Jin divide the computational domain into two subdomains separated by a surface  $C$  containing the scatterer  $D$  in its interior. Then they use a variational formulation computing  $E$  inside  $C$  and  $H$  outside  $C$ . The integral representation (12.2) is used on  $C$  to obtain the boundary condition on the artificial boundary  $\Sigma$ . Since  $E$  has a tangential trace from inside  $C$  and  $H$  has a tangential trace from outside  $C$ , enough data are available on  $C$  to apply (12.2) directly. They then solve the coupled problems

by an iterative scheme. Applied to our case, this iterative scheme would require to guess  $\vec{E}^{(0)}$  and then compute, for some  $\gamma$  with  $0 < \gamma \leq 1$ , and  $n = 0, 1, 2, \dots$ ,

$$\vec{E}^{(n+1)} = \vec{E}^{(n)} - \gamma \left[ (I + A^{-1}L) \vec{E}^{(n)} - A^{-1} \vec{F} \right].$$

At each step of this iterative scheme, the term  $L \vec{E}^{(n)}$  can be evaluated using the fast multipole method (without the need for evaluating near interactions in the fast multipole method). Then  $A^{-1}(\vec{F} - L \vec{E}^{(n)})$  is evaluated by solving the interior finite element problem (in Liu and Jin's case using a multi-frontal solver). At least for Liu and Jin's formulation, fast convergence is observed.

## 12.3 Perfectly conducting half space

Next we consider the case where the electromagnetic field is confined to the upper half space denoted by

$$\mathbb{R}_+^3 = \{x \in \mathbb{R}^3 \mid x_3 > 0\}.$$

The lower half space is assumed to be occupied by a perfect conductor and that the scattered electric field  $E^s$  in the upper half space satisfies the boundary condition(12.25)

$$\nu \times E^s = 0 \text{ on } \Sigma_0,$$

where  $\Sigma_0 = \{x \in \mathbb{R}^3 \mid x_3 = 0\}$ . We denote the lower half space by  $\mathbb{R}_-^3 = \{x \in \mathbb{R}^3 \mid x_3 < 0\}$ .

The choice of boundary condition in (12.25) requires some comment. We shall use an incident field  $E^i$  that satisfies the perfect conducting boundary condition on  $\Sigma_0$ . Thus  $E^i \times \nu = 0$  on  $\Sigma_0$  and so if the total field  $E = E^i + E^s$ , then  $E^s \times \nu = (E^i + E^s) \times \nu = E \times \nu = 0$ .

Our goal is to obtain an integral representation of the electric field in  $\mathbb{R}_+^3$  outside any scatterers present there. Let  $D \subset \mathbb{R}_+^3$  be a bounded Lipschitz domain with connected complement such that  $D \subset \mathbb{R}_+^3$ . We suppose that  $E^s$  satisfies Maxwell's equations in  $\Omega = \mathbb{R}_+^3 \setminus D$  so that(12.26)

$$\nabla \times \nabla \times E^s - \kappa^2 E^s = 0 \text{ in } \mathbb{R}_+^3 \setminus D,$$

with the perfectly conducting boundary condition (12.25) on  $\Sigma_0$  and the Silver–Müller radiation condition (9.14) holding uniformly for all directions in  $\partial B_1 \cap \mathbb{R}_+^3$  (i.e. the upper half of the unit sphere).

This field can be represented by an integral of the form (12.2) provided we obtain a suitable dyadic Green's function for the half space problem. In particular we seek a  $3 \times 3$  matrix function  $G_{\text{pec}}(x, y)$  such that for each fixed  $\mathbb{R}_+^3$ ,(12.27)

$$\nabla_y \times (\nabla_y \times G_{\text{pec}}) - \kappa^2 G_{\text{pec}} = \delta_x \mathbb{I} \text{ for } y \in \mathbb{R}_+^3.$$

Here  $\nabla_y \times G_{\text{pec}}$  is understood column by column as in the previous section. In addition, again defining the indicated quantity column by column(12.28)

$$\nu \times G_{\text{pec}} = 0 \text{ on } \Sigma_0,$$

$$\lim_{P_y \rightarrow \infty} \rho y ((\nabla \times \mathbb{G}_{\text{pec}}) \times \hat{y} - i\kappa \mathbb{G}_{\text{pec}}) = 0, \quad (12.29)$$

where the limit holds uniformly for  $x$  in a compact subset of  $\mathbb{R}_+^3$  and for all directions  $\hat{y}$  in the upper half of the unit sphere.

This problem is considered by Sommerfeld in [273] where equations (12.27)–(12.29) are suggested as a model of a dipole antenna over sea water (obviously a calm day!). Following Sommerfeld, we can construct  $\mathbb{G}_{\text{pec}}$  from our existing Green's dyadic  $\mathbb{G}$ . To do this, we shall need to use mirror image points reflected by the plane  $\sum_0$  so that (12.30)

$$\text{if } x = (x_1, x_2, x_3)^T \in \mathbb{R}_+^3 \text{ we define } x' = (x_1, x_2 - x_3)^T.$$

Suppose  $\mathbb{G}_{\text{pec}}$  has columns  $g_{\text{pec},1}$ ,  $g_{\text{pec},2}$  and  $g_{\text{pec},3}$  then (12.31a)

$$\begin{aligned} g_{\text{pec},1}(x, y) &= g_1(x, y) - g_1(x', y), \\ g_{\text{pec},2}(x, y) &= g_2(x, y) - g_2(x', y), \end{aligned} \quad (12.31b)$$

$$g_{\text{pec},3}(x, y) = g_3(x, y) + g_3(x', y), \quad (12.31c)$$

Thus,

$$\mathbb{G}_{\text{pec}}(x, y) = \mathbb{G}(x, y) - \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \mathbb{G}(x', y).$$

To verify that this Green's dyadic actually solves (12.27)–(12.29), we note that both  $\mathbb{G}(x, \cdot)$  and  $\mathbb{G}(x', \cdot)$  satisfy (12.27) for  $y \neq x$ , and  $\mathbb{G}(x', \cdot)$  is smooth for all  $y \in \mathbb{R}_+^3$  so only  $\mathbb{G}(x, \cdot)$  contributes to the singularity at  $y = x$  in (12.27). Also each term of  $\mathbb{G}_{\text{pec}}$  satisfies the radiation condition (12.29). Finally, to show that (12.28) is satisfied, we can perform a direct calculation (by MAPLE preferably) or use the diagram in Fig. 12.1 (similar to Fig. 27 from [273]).

Now using essentially the proof of Theorem 12.2 (and using an argument like that in the proof of Theorem 9.1), we can obtain the following result.

**Theorem 12.17** Suppose  $E^s \in H_{\text{loc}}(\text{curl}; \mathbb{R}^3 \setminus \bar{D})$  is a radiating solution of the homogeneous Maxwell's equations in  $\mathbb{R}_+^3 \setminus D$  satisfying the perfectly conducting boundary condition on  $\sum_0$ . Then for each  $x \in \mathbb{R}_+^3 \setminus D$ , the representation formula (12.2) holds with  $\mathbb{G}$  replaced by  $\mathbb{G}_{\text{pec}}$  and  $\mathbb{R}^3 \setminus D$  replaced by  $\mathbb{R}_+^3 \setminus \bar{D}$ .

Now we describe two commonly used incident waves. The incident waves must satisfy Maxwell's equations in the background medium (i.e. in the upper half plane) and obey the perfectly conducting boundary condition on  $\sum_0$ .

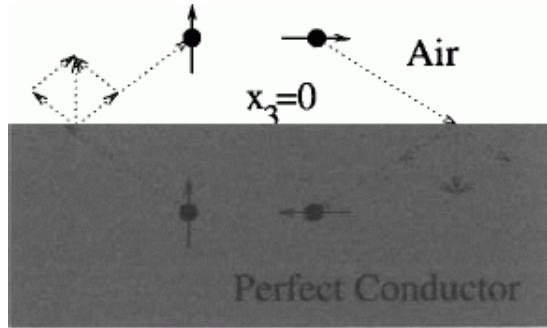
- (1) *Point Source* We suppose that a dipole point source with polarization  $p \in \mathbb{R}^3$ ,  $|p| \neq 0$ , is located at  $x_p \in \mathbb{R}_+^3 \setminus \bar{D}$ . The field due to this dipole point source is given by (12.32)

$$E^i(x) = \mathbb{G}_{\text{pec}}(x_p, x)p.$$

Clearly, this field satisfies (12.33)

$$\nabla \times (\nabla \times E^i) - \kappa^2 E^i = p \delta_{xp}, \quad \text{in } \mathbb{R}_+^3,$$

Fig. 12.1. How the dipole source and its mirror image sum to produce a field satisfying the perfect conducting boundary condition on  $\Sigma_0$  (i.e. where  $x_3 = 0$ ). Two cases are shown: a vertically polarized dipole, and a horizontally polarized dipole. The vector construction indicates how the fields due to the dipole at  $x'$  and its mirror point cancel on  $\Sigma_0$ . This explains why the sign change is needed comparing the expressions for  $g_{\text{pec},1}$  in (12.31a) and  $g_{\text{pec},3}$  in (12.31c).



(12.34)

$$\nu \times E^i = 0 \text{ on } \Sigma_0,$$

(12.35)

$$\lim_{x \rightarrow \infty} \rho \left( (\nabla \times E^i) \times \hat{x} - i\kappa E^i \right) = 0.$$

Thus,  $E^i$  satisfies the homogeneous Maxwell's equations in  $D$  and in a neighborhood of  $D$  and the perfectly conducting boundary condition on  $\Sigma_0$ . It is important that the incident field satisfies Maxwell's equations and the boundary condition on  $\Sigma_0$  for the method we shall describe to work. The Silver–Müller radiation condition is specified just to allow a unique identification of  $E^i$  in terms of  $G_{\text{pec}}$ . Of course, by adding or integrating incident fields from point sources, we can also handle multiple sources or even a distributed current density.

- (2) *Plane Wave* If the source point  $x_p$  is very far from the scatterer  $D$ , the incident waves are approximately plane waves. The basic plane wave (1.20) has to be modified to allow for reflection at  $\Sigma_0$  to satisfy (12.34). In this case, we have the incident wave(12.36)

$$E^i(x) = p \exp(i\kappa x \cdot d) - p' \exp(i\kappa x \cdot d').$$

Here the polarization  $p$  and direction of propagation  $d$  satisfy  $|d| = 1$ ,  $p \neq 0$  and  $p \cdot d = 0$ . It is easy to see that

$$\nabla \times (\nabla \times E^i) - \kappa^2 E^i = 0 \text{ in } \mathbb{R}_+^3.$$

If  $x_3 = 0$ , taking into account that  $\nu = (0, 0, 1)^T$ ,

$$\nu \times E^i = \nu \times (p - p') \exp(i\kappa(x_1 d_1 + x_2 d_2)) = 0.$$

Thus, the incident field given in (12.36) satisfies Maxwell's equations in a neighborhood of  $D$  together with the perfectly conducting boundary condition on  $\Sigma_0$  as required for our integral representation.

Now given an incident field  $E^i$  defined by (12.32) or (12.36) the total field  $E$  and scattered field  $E^s$  satisfy(12.37a)

$$\begin{aligned} \nabla \times (\nabla \times E) - \kappa^2 E &= F \text{ in } \mathbb{R}_+^3 \setminus D, \\ v \times E &= 0 \text{ on } \Gamma = \Gamma \text{ and on } \Sigma_0, \end{aligned} \quad (12.37b)$$

$$E = E^i + E^s \text{ in } \mathbb{R}_+^3 \setminus D, \quad (12.37c)$$

$$\lim_{p \rightarrow \infty, \hat{x} \in \mathbb{R}_+^3} \rho(\{\nabla \times E^s\} \times \hat{x} - i\kappa E^s) = 0, \quad (12.37d)$$

where  $F = p\delta_{x_0}$ ,  $x_0 \notin D$ , if (12.32) is used and  $F = 0$  if (12.36) is used. Note that on  $\Sigma_0$

$$0 = v \times E = v \times (E^i + E^s) = v \times E^s,$$

so  $v \cdot E^s$  also satisfies the perfectly conducting boundary condition on  $\Sigma_0$ . Thus,  $E^s$  can be represented by the integral formula in Theorem 12.17 which is our goal. We could now proceed to verify uniqueness of the solution of (12.37). Truncating the problem as in the previous section with  $G$  replaced by  $G_{\text{pec}}$  (see (12.11)), and then applying the Fredholm theory as in the previous section, we could verify existence of a solution to this problem. The finite element method given by (12.13) can then be used (and proved to converge), provided the finite element mesh is contained in  $\mathbb{R}_+^3$  and  $G$  is replaced by  $G_{\text{pec}}$ .

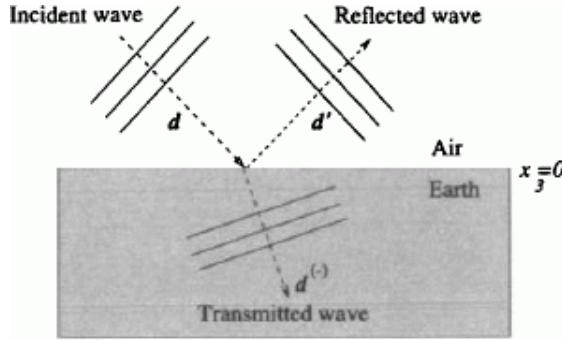
## 12.4 Layered medium

In this section, we study scattering from objects in a layered background medium. For simplicity (it also corresponds to our research interests), we will only study a medium with two layers. For multiple layers, see [288, 248, 67]. We start by deriving a special solution of Maxwell's equations in a layered medium (without scatterers). This will later be used as an incident or incoming wave for the scattering calculation. Next we derive the dyadic Green's function for the layered medium. As in the previous sections of this chapter, each column of the dyadic Green's function is the solution of Maxwell's equations due to a dipole point source. Finally, we use the dyadic Green's function in a representation theorem as in Corollary 12.2.

### 12.4.1 Incident plane waves

We are going to calculate how a plane waves interact with the layered medium. This solution will serve as an incident field for the scattering problem. We emphasize that we must always use incident fields that satisfy the Maxwell's equations for the background medium. The scattered field is then the perturbation of the incident field due to the scatterer alone (some authors differ on the splitting of incident and scattered fields). The geometry of the problem is shown in Fig. 12.2 .

Fig. 12.2. The directions and geometry of scattering of plane waves from the interface between two different media in the upper and lower half space.



A plane wave propagating with direction vector  $d$  ( $|d| = 1$ ) is incident from above (so  $d_3 < 0$ ). It gives rise to a reflected wave with direction  $d'$  and a transmitted wave with direction  $d^{(-)}$ .

For now we assume that  $d \times e_3 \neq 0$ . Then we can define vectors  $l$  and  $m$  by

$$l = \frac{d \times e_3}{|d \times e_3|} \text{ and } m = d \times l.$$

The three-tuple  $(d, l, m)$  forms an orthonormal coordinate system in  $\mathbb{R}^3$ . In the absence of the layered medium, the incident field (which must be polarized orthogonal to  $d$ ) can be written as  $(\alpha_0 l + \beta_0 m) \exp(i\kappa x \cdot d)$ , where  $\kappa$  is the wavenumber in the region  $x_3 > 0$  (assumed real and positive). With the layer present, this field will be reflected and transmitted at the interface  $x_3 = 0$  and we want to compute this field in  $\mathbb{R}^3$ . By linearity we can consider the two polarizations  $l$  and  $m$  separately.

*Parallel incidence* In this case  $\alpha_0 = 0$  and  $\beta_0 \neq 0$ . Again, in the absence of the layer, the incident magnetic field is given by

$$H = \frac{1}{ik} \nabla \times (\beta_0 m \exp(i\kappa x \cdot d)) = \beta_0 d \times m \exp(i\kappa x \cdot d) = \beta_0 l \exp(i\kappa x \cdot d).$$

Thus, the polarization of the incident magnetic field is parallel to the plane  $x_3 = 0$  (hence the term “parallel incidence”) and it turns out to be easier to work in terms of the magnetic field. In the presence of the layer, the magnetic field  $H$ , now including reflected and transmitted components, is given by(12.38)

$$H^i(x) = \begin{cases} -\beta_0 l \exp(i\kappa x \cdot d) - \beta_1 l \left( i\kappa x \cdot d' \right) & \text{if } x_3 > 0, \\ -\beta_0 l \exp \left( i\kappa n x \cdot d^{(-)} \right) & \text{if } x_3 < 0. \end{cases}$$

Here  $d = (d_1, d_2, d_3)^\top$  and, as usual, the image point  $\tilde{d} = (d_1, d_2, -d_3)^\top$ . The index of refraction of the lower half plane is  $n = \sqrt{\epsilon_r}$  with  $\Im(\sqrt{\epsilon_r}) \geq 0$ . The vector  $d'$   $\in \mathbb{C}^3$  satisfies  $d' \cdot d' = 1$ . The unknown coefficients  $\beta_1$  and  $\beta_2$  measure the magnitude of the reflected and transmitted waves respectively.

To determine  $\beta_1$ ,  $\beta_2$  and  $d'$ , we impose the continuity conditions. Continuity of  $e_3 \times H$  on  $\Sigma_0$  implies

$$\begin{aligned} & (\beta_0 \tilde{z} \times l \exp(i\kappa x \cdot d) + \beta_1 \tilde{z} \times l \exp(i\kappa x \cdot d')) \Big|_{x_3=0} \\ &= \beta_2 \tilde{z} \times l \exp(i\kappa n x \cdot d^{(-)}) \Big|_{x_3=0}. \end{aligned}$$

Since this must hold for all  $x_1$  and  $x_2$  we need(12.39)

$$\kappa d_1 = \kappa n d_1^{(-)} \text{ and } \kappa d_2 = \kappa n d_2^{(-)}.$$

From this we can compute  $d^{(+)}$ , since

$$\left(d_3^{(-)}\right)^2 = 1 - \left(\frac{1}{n}\right)^2 (d_1^2 + d_2^2) = \left(1 - \frac{1}{\epsilon_r}\right) + \frac{d_3^2}{\epsilon_r}.$$

Hence,  $d_3^{(-)}$  is determined by requiring that  $\Im(\sqrt{\epsilon_r} d_3^{(-)}) \geq 0$  and that the sign of  $\Re(d_3^{(-)})$  and  $\Re(d_3)$  agree so that the transmitted wave propagates downwards and does not grow as  $x_3$  tends to  $-\infty$ .

In addition, we need(12.40)

$$\beta_0 + \beta_1 = \beta_2.$$

Next we impose the boundary condition on  $E^i = -(1/i\kappa\epsilon_r)\nabla \times H^i$  (where  $\epsilon_r = 1$  if  $x_3 > 0$ ). Using (12.39), the continuity of  $e_3 \times E^i$  at  $x_3 = 0$  implies

$$\beta_0 e_3 \times (d \times l) + \beta_1 e_3 \times (d' \times l) = \frac{\beta_2}{n} e_3 \times (d^{(-)} \times l).$$

Taking the dot product with  $l$  and using the fact that  $e_3 \cdot l = 0$  gives(12.41)

$$\beta_0 e_3 \cdot d + \beta_1 e_3 \cdot d' = \frac{\beta_2}{n} e_3 \cdot d^{(-)}$$

Solving (12.40) and (12.41) for  $\beta_1$  and  $\beta_2$  gives

$$\frac{\beta_1}{\beta_0} = \frac{d_3 - d_3^{(-)}}{d_3 + d_3^{(-)}} / n, \quad \frac{\beta_2}{\beta_0} = \frac{2d_3}{d_3 + d_2^{(-)}} / n.$$

This completes our determination of  $H^i$  (and hence  $E^i$ ) under parallel incidence.

*Perpendicular Incidence:* Now we consider the case when  $\beta_0 = 0$  and  $a_0 \neq 0$ . In the absence of the layer the full incident electric field is given by  $a_0 / \exp(i\kappa x \cdot d)$ . Then in the presence of the layer the full incident field, including transmitted and reflected waves, is given by(12.42)

$$E^i = \begin{cases} -a_0 l \exp(i\kappa x \cdot d) + a_1 l (i\kappa x \cdot d') & \text{if } x_3 > 0 \\ a_2 l \exp(i\kappa n x \cdot d^{(-)}) & \text{if } x_3 < 0, \end{cases}$$

with the same notation as in the previous section. Continuity of  $e_3 \times E$  at  $x_3 = 0$  implies that

$$\begin{aligned} & (a_0 e_3 \times l \exp(i\kappa x \cdot d) + a_1 e_3 \times l \exp(i\kappa x \cdot d')) \Big|_{x_3=0} \\ &= a_2 e_3 \times l \exp(i\kappa n x \cdot d^{(-)}) \Big|_{x_3=0}. \end{aligned}$$

Hence, as before, (12.39) holds and we are correct in using the same notation  $\mathbf{d}^\top$  in both cases. In addition, similarly to (12.40), we have

$$\alpha_0 + \alpha_1 = \alpha_2.$$

But  $\mathbf{H} = (1/i\kappa)\nabla \times \mathbf{E}$ , so continuity of  $e_3 \times \mathbf{H}$  at  $x_3 = 0$  implies

$$\alpha_0 e_3 \times (d \times l) + \alpha_1 e_3 \times (d' \times l) = \alpha_2 n e_3 \times (d^{(-)} \times l).$$

Proceeding as in the previous section we compute

$$\frac{\alpha_1}{\alpha_0} = \frac{d_3 - nd_3^{(-)}}{d_3 + nd_3^{(-)}}, \quad \frac{\alpha_2}{\alpha_0} = \frac{2d_3}{d_3 + nd_3^{(-)}}.$$

This completes our determination of plane wave scattering by a plane interface. The incident field including transmitted and reflected components is considered to be the “incident wave” for this formulation.

## 12.4.2 The dyadic Green's function

Next we turn our attention to computing the dyadic Green's function for the layered medium. First we need to understand the free space Green's dyadic  $G$  a little more. Suppose  $\mathbf{u}$  is any locally smooth solution of the vector Helmholtz equation so that  $\nabla \mathbf{u} + \kappa^2 \mathbf{u} = 0$  component-wise. Then if we define

$$\mathbf{v} = \mathbf{u} + \frac{1}{k^2} \nabla \nabla \cdot \mathbf{u}$$

and use the fact that  $\nabla \times \nabla \times \mathbf{v} = -\Delta \mathbf{v} + \nabla \nabla \cdot \mathbf{v}$  we can easily verify that  $\mathbf{v}$  is a divergence free solution of Maxwell's equations  $\nabla \times (\nabla \times \mathbf{v}) - \kappa^2 \mathbf{v} = 0$ . The vector  $\mathbf{u}$  is called a *Hertz* vector. For free space, the Green's dyadic is  $G$ , and we see that the first column  $\mathbf{g}_1$  is given by

$$\mathbf{g}_1 = \Pi + \frac{1}{k^2} \nabla \nabla \cdot \Pi,$$

where the Hertz vector  $\Pi$  is given by  $\Pi = (\Phi(\mathbf{x}, \mathbf{y}), 0, 0)^\top$ . A similar representation can be given for the other two columns of  $G$ .

Thus, the columns of the free space dyadic Green's function can be constructed from solutions of the vector Helmholtz equation which simplifies computing the expressions for this dyadic. Of course, this is essentially a formal process and once the Green's dyadic has been computed, it is then necessary to check its properties to ensure that the formal approach has computed the desired matrix. This is the approach followed by Sommerfeld [273] in the case we are considering, and generalized in [288] for multiple stratifications.

Let us denote by  $G_L$  the Green's dyadic for the layered medium with the  $\ell$ th column denoted by  $\mathbf{g}_{L,\ell}$ ,  $1 \leq \ell \leq 3$ . We consider two cases: the first is a vertically polarized dipole source ( $\ell = 3$ ) and the second more complex case is horizontally polarized ( $\ell = 1, 2$ ). We follow [273] but modify the result in order to obtain the dyadic.

First we wish to compute the third column of the dyadic Green's function  $G_L$ , denoted by  $\mathbf{g}_{L,3}$ , which satisfies(12.43)

$$\nabla_y \times (\nabla_y \times \mathbf{g}_{L,3}) - k^2 n^2 \mathbf{g}_{L,3} = e_3 \delta_x, \quad (12.44)$$

$$\lim_{x \rightarrow \infty} \int_{B_R} \left| (\nabla_y \times \mathbf{g}_{L,3}) \times \hat{\mathbf{y}} - ikn \mathbf{g}_{L,3} \right|^2 dA(y) = 0,$$

where  $\hat{\mathbf{y}} = \mathbf{y}/|\mathbf{y}|$ , and  $n(\mathbf{y}) = n(y_3)$ , is the index of refraction given by

$$n(y) \begin{cases} 1 & \text{if } y_3 > 0, \\ \sqrt{\epsilon_r^e} & \text{if } y_3 < 0 \quad (\text{positive imaginary part}) . \end{cases}$$

Here we are using  $\mathbf{y}$  as the independent variable, since, by tradition,  $\mathbf{x}$  denotes the position of the source (vertically polarized since it is in the direction  $e_3$ ).

Because the source is normal to the interface  $\Sigma_0$ , symmetry considerations suggest that the Hertz vector  $\Pi$  for  $\mathbf{g}_{L,3}$  will be  $\Pi = (0, 0, \Pi_3)^\top$ . Near the source, we wish  $\Pi$  to have the singular behavior of  $\Phi(\mathbf{x}, \mathbf{y})$ , but this needs to be corrected by a “secondary field” to allow for reflections from the interface at  $y_3 = 0$ . So we write

$$\Pi_3(y) \begin{cases} \Pi_3^+(y) & \text{if } y_3 > 0, \\ \Pi_3^-(y) & \text{if } y_3 < 0, \end{cases}$$

and  $\Pi_3^+ = \Phi(x, y) + \tilde{\Pi}_3^+(y)$ , where  $\tilde{\Pi}_3^+$  is a smooth solution of the Helmholtz equation

$$\Delta \tilde{\Pi}_3^+ + \kappa^2 \tilde{\Pi}_3^+ = 0 \text{ in } \mathbb{R}_+^3 .$$

Similarly,  $\Pi_3^-$  is a smooth solution of

$$\Delta \Pi_3^- + \kappa^2 \Pi_3^- = 0 \text{ in } \mathbb{R}_-^3 .$$

Using cylindrical polar coordinates with origin at  $(\mathbf{x}_1, \mathbf{x}_2, 0)^\top$  having coordinates denoted by  $(\rho, \theta, y_3)$ , where

$$\rho = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2},$$

we see that  $\tilde{\Pi}_3^+$  and  $\Pi_3^-$  must satisfy(12.45a)

$$\left( \frac{1}{\rho} \frac{\partial}{\partial \rho} \left( \rho \frac{\partial}{\partial \rho} \right) + \frac{\partial^2}{\partial y_3^2} - \kappa^2 \right) \tilde{\Pi}_3^+ = 0 \text{ in } \mathbb{R}_+^3,$$

(12.45b)

$$\left( \frac{1}{\rho} \frac{\partial}{\partial \rho} (\rho \frac{\partial}{\partial \rho}) + \frac{\partial^2}{\partial y_3^2} - \kappa^2 n^2 \right) \Pi_3^- = 0 \text{ in } \mathbb{R}_-^3.$$

Here we have used symmetry to conclude that  $\bar{\Pi}_3^+ = \bar{\Pi}_3(\rho, y_3)$  and  $\Pi_3^- = \Pi_3^-(\rho, y_3)$  (i.e. there is no dependence on the angle  $\theta$  of the polar coordinates). If we define

$$\mu_+ = \sqrt{\lambda^2 - \kappa^2} \text{ and } \mu_- = \sqrt{\lambda^2 - \kappa^2 n^2},$$

with real part of  $\mu_{\pm}$  positive (take care:  $\mu_{\pm}$  is not magnetic permeability!), we see that the equations in (12.45) have linearly independent solutions,

$$\begin{aligned} & \exp(\mu \pm y_3) J_0(\lambda \rho), \quad \exp(-\mu \pm y_3) J_0(\lambda \rho), \\ & \exp(\mu \pm y_3) Y_0(\lambda \rho), \quad \exp(-\mu \pm y_3) Y_0(\lambda \rho), \end{aligned}$$

where  $J_0$  and  $Y_0$  are cylindrical Bessel functions of order zero (see [93]). Because the functions  $\bar{\Pi}_3^+$  and  $\Pi_3^-$  are bounded at  $\rho = 0$ , we reject the solution involving  $Y_0$ . Furthermore, for  $\bar{\Pi}_3^+$  to be bounded as  $y_3 \rightarrow -\infty$ , we must choose the negative exponential solution and write by superposition (a similar argument picks the function for  $\Pi_3^-$ ) (12.46a)

$$\begin{aligned} \Pi_3^+ &= \Phi(x, y) + \int_0^\infty a(\lambda) J_0(\lambda \rho) \exp(-\mu_+(x_3 + y_3)) d\lambda, \\ & \quad (12.46b) \end{aligned}$$

$$\Pi_3^- = \int_0^\infty b(\lambda) J_0(\lambda \rho) \exp(\mu_- y_3 - \mu_+ x_3) d\lambda.$$

Here we have introduced a convenient factor  $\exp(-\mu_+ x_3)$  independent of  $y$  in both integrals. To complete our determination of  $\mathbf{g}_{L,3}$  when  $x_3 > 0$ , we need to express  $\Phi(x, y)$  as an integral (Fourier-Bessel expansion). From Sommerfeld [273] (see Sections 21B and 31B),

$$\frac{\exp(i\kappa|x-y|)}{4\pi|x-y|} = \frac{1}{4\pi} \int_0^\infty J_0(\lambda \rho) \exp(-\mu_+ |x_3 - y_3|) \frac{\lambda}{\mu_+} d\lambda,$$

where, as before,  $\rho = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$  and  $\mu_+ = \sqrt{\lambda^2 - \kappa^2}$  (positive real part). Thus, for  $y_3 > 0$ , (12.47)

$$\begin{aligned} \Pi_3^+ &= \int_0^\infty \left( \frac{1}{4\pi} \exp(-\mu_+ |x_3 - y_3|) \frac{\lambda}{\mu_+} \right. \\ & \quad \left. + a(\lambda) \exp(-\mu_+ (x_3 + y_3)) \right) J_0(\lambda \rho) d\lambda. \end{aligned}$$

Next, we determine  $a(\lambda)$  and  $b(\lambda)$  from the transmission conditions at  $y_3 = 0$ . Recalling that (12.48a)

$$\mathbf{g}_{L,3} = \mathbf{\Pi} + \frac{1}{\kappa^2} \nabla \nabla \cdot \mathbf{\Pi} \text{ if } y_3 > 0,$$

$$\mathbf{g}_{L,3} = \Pi + \frac{1}{\kappa^2 \epsilon_r^e} \nabla \nabla \cdot \Pi \text{ if } y_3 > 0, \quad (12.48b)$$

and using the fact that  $\Pi$  has only a third component depending on  $Q$  and  $y_3$ , we see that the continuity of  $e_3 \times \mathbf{g}_{L,3}$  and  $e_3 \times (\nabla \times \mathbf{g}_{L,3})$  is implied by the conditions

$$\begin{aligned} \frac{1}{\kappa^2} \frac{\partial}{\partial \rho} \frac{\partial}{\partial y_3} \Pi_3^+ &= \frac{1}{\kappa^2 \epsilon_r^e} \frac{\partial}{\partial \rho} \frac{\partial}{\partial y_3} \Pi_3^- \text{ at } y_3 = 0, \\ \frac{\partial}{\partial \rho} \Pi_3^+ &= \frac{\partial}{\partial \rho} \Pi_3^- \text{ at } y_3 = 0. \end{aligned}$$

Integrating with respect to  $Q$ , we see that these conditions are satisfied if

$$\frac{\partial}{\partial y_3} \Pi_3^+ = \frac{1}{\epsilon_r^e} \frac{\partial}{\partial y_3} \Pi_3^- \text{ and } \Pi_3^+ = \Pi_3^- \text{ at } y_3 = 0.$$

Hence, using (12.46) and (12.47), we require

$$\frac{1}{4\pi} \frac{\lambda}{\mu_+} + a(\lambda) - b(\lambda) = 0 \text{ and } \frac{1}{4\pi} \lambda - \mu_+ a(\lambda) - \frac{\mu_-}{\epsilon_r^e} b(\lambda) = 0.$$

These equations can be solved to obtain

$$a(\lambda) = \frac{\lambda}{4\pi\mu_+} \left( 1 - \frac{2\mu_-}{\epsilon_r^e \mu_+ + \mu_-} \right) \text{ and } b(\lambda) = \frac{\epsilon_r^e \lambda}{2\pi(\epsilon_r^e \mu_+ + \mu_-)}.$$

Hence,

$$\begin{aligned} \Pi_3^+ &= \int_0^\infty \left\{ \frac{\lambda}{4\pi\mu_+} \exp(-\mu_+|x_3 - y_3|) + \frac{\lambda}{4\pi\mu_+} \exp(-\mu_+(x_3 + y_3)) \right. \\ &\quad \left. - \frac{2\mu_- \lambda}{4\pi\mu_+(\epsilon_r^e \mu_+ + \mu_-)} \exp(-\mu_+(x_3 + y_3)) \right\} J_0(\lambda\rho) d\lambda \end{aligned}$$

and this can be rewritten, for  $x_3 > 0, y_3 > 0$ , as

$$\begin{aligned} \Pi_3^+ &= \Phi(x, y) + \Phi(x', y) \\ &\quad - \frac{1}{2\pi} \int_0^\infty J_0(\lambda\rho) \exp(-\mu_+(x_3 + y_3)) \left( \frac{\mu_-}{\epsilon_r^e \mu_+ + \mu_-} \right) \frac{\lambda}{\mu_+} d\lambda. \end{aligned}$$

As we might expect,  $\mathbf{g}_{L,3}$  is a perturbation of the third column of the Green's dyadic  $G_{pec}$  for a perfectly conducting half space problem.

For  $x_3 > 0$  and  $y_3 < 0$ ,

$$\Pi_3^- = \frac{1}{2\pi} \int_0^\infty \frac{\epsilon_r^e \lambda}{\epsilon_r^e \mu_+ + \mu_-} J_0(\lambda\rho) \exp(\mu_- y_3 - \mu_+ x_3) d\lambda.$$

A similar calculation gives  $\mathbf{g}_{L,3}$  for  $x_3 < 0$  and any  $y_3$ . We obtain

$$\Pi_3^+ = \frac{1}{2\pi} \int_0^\infty J_0(\lambda\rho) \exp(-\mu_+ y_3 + \mu_- x_3) \frac{\lambda}{\mu_- + \mu_+ \epsilon_r^e} d\lambda \text{ for } y_3 > 0,$$

$$\begin{aligned} \Pi_3^- = & \frac{\exp(i\kappa n|x-y|)}{4\pi|x-y|} + \frac{\exp(i\kappa n|x'-y|)}{4\pi|x'-y|} \\ & - \frac{1}{2\pi} \int_0^\infty J_0(\lambda\rho) \exp(\mu_-(y_3 + x_3)) \frac{\epsilon_r^e \mu_+}{\mu_-(\mu_- + \mu_+ \epsilon_r^e)} d\lambda \text{ for } y_3 < 0, \end{aligned}$$

where we recall  $n = \sqrt{\epsilon_r^e}$  for  $y_3 < 0$ .

Now we want to compute the field due to a horizontal dipole. With no loss of generality, we can assume that the dipole is directed along the  $e_1$ -axis. The other case, directed along the  $e_2$ -axis, is obtained by rotation. Thus, we want to compute  $\mathbf{g}_{L,1}$  which satisfies (12.49a)

$$\nabla_y \times (\nabla_y \times \mathbf{g}_{L,1}) - \kappa^2 n g_{L,1} = e_1 \delta_x \text{ in } \mathbb{R}^3,$$

$$\lim_{x \rightarrow \infty} \int_{\partial B_R} \left| (\nabla_y \times \mathbf{g}_{L,1}) \times \hat{y} - i\kappa n g_{L,1} \right|^2 dA(y) = 0, \quad (12.49b)$$

where  $\mathbf{g}_{L,1}$  is considered as a function of  $y$  and  $x \in \mathbb{R}_+^3$ , is fixed. In this case, we cannot simply assume that the Hertz vector  $\Pi$  has one component as we did in the previous section. Instead, it turns out that  $\Pi = (\Pi_1, 0, \Pi_3)^\top$ . As before, we separate into primary (singular fields) and secondary fields. We write

$$\Pi_1 = \begin{cases} \Pi_1^+ = \Phi(x, y) + \tilde{\Pi}_1^+ & \text{if } y_3 > 0, \\ \Pi_1^- & \text{if } y_3 < 0. \end{cases}$$

Here we have again used the fundamental solution of the Helmholtz equation to provide the necessary singularity. In addition

$$\Pi_3 = \begin{cases} \Pi_3^+ & \text{if } y_3 > 0, \\ \Pi_3^- & \text{if } y_3 < 0. \end{cases}$$

The functions  $\tilde{\Pi}_1^+$ ,  $\Pi_1^-$  and  $\Pi_3^+$  are non-singular solutions of the Helmholtz equations  $\Delta \tilde{\Pi}_1^+ + \kappa^2 \tilde{\Pi}_1^+ = 0$  in  $\mathbb{R}_+^3$ ,  $\Delta \Pi_1^- + \kappa^2 n^2 \Pi_1^- = 0$  in  $\mathbb{R}_-^3$  and so on (the solutions are smooth away from the plane  $y_3 = 0$ ). The components  $\Pi_1$  and  $\Pi_3$  are only coupled at the interface and as before boundary conditions are provided by requiring that tangential components of  $\mathbf{g}_{L,1}$  given by the analogue of (12.48) are continuous. By requiring continuity of the first component of  $\mathbf{g}_{L,1}$ , we have, at  $y_3 = 0$ ,

$$\Pi_1^+ + \frac{1}{\kappa^2} \frac{\partial}{\partial y_1} \nabla \cdot \Pi^+ = \Pi_1^- + \frac{1}{\kappa^2 \epsilon_r^e} \frac{\partial}{\partial y_1} \nabla \cdot \Pi^-,$$

and by enforcing continuity of the second component,

$$\frac{1}{\kappa^2} \frac{\partial}{\partial y_2} \nabla \cdot \Pi^+ = \frac{1}{\kappa^2 \epsilon_r^e} \frac{\partial}{\partial y_2} \nabla \cdot \Pi^-.$$

These conditions are satisfied if, at  $y_3 = 0$ , we have (12.50)

$$\Pi_1^+ = \Pi_1^-,$$

(12.51)

$$\nabla \cdot \Pi^+ = \frac{1}{\epsilon_r} \nabla \cdot \Pi^-.$$

In addition, continuity of tangential components of  $\nabla_y \times g_{l,1}$  requires

$$\begin{aligned} \frac{\partial \Pi_3^+}{\partial y_2} &= \frac{\partial \Pi_3^-}{\partial y_2} \\ \left( \frac{\partial \Pi_3^+}{\partial y_1} = \frac{\partial \Pi_1^+}{\partial y_3} \right) &= \left( \frac{\partial \Pi_3^-}{\partial y_1} = \frac{\partial \Pi_1^-}{\partial y_3} \right) \end{aligned}$$

These are satisfied if, at  $y_3 = 0$ , (12.52)

$$\Pi_3^+ = \Pi_3^-,$$
(12.53)

$$\frac{\partial \Pi_1^+}{\partial y_3} = \frac{\partial \Pi_1^-}{\partial y_3}$$

The boundary conditions (12.50)–(12.53) are such that we can solve for  $\Pi_1^\pm$  and then compute  $\Pi_3^\pm$ . Furthermore, it turns out that we can assume that  $\Pi_1^\pm$  is independent of the angular coordinate  $\theta$ .

As in our derivation of  $\Pi_3$  in the case of a vertical dipole, we can write, for  $y_3 > 0$ ,

$$\Pi_1^+ = \int_0^\infty \left( \frac{1}{4\pi} \exp(-\mu_+ |x_3 - y_3|) \frac{\lambda}{\mu_+} + a_+(\lambda) \exp(-\mu_+(x_3 + y_3)) \right) J_0(\lambda\rho) d\lambda,$$

and, for  $y_3 < 0$ ,

$$\Pi_1^- = \int_0^\infty b(\lambda) J_0(\lambda\rho) \exp(\mu_- y_3 - \mu_+ x_3) d\lambda.$$

The transmission conditions (12.50) and (12.53) are satisfied if

$$\frac{1}{4\pi} \frac{\lambda}{\mu_+} + a(\lambda) = b(\lambda) \text{ and } \frac{1}{4\pi} \lambda - \mu_+ a_+(\lambda) = \mu_- b(\lambda).$$

Solving for  $a_+(\lambda)$  and  $b(\lambda)$ , we obtain

$$\begin{aligned} \Pi_1^+ &= \Phi(x, y) - \Phi(x', y) + \frac{1}{2\pi} \int_0^\infty J_0(\lambda\rho) \frac{\exp(-\mu_+(y_3 + x_3))\lambda}{\mu_+ + \mu_-} d\lambda \text{ for } y_3 > 0, \\ \Pi_1^- &= \frac{1}{2\pi} \int J_0(\lambda\rho) \exp(\mu_- y_3 - \mu_+ x_3) \frac{\lambda}{\mu_+ + \mu_-} d\lambda \text{ for } y_3 < 0. \end{aligned}$$

Again we see that  $\Pi_1$  is a perturbation of the corresponding Hertz vector for the perfectly conducting half space.

We now wish to determine  $\Pi_3^\pm$ . The boundary condition (12.51) yields

$$\frac{\partial \Pi_3^+}{\partial y_3} - \frac{1}{\epsilon_r^\theta} \frac{\partial \Pi_3^-}{\partial y_3} = \frac{1}{\epsilon_r^\theta} \frac{\partial \Pi_1^-}{\partial y_1} - \frac{\partial \Pi_1^+}{\partial y_1} \text{ at } y_3 = 0.$$

But we are trying to solve (12.49) using cylindrical polar coordinates, so we need to express  $\partial/\partial y_i$  in that coordinate system:

$$\frac{\partial}{\partial y_1} = \frac{\partial r}{\partial y_1} \frac{\partial}{\partial r} = \cos \theta \frac{\partial}{\partial r},$$

where  $\theta$  is the angle coordinate (i.e. angle between  $e_1$  and  $y$ ). So (12.54)

$$\frac{\partial}{\partial y_3} \Pi_3^+ - \frac{1}{\epsilon_r^\theta} \frac{\partial \Pi_3^-}{\partial y_3} = \cos \theta \frac{\partial}{\partial r} \left( \frac{\Pi_1^-}{\epsilon_r^\theta} - \Pi_1^+ \right).$$

We thus cannot assume that  $\Pi_3^\pm$  has no angular dependence. Instead, we build the angular dependence by using  $J_1(\lambda Q)$   $\cos \theta \exp(-\mu y_3)$  as the basic solution of the Helmholtz equations satisfied by  $\mathbb{R}_+^3$ . Hence,

$$\begin{aligned} \Pi_3^+ &= \int_0^\infty c(\lambda) J_1(\lambda \rho) \cos \theta \exp(-\mu_+(y_3 + x_3)) d\lambda, \\ \Pi_3^- &= \int_0^\infty d(\lambda) J_1(\lambda \rho) \cos \theta \exp(\mu_- y_3 - \mu_+ x_3) d\lambda. \end{aligned}$$

Here  $c(\lambda)$ ,  $d(\lambda)$  are coefficients to be determined. Condition (12.52) yields  $c(\lambda) = d(\lambda)$  and condition (12.54) yields

$$\begin{aligned} -\mu_+ J_1'(\lambda \rho) \cos \theta c(\lambda) - \frac{1}{\epsilon_r^\theta} \mu - J_1(\lambda \rho) \cos \theta d(\lambda) \\ = \cos \theta \lambda J_0'(\lambda \rho) \left( \frac{\lambda}{2\pi(\mu_+ + \mu_-)} - \frac{1}{\epsilon_r^\theta} \frac{\lambda}{2\pi(\mu_+ + \mu_-)} \right). \end{aligned}$$

Since  $J'_0 = -J_1$  [203], we have

$$c(\lambda) = \frac{\lambda^2 (\epsilon_r^\theta - 1)}{2\pi (\epsilon_r^\theta \mu_+ + \mu_-)(\mu_+ + \mu_-)}.$$

Multiplying top and bottom by  $\mu_+ - \mu_-$  and using the definition of  $\mu_+$  and  $\mu_-$ , we have

$$d(\lambda) = c(\lambda) \frac{1}{2\pi} \frac{\lambda^2}{k^2} \frac{(\mu_+ - \mu_-)}{(\epsilon_r^\theta \mu_- + \mu_-)}.$$

This completes our determination of  $g_{l,1}$  for  $x_3 > 0$  (and by rotation  $g_{l,2}$  for  $x_3 > 0$ ). The case of  $g_{l,1}$  and  $g_{l,2}$  for  $x_3 < 0$  is left to the reader!

### 12.4.3 Reduction to a bounded domain

We now wish to use a representation theorem like that used in the previous sections to reduce the scattering problem to a problem posed on a bounded domain. We assume, as discussed in Section 1.3, that the background medium has two regions. In the upper half space  $\mathbb{R}_+^3$ , the medium is air, while the lower half space is occupied by earth, which is modeled as a uniform conducting medium. We assume that the air and earth have the same magnetic properties, so that  $\mu_r = 1$  [273].

Let the refractive index  $n$  be defined by

$$n^2 = \begin{cases} \epsilon_r & \text{if } x_3 < 0, \\ 1 & \text{if } x_3 > 0, \end{cases}$$

with  $\Im(n) \geq 0$ . Then we seek to compute a total field  $E \in H_{\text{loc}}(\text{curl}; \mathbb{R}^3)$  such that(12.55)

$$\nabla \times \nabla \times E - k^2 n^2 E = F \text{ in } \mathbb{R}^3 \setminus D.$$

Here  $F$  is a function of compact support in  $x_0 \in \mathbb{R}_+^3$ . In general, we could allow a distributed source, but we have in mind either  $F = 0$  (plane wave) or  $F = p\delta_{x_0}$  for some  $p \in \mathbb{R}^3$ ,  $p \neq 0$  and  $D \subset \mathbb{R}_+^3$  (point source). The scatterer  $E^i = G_L^T(x, x_0)p$  is a bounded, simply connected domain such that  $\mathbb{R}^3 \setminus D$  is simply connected. Extensions to more general topological settings are possible. On the surface of the scatter  $\Gamma$ , we impose the perfect conducting boundary condition(12.56)

$$\nu \times E = 0 \text{ on } \Gamma.$$

The total field  $E$  is the sum of the incident and scattered field ( $E^i$  and  $E^s$  respectively)(12.57)

$$E = E^i + E^s,$$

where  $E^i$  is a smooth solution of Maxwell's equations in a neighborhood of  $D$  and satisfies

$$\nabla \times \nabla \times E^i - k^2 n^2 E^i = F \text{ in } \mathbb{R}^3.$$

Note that the usual transmission conditions are satisfied on  $\sum_0$ . In the case  $F = 0$ , we have  $E^i$  given by (12.38) or (12.42), whereas if  $F = p\delta_{x_0}$ , we have  $G_L^T(x, x_0)p$  where  $G_L$  is the dyadic Green's function for the layered medium. Finally,  $E^s$  satisfies the integral radiation condition(12.58)

$$\lim_{x \rightarrow \infty} \int_{\partial B_R} |\nabla \times E^s \times \hat{x} - i \kappa n E^s|^2 dA = 0.$$

To reduce this problem to a variational problem posed on a bounded domain we introduce a Lipschitz smooth and connected surface  $\Sigma$  containing  $D$  in its

interior. We assume the added restrictions that  $\Sigma \subset \mathbb{R}^3_-$  and that the domain  $\Omega$  inside  $\Sigma$  and outside  $D$  is simply connected.

We now need to derive a boundary condition on  $\Sigma$ . For this we follow Cutzach and Hazard [111] and use the integral representation proposed previously in Section 12.2. Using  $\Gamma = \partial D$  as the surface for the integral representation (12.2), we obtain, for  $x \in \mathbb{R}^3 \setminus \bar{D}$ ,

$$\begin{aligned} E^s(x) &= \int_{\Gamma} \left( \mathbb{G}_L^T(x, y) (\nu \times \nabla \times E^s)(y) \right. \\ &\quad \left. + (\nabla_y \times \mathbb{G}_L)^T(x, y) (\nu \times E^s)(y) \right) dA(y). \end{aligned}$$

Now, since both the columns of  $\mathbb{G}_L$  and the function  $E$  are smooth solutions of the homogeneous Maxwell system in  $D$  we have

$$\int_{\Gamma} \mathbb{G}_L^T(x, y) (\nu \times \nabla \times E^i)(y) + (\nabla_y \times \mathbb{G}_L)^T(x, y) (\nu \times E^i)(y) dA(y) = 0$$

and so

$$\begin{aligned} E &= E^i + E^s \\ &= E^i + \int_{\Gamma} \left( \mathbb{G}_L^T(x, y) (\nu \times \nabla \times E)(y) \right. \\ &\quad \left. + (\nabla_y \times \mathbb{G}_L)^T(x, y) (\nu \times E)(y) \right) dA(y). \end{aligned}$$

But, using the perfectly conducting boundary condition, the last term in the integral vanishes, so we have (12.59)

$$E = E^i + \int_{\Gamma} \mathbb{G}_L^T(x, y) (\nu \times \nabla \times E)(y) dA(y).$$

This representation is not sufficient for deriving a finite element method, since we need to allow more general fields  $E$  in  $H(\text{curl}; \Omega)$ . Hence, as in Section 12.2, we extend the domain of the integral operator in (12.59). To this end let  $\chi \in C_0^\infty(\mathbb{R}^3)$  be such that  $\chi = 0$  in a neighborhood of  $\Sigma$  and outside  $\Sigma$  and suppose  $\chi = 1$  on  $\Gamma$ . Then

$$\begin{aligned} &\int_{\Gamma} \mathbb{G}_L^T(x, y) (\nu \times \nabla \times E)(y) dA(y) \\ &= \int_{\Gamma} \chi(y) \mathbb{G}_L^T(x, y) (\nu \times \nabla \times E)(y) dA(y) \\ &= - \int_{\Omega} \chi(y) \mathbb{G}_L^T(x, y) (\nabla \times \nabla \times E)(y) \\ &\quad + \left( \nabla_y \times (\chi(y) \mathbb{G}_L)^T(x, y) \nabla \times E(y) \right) dA(y). \end{aligned}$$

Here we have used the integration-by-parts formula (3.51) and have taken into account that  $\nu$  points out of  $D$ . As in Section 12.2, let  $\mathbb{G}_L(x, y) = \chi(y) \mathbb{G}_L(x, y)$

where  $X \in C_0^\infty(\mathbb{R}^3)$  is such that  $X = 1$  in a neighborhood of  $\Gamma$  and  $X = 0$  in a neighborhood of  $\Sigma$ . Then using the fact that  $E$  satisfies the Maxwell system (12.55) with  $F = 0$  in  $\Omega$ , we have

$$E(x) = E^i(x) + \int_{\Omega} \left( \nabla_y \times \tilde{\mathbb{G}}_L^T(x, y) \nabla \times E(y) - \kappa^2 \epsilon_r \tilde{\mathbb{G}}_L^T(x, y) E(y) \right) dV(y).$$

As before, we define, for  $u \in H(\text{curl}; \Omega)$  and  $x$  outside the support of  $X$ , (12.60)

$$\begin{aligned} \mathcal{G}(u)(x) &= \int_{\Omega} \left( \nabla_y \times \tilde{\mathbb{G}}_L^T(x, y) \nabla \times u(y) \right. \\ &\quad \left. - \kappa^2 \epsilon_r \tilde{\mathbb{G}}_L^T(x, y) u(y) \right) dV(y). \end{aligned}$$

We have proved the following lemma.

**Lemma 12.18** *Let  $E \in H_{\text{loc}}(\text{curl}; \mathbb{R}^3 \setminus D)$  satisfy (12.55)-(12.58). Then, for  $x$  outside the support of  $X$ , we have* (12.61)

$$E(x) = E^i(x) + \mathcal{G}(E)(x).$$

In order to apply the arguments developed in Section 12.2, we need to extend the operator  $\mathcal{J}$  to functions in  $(L^2(\Omega))^3$ . We note that  $\tilde{\mathbb{G}}_L(x, y)$  is smooth when  $x$  is not in the support of  $X$ . Hence, using integration by parts, for  $u \in H(\text{curl}; \Omega)$  such that  $v \times u = 0$  on  $\Gamma$ , we obtain the regularized operator  $\mathcal{J}^R$  defined by

$$\mathcal{G}^R(u) = \int_{\Omega} \left( \nabla_y \times \nabla_y \times \tilde{\mathbb{G}}_L(x, y) - \kappa^2 n^2 \tilde{\mathbb{G}}_L(x, y) \right)^T u(y) dV(y).$$

Using this identity we see that  $\mathcal{J}^R(u)$  is well defined and continuous for  $u \in (L^2(\Omega))^3$  and agrees with  $\mathcal{J}(u)$  defined in (12.60) when  $u \in H(\text{curl}; \Omega)$ .

Now using the space  $X$  defined in (4.3), and using the representation for  $E$  given by (12.61), we see that we need to compute  $E \in X$  such that (12.62)

$$\begin{aligned} \nabla \times \nabla \times E - \kappa^2 n^2 E &= F \text{ in } \Omega, \\ (12.63) \end{aligned}$$

$$\begin{aligned} v \times E &= 0 \text{ on } \Gamma, \\ (12.64) \end{aligned}$$

$$\begin{aligned} (\nabla \times E \times v - i\kappa n E_T) &= \left( \nabla \times E^i \times v - i\kappa n E_T^i \right) \\ &\quad + \left( \nabla \times \mathcal{G}^R(E) \times v - i\kappa n \mathcal{G}^R(E)_T \right) \text{ on } \Sigma, \end{aligned}$$

where  $F = p\delta x_0$  in the case of a point source (in the upper half space) and  $F = 0$  for a plane wave. Here we have imposed the impedance boundary condition on  $\Sigma$ . Thus, the  $\mathcal{J}^R$  operator gives a perturbation of the absorbing boundary condition considered in Section 13.5.

Using the standard Galerkin strategy of multiplying (12.62) by the complex conjugate of  $\varphi \in X$  and integrating by parts, then using (12.63) and (12.64), we obtain the problem of finding  $E \in X$  such that(12.65)

$$\begin{aligned} (\nabla \times E, \nabla \times \varphi) &= -\kappa^2(n^2 E, \varphi) - i\kappa \langle E_T, \varphi_T \rangle - \left\langle T_n(\mathcal{F}^R(E)), \varphi_T \right\rangle \\ &= \left\langle T_n(E^i), \varphi_T \right\rangle \text{ for all } \varphi \in X, \end{aligned}$$

where the impedance operator  $T_n$  is given by

$$T_n(u) = (\nabla \times u) \times v - i\kappa n u_T \text{ on } \Sigma.$$

This is essentially the operator  $T$  from (12.10) but allowing for the fact that  $\epsilon_r \neq 1$ . Note that, if  $F = p\delta_{x_0}$ , then  $E^i(x) = G_L(x_0, x)p$ , and if  $E^i$  is a plane wave as constructed in Section 12.4.1 then  $F = 0$ . In either case  $E^i$  satisfies the background layered medium Maxwell system.

We could now proceed to analyze this variational problem as we did for the simpler problem in Section 12.2. Using asymptotic methods, Cutzach and Hazard [111] show that any solution  $E$  of (12.65), extended to  $\mathbb{R}^3 \setminus \Omega$  by (12.61), satisfies the radiation condition (12.58). Hence, a solution of (12.65) is a weak solution of the scattering problem. As in Section 12.2, eqn (12.65) can then be expressed as a Fredholm equation on  $X_0$ . The uniqueness result of [111] then allows us to conclude the existence of a solution to (12.65). The finite element error analysis can be applied to the discretization of (12.65) obtained by replacing  $X$  by  $X_h$  (see (7.1)), and we can conclude that Theorem 12.9 holds provided  $\Omega$  is covered by a regular mesh of elements that are quasi-uniform on  $\Sigma$ .

# 13 ALGORITHMIC DEVELOPMENT

## 13.1 Introduction

The intention of this rather ambitious chapter is to discuss some issues related to practical aspects of solving Maxwell's equations. We start with a very important problem: how to solve the linear system resulting from an edge finite element discretization of Maxwell's equations (see Section 13.2). This is a major problem since the matrix for this linear system is complex and sparse (and often symmetric) but not Hermitian or definite. Hence many of the standard approaches are not applicable. We discuss the use of an overlapping Schwarz algorithm. Multigrid methods have also been tried [118]. Nevertheless, much remains to be done to arrive at a fast solver.

After discussing the solution of the linear system, we consider the problem of the wavenumber dependence of error estimates. This forces us to confront the problem of “phase error” which is, perhaps, the dominant cause of error in the computed solution for coarse grids. In particular, the wavelength of the wave in a numerical simulation will not be precisely correct. Thus, the numerical wave will become out of phase with the true solution as the wave transits a region (see Section 13.3).

Related to phase error is the difficulty of assessing the error in a numerical solution. There are many schemes for providing an *a posteriori* error estimate for finite element methods for parabolic and uniformly elliptic problems. These tend not to work so well for time-harmonic problems on coarse grids, since constants in the *a posteriori* estimates depend on the wavenumber  $\omega$ . In addition, local error indicators do not necessarily show where to refine a mesh, since phase error can build up slowly across a domain. In Section 13.4 we derive a residual-based error estimator and discuss the problem further.

In Section 13.5 we return to the problem of how to approximate the solution of an exterior scattering problem via a boundary value problem posed on a bounded domain. In Chapters 10 - 12 this was done by using elaborate schemes for approximating the Calderon operator. Here we consider three other approaches. In particular, we examine further the standard absorbing boundary condition, a less standard infinite element approach and finally the justly celebrated Bérenger perfectly matched layer (PML).

Lastly, in Section 13.6 we describe a special post-processing issue related to Maxwell's equations. Often the desired output from a Maxwell solver is an estimate of the far field pattern of the scattered wave. We show how to apply flux recovery techniques to extract a high order approximation to the far field

pattern.

Necessarily a great deal of important work has been neglected in this section. I have tried to put the issues examined here into perspective, but there is much less analysis here, and much that remains to be done.

## 13.2 Solution of the linear system

We now want to discuss how to solve the linear system derived from the finite element approximation of Maxwell's equations. Unlike the situation for uniformly elliptic problems where methods such as multigrid [282] are available, the problem of solving our linear system is less well understood.

For simplicity we shall not discuss the full scattering problem here. The principal difficulty is visible on a simple model problem. Thus, let us consider the perfect conducting cavity problem of finding  $E$  on a bounded domain  $\Omega$  such that

$$\begin{aligned} \nabla \times \nabla \times E - \kappa^2 E &= F \quad \text{in } \Omega, \\ v \times E &= 0 \quad \Gamma = \partial\Omega. \end{aligned}$$

As usual, this is cast into variational form as the problem of finding  $E \in H_0(\text{curl}; \Omega)$  such that

$$(\nabla \times E, \nabla \times \varphi) - \kappa^2(E, \varphi) = (F, \varphi) \quad \text{for all } \varphi \in H_0(\text{curl}; \Omega).$$

For this section we have no impedance boundary condition, and  $\varepsilon_r = \mu_r = 1$ . Hence the general sesquilinear form  $a$  defined in (4.5) reduces to  $a(u, v) = (\nabla \times u, \nabla \times v) - \kappa^2(u, v)$ . In Chapter 4 we presented a detailed existence and uniqueness study of this problem. In this chapter we assume that  $\kappa$  is not an interior Maxwell eigenvalue and  $\Omega$  is a simply connected Lipschitz polyhedron with connected boundary  $\Gamma = \partial\Omega$ .

As we have seen in Chapter 7 the problem of finding a finite element approximation is developed as follows. We suppose that we have a family of finite element meshes  $\tau_b$ ,  $b > 0$ , of regular geometric elements. For concreteness we assume that they are regular tetrahedra, but hexahedra with edges parallel to the coordinate axis are also permissible. Let  $X_b \subset H_0(\text{curl}; \Omega)$  be the degree- $k$  edge space on these elements given by eqn (7.1). We then want to find  $E_b \in X_b$  such that(13.1)

$$(\nabla \times E_b, \nabla \times \varphi_b) - \kappa^2(E_b, \varphi_b) = (F, \varphi_b) \quad \text{for all } \varphi_b \in X_b.$$

From Chapter 7 we know that this problem is well posed provided  $b$  is small enough, and  $E_b \rightarrow E$  as  $b \rightarrow 0$  in  $H(\text{curl}; \Omega)$ .

Now let us expand  $E_b$  using the edge finite element space basis functions for  $X_b$  defined via the degrees of freedom for the element (having first introduced a numbering of those degrees of freedom). Then we can write

(13.2)

$$E_h(x) = \sum_{l=1}^{N_h} E_l \chi_l(x),$$

where  $N_b$  is the number of degrees of freedom for functions in  $X_b$ , and  $\vec{E}_h^{(n)}$  are the basis functions. For example when  $k = 1$ , the degrees of freedom are associated with the edges in the mesh and hence  $N_b$  is given by the number of interior edges in the mesh. Using (13.2) in (13.1), we see that the problem of solving (13.1) is equivalent to solving the matrix problem (13.3)

$$A_h \vec{E}_h = \vec{F}_h,$$

where  $A_b$  is the  $N_b \times N_b$  matrix with

$$(A_h)_{l,m} = (\nabla \times \chi_m, \nabla \times \chi_l) - \kappa^2(\chi_m, \chi_l), \quad 1 \leq l, m \leq N_h.$$

The vector  $\vec{F}_h$  has  $l$ th entry  $F_l = (F, \chi_l)$ ,  $1 \leq l \leq N_b$ , and

$$\vec{E}_h = (E_1, E_2, \dots, E_{N_h})^\top.$$

From Corollary 7.3, we know that  $A_b$  is invertible if  $b$  is small enough but although, in this case,  $A_b$  is symmetric it is not positive definitive due to the presence of the term  $-\kappa^2(\chi_m, \chi_l)$  in its definition. Usually, for scattering problems,  $A_b$  is sparse, complex, but neither symmetric nor Hermitian. For the Helmholtz equation, the corresponding bilinear form is positive definite for  $\kappa$  sufficiently small and only loses definiteness as  $\kappa$  increases beyond the first Dirichlet eigenvalue of the domain. For Maxwell's equations, due to the fact that the curl of the gradient of a function vanishes, the bilinear form is indefinite for any  $\kappa > 0$ , no matter how small.

One approach is to solve (13.3) using, for example, sparse LU factorization [146]. However, as  $N_b$  grows, the amount of work and memory (usually memory is the deciding factor) needed for the factorization renders this approach infeasible. We are thus motivated to consider using an iterative solver. We could hope to use the simple Richardson scheme of generating a sequence of vectors  $\vec{E}_h^{(n)}$ ,  $n = 0, 1, 2, \dots$ , from an initial guess  $\vec{E}_h^{(0)}$  using the iteration

$$\vec{E}_h^{(n+1)} = \vec{E}_h^{(n)} + \alpha_n \left( \vec{F}_h - A_h \vec{E}_h^{(n)} \right),$$

where  $\alpha_n > 0$  is a parameter to be chosen. The iteration matrix for this procedure is  $M = I - \alpha_n A_b$ , and since  $A_b$  is indefinite there is no reason to suppose it is possible to choose  $\alpha_n$  so that the spectral radius of  $I - \alpha_n A_b$ , denoted by  $\varrho(I - \alpha_n A_b)$ , satisfies the necessary condition for convergence that  $\varrho(I - \alpha_n A_b) < 1$ . Even if such a choice is possible, we need to choose  $\alpha_n = O(b^2)$ , since the largest eigenvalues of  $A_b$  are  $O(1/b)$ , as is suggested by Gershgorin's theorem. Thus, convergence will be extremely slow.

One way around this is to construct a matrix  $B_h$ , called a preconditioner, and use the iteration(13.4)

$$\vec{E}_h^{(n+1)} = \vec{E}_h^{(n)} + B_h \left( \vec{F}_h - A_h \vec{E}_h^{(n)} \right).$$

We want to choose  $B_h$  such that  $B_h A_h \simeq I$  or, more precisely, such that  $\rho(I - B_h A_h) < 1$  uniformly in  $h$ , so that the iterative method converges at roughly the same rate regardless of  $h$ . In order to be useful,  $B_h$  must also be easy to compute so that each step of the iteration is relatively efficient. For example, it would be desirable to have the number of floating point operations expended to compute  $A_h \vec{y}$  be roughly the same as the number needed to compute  $B_h \vec{y}$  for general vectors  $\vec{x}$  and  $\vec{y}$ .

One easy method to construct  $B_h$  is the incomplete LU (ILU) decomposition. It is necessary to allow for considerable fill in for this approach to work (in particular, the level-zero ILU decomposition is often singular), but such a preconditioner can be effective. Of course, for large problems, this approach also becomes prohibitively expensive. For standard uniformly elliptic problems like Laplace's equation, typical methods for constructing  $B_h$  include the multigrid method and the Schwarz methods.

Multigrid methods have been derived and analyzed for finite element subspaces in  $H(\text{div};\Omega)$  and  $H(\text{curl};\Omega)$ , by Hiptmair [161] and Arnold *et al.* [18]. The analysis of these methods, which uses tools like those found in Chapter 7, shows that for the coercive problem of finding  $E \in H_0(\text{curl};\Omega)$  such that(13.5)

$$\nabla \times \nabla \times E + \gamma E = F, \quad \gamma > 0,$$

these multigrid methods can be very effective. There is also ample computational evidence to this effect [262]. We should point out that, although (13.5) does not arise in time-harmonic scattering theory, it arises as part of an implicit method for solving the time domain Maxwell equations.

For the non-coercive Helmholtz and Maxwell equations, a new problem appears. Due to the indefiniteness of the problem there is a coarse level mesh below which the discrete equations lose the necessary properties for convergence of the iterative scheme (we know they also lose accuracy if the number of grid points per wavelength drops too low — see Section 13.3). Thus, the multigrid mesh coarsening strategy must stop at a grid that is still fine enough to resolve the wave and this has an adverse impact on multigrid efficiency. Non-standard methods have been suggested for avoiding this problem (see [204] for an interesting suggestion applied to the Helmholtz equation).

The same problem also afflicts the Schwarz methods for Maxwell's equations [68, 280, 281, 147]. In these methods, a global coarse grid (but not too coarse) is needed to obtain a convergent iteration. Here we choose to describe the overlapping Schwarz method rather than the multigrid method, because one of the few general codes to use either technique uses a Schwarz procedure [256]. In fact the one we shall present here is a little simpler than the one used in that

code. In particular, we present the method of Gopalkrishnan and Pasciak [147] and follow directly their proofs with the exception that we check that they are applicable for a general Lipschitz polyhedron (rather than a convex polyhedron assumed in [147]).<sup>3</sup>

The overlapping Schwarz algorithm is based on two levels of partitioning of  $\Omega$  (at least in the variant proposed in [147]). The first level is based on a coarse mesh  $\tau_H$  of regular elements of maximum diameter  $H$  with elements denoted  $K_1, \dots, K_{M_H}$ . Next, each coarse-level tetrahedron  $K_m$  is partitioned into fine level tetrahedra  $\kappa_l^m, l = 1, \dots, L_m$ , where  $\kappa_l^m \subset K_m, 1 \leq l \leq L_m$ . This is obviously a fairly practical setup in which a coarse grid is generated first, and is then refined to try to obtain a more accurate solution.

We denote by  $X_H$  the subspace of  $H_0(\text{curl}; \Omega)$  constructed using edge elements of degree  $k$  on  $\tau_H$ . To form the overlapping grid, each  $K_i$  is enlarged to form  $\kappa'_i$  by adding fine-level tetrahedra in such a way that  $\partial\kappa'_i$  is a union of faces of the level  $b$  mesh. Obviously, each  $\kappa'_i$  is covered by a mesh of level  $b$  tetrahedra, and if  $X_b$  denotes the global edge element space of elements of degree  $k$  on the  $b$  level grid we may define

$$X_{h,l} = \left\{ u_h \in X_h \mid u_h(x) = 0 \text{ for } x \notin \kappa'_l \right\}, \quad 1 \leq l \leq M_H.$$

We can think of this edge finite element subspace as the intersection of  $X_b$  with  $H_0(\text{curl}; \kappa'_i)$  where functions in  $H_0(\text{curl}; \kappa'_i)$  are extended to  $\Omega$  by zero. It is also convenient to introduce the corresponding space of scalar

$$S_{h,l} = \left\{ p_h \in S_h \mid p_h(x) = 0 \text{ for all } x \notin \kappa'_l \right\}$$

so that roughly  $S_{h,l} = H_0^1(\kappa'_l) \cap S_h$ , where  $S_b$  is the usual scalar space associated with  $X_b$ .

Of course, this overlapping grid has to obey some rules, and we make the following assumptions [147] :

- (1) *Generous overlap* There exists  $\delta > 0$  such that  $\text{dist}(\partial\kappa'_l \cap \Omega, \partial\kappa_i \cap \Omega) \geq \delta H$  for  $l = 1, 2, \dots, M_H$ .
- (2) *Finite covering* Every point of  $\Omega$  belongs to at most  $\varrho$  subdomains  $\kappa'_i$  independent of  $b$  and  $H$ .
- (3) *H-independent uniformity* There are a fixed number of Lipschitz polyhedral reference domains  $\{\mathcal{K}_m\}$  such that each subdomain  $K'_i$  is the image under an affine transformation  $F_{m,l} : \mathcal{K}_m \rightarrow K'_i$  of the form  $F_{m,l}(\hat{x}) = B_{m,l}\hat{x} + b_{m,l}$  where  $B_{m,l}$  is an invertible matrix and  $b_{m,l} \in \mathbb{R}^3$ . The transformations are assumed to be such that there are constants  $c_0$  and  $c_1$  independent of  $H$  such that

$$c_0H \leq |B_{m,l}| \leq c_1H, \quad (\text{where } |\cdot| \equiv \text{matrix spectral norm}),$$

$$c_0H \leq \left| \det(B_{m,l}) \right|^{1/3} \leq c_1H.$$

<sup>3</sup> While proof reading this text, I found that a theory of Schwarz methods for general polyhedra has also been developed by Pasciak and Zhao. Their report “Overlapping Schwarz methods in  $H(\text{curl})$  on nonconvex domains” is available at <http://www.math.tamu.edu/~joe.pasciak/>.

**Remark 13.1** It appears that the first two assumptions are rather standard in the Schwarz literature [147, 280]. The third assumption is a generalization of the one in [147], although it has to be admitted that the generalization is not particularly useful, and that it is difficult to satisfy this assumption (or the one in [147]) except for very special uniform meshes.

Now we define the two-level multiplicative Schwarz preconditioner. During the course of the iterative algorithm to compute  $\vec{E}_h^{(n+1)}$  from  $\vec{E}_h^{(n)}$  via (13.4), we first compute the residual vector on the fine grid

$$\vec{r}_h^{(n)} = \vec{F}_h - A_h \vec{E}_h^{(n)}.$$

In the two-level multiplicative Schwarz method, we apply the preconditioner  $B_b$  to  $\vec{r}_h^{(n)}$  in the following way. The vector  $\vec{r}_h^{(n)}$  corresponds, using the basis functions, to a function  $r_h^{(n)} \in X_h$ . We then solve the discrete Maxwell's equations on the coarse grid to find  $v_H \in X_H$  such that

$$a(v_H, \varphi_H) = (r_h^{(n)}, \varphi_H) \text{ for all } \varphi_H \in X_H.$$

Of course, this requires inverting the coarse grid matrix  $A_H$  (or, more precisely, solving a linear system involving  $A_H$ ) and this in itself is expensive. We could do this, for example, using an incomplete LU preconditioned Richardson procedure on the coarse grid space  $X_H$ .

Now, for each  $l = 1, \dots, M_H$  we solve the local fine grid problem of finding  $v_{b,l} \in X_{b,l}$  such that

$$a(v_{h,l}, \varphi_{h,l}) = (r_h^{(n)}, \varphi_{h,l}) - a(v_H, \varphi_{h,l}) \text{ for all } \varphi_{h,l} \in X_{h,l}.$$

This corresponds to solving  $M_H$  standard discrete Maxwell problems on each of the overlapping subdomains  $K_i$  using the fine level grid. If these subdomains are small (within the generous overlap assumption), each of these problems should be much more rapid to solve than the global problem.

Let  $\vec{v}_{h,l}$  and  $\vec{V}_H$  be the coefficients of  $v_{b,l}$  and  $v_H$  in the usual basis function expansion, then we define

$$B_h \vec{r}_h^{(n)} = \vec{V}_H + \alpha \sum_{l=1}^{M_H} \vec{V}_{h,l}.$$

As we can see from this discussion, it is easier to think of this algorithm in terms of functions and operators. In this view the residual  $\vec{r}_h^{(n)}$  is thought of as a functional in the dual space  $K'_i$ . More precisely, for  $\varphi \in X_b$ , we define  $R_h^{(n)}: X_h \rightarrow \mathbb{R}$  by

$$R_h^{(n)}(\varphi) = (r_h^{(n)}, \varphi).$$

Since  $R_h^{(n)}$  is obviously bounded and linear,  $R_h^{(n)} \in X_h'$ . Then the preconditioner  $B_b$  is a map from  $X_b' \rightarrow X_b'$ . To compute  $B_b(R_b)$  for some  $R_b \in X_b'$ , we carry out the following three steps (equivalent to the previously presented algorithm):

1. Solve for  $v_H \in X_H$ ,

$$\alpha(v_H, \varphi_H) = R_h(\varphi_H) \text{ for all } \varphi_H \in X_H.$$

2. For  $l = 1, \dots, M_H$  define  $v_{h,l} \in X_{h,l}$  by

$$a(v_{h,l}, \varphi_{h,l}) = R_h(\varphi_{h,l}) - a(v_H, \varphi_{h,l}) \text{ for all } \varphi_{h,l} \in X_{h,l}.$$

3. Set  $B_h(R_h) = v_H + \alpha \sum_{i=1}^{M_H} v_{h,i}$ .

We shall prove the following theorem from [147] (but now extended to hold on Lipschitz polyhedral domains), which shows that  $B_h$  is a good preconditioner.

**Theorem 13.2** There exists  $a_0 > 0$  such that for all  $a \leq a_0$ , there is a constant  $H_2 > 0$ ,  $H_2 = H_2(a)$ , such that if  $H \leq H_2$  then

$$\| (I - B_h A_h) u_h \|_{H(\text{curl}; \Omega)} \leq \gamma \| u_h \|_{H(\text{curl}; \Omega)}$$

for all  $u_h \in X_h$  with  $\gamma < 1$  independent of  $h$  and  $H$ . Here  $A_h : X_h \rightarrow X_h$  is the operator such that  $(A_h \varphi_h)(\xi_h) = a(\varphi_h, \xi_h)$  for all  $\varphi_h \in X_h$ , and  $\xi_h \in X_h$  (i.e. the operator corresponding to the matrix  $A_h$  introduced previously).

To prove this theorem, we shall first prove a number of lemmas. Following [147] we define the operators  $T_H : H(\text{curl}; \Omega) \rightarrow X_H$  and  $T_{h,l} : H(\text{curl}; \Omega) \rightarrow X_{h,l}$  such that  $T_H w \in X_H$  satisfies(13.6)

$$a(T_H \omega, \varphi_H) = a(\omega, \varphi_H) \text{ for all } \varphi_H \in X_H$$

and that  $T_{h,l} w \in X_{h,l}$  satisfies(13.7)

$$a(T_{h,l} \omega, \varphi_{h,l}) = a(\omega, \varphi_{h,l}) \text{ for all } \varphi_{h,l} \in X_{h,l}$$

for  $l = 1, 2, \dots, M_H$ . From our study in Chapter 7, Theorem 7.1, we know that since  $\kappa$  is not an interior Maxwell eigenvalue for  $\Omega$ , (13.6) can be uniquely solved provided  $H$  is small enough. We shall show later that  $T_{h,l}$ ,  $l = 1, \dots, M_H$ , are also well defined for  $H$  small enough and derive a stability bound.

Assuming for the moment that these operators are well defined, we have the following lemma.

**Lemma 13.3** Using the operators defined in (13.6) and (13.7), we have

$$(I - B_h A_h) = (I - \alpha \bar{T}_h)(I - T_H),$$

where

$$\bar{T}_h = \sum_{l=1}^{M_H} T_{h,l}.$$

**Proof** Take  $u_b \in X_b$  and let  $R_b(\varphi) = a(u_b, \varphi)$ , for all  $\varphi \in X$ . We shall compute  $(I - B_b^T A_b)u_b$ . In step (2) of the algorithm,  $v_{b,l} \in X_{b,l}$  satisfies

$$a(v_{b,l}, \varphi_{b,l}) = a(u_b, \varphi_{b,l}) - a(v_H, \varphi_{b,l}) \text{ for all } \varphi_{b,l} \in X_{b,l},$$

so  $v_{b,l} = T_{b,l}(u_b - v_H)$ . Hence,  $v_{b,l} \in X_H$ . But  $v_H \in X_H$  satisfies  $a(v_H, \varphi_H) = a(u_b, \varphi_H)$  for all  $\varphi_H \in X_H$  so  $v_H = T_H u_b$  and we are done.  $\square$

One of the main components of the analysis in [147] is to prove that the discrete Friedrichs inequality holds on each subdomain  $K_l$  and estimate the constant. We use the same approach but allow more general transformations; hence our proof uses different estimates on the local domain.

**Lemma 13.4** For  $l = 1, \dots, M_H$  there is a constant  $C$  independent of  $h$  and  $H$  such that, provided  $H$  is small enough, (13.8)

$$\|v_{h,l}\|_{\left(L^2(K_l)\right)^3} \leq CH \|\nabla \times v_{h,l}\|_{\left(L^2(K_l)\right)^3}, \quad 1 \leq l \leq M_H$$

for all  $v_{h,l} \in X_{0,b,l}$ , where  $X_{0,b,l}$  is the space of discrete divergence-free functions in  $X_{b,l}$ . More precisely,

$$X_{0,b,l} = \left\{ u_{h,l} \in X_{h,l} \mid \int_{K_l} u_{h,l} \cdot \nabla \xi_{h,l} dV = 0 \text{ for all } \xi_{h,l} \in S_{h,l} \right\}.$$

**Proof** As in [147] this is proved by mapping, but we adopt a slightly different approach to the one in that paper. By the  $H$ -independent uniformity assumption, there is a reference domain  $K_m$  and an affine map  $F_{m,l} : K_m \rightarrow K_l$ . Scalar functions in  $\xi \in H_0^1(K_l)$  are identified as usual with scalar in  $H_0^1(K_m)$  by  $\tilde{\xi} = \xi \circ F_{m,l}$  and according to (5.16) we have  $\tilde{\nabla} \tilde{\xi} = B_{m,l}^T \nabla \xi$ . Here  $\tilde{\nabla}$  denotes the gradient on  $K_m$  and  $F_{m,l}(\tilde{x}) = B_{m,l} \tilde{x} + b_{m,l}$ . Similarly, a vector function  $u_0 \in H_0(\text{curl}; K_l)$  is identified with  $\hat{u} \in H_0(\text{curl}; K_m)$  by (13.9)

$$u = (B_{m,l})^\top u \circ F_{m,l},$$

and, via Corollary 3.58, we have (13.10)

$$\hat{\nabla} \times \hat{u} = \det(B_{m,l})(B_{m,l})^{-1} \nabla \times u.$$

Now suppose  $v_{b,l} \in X_{0,b,l}$  is discrete divergence free and is mapped to  $\hat{v}$  on  $K_m$ . For all  $\xi_h \in H_0^1(K_l) \cap S_h$  (mapped to  $\tilde{\xi}$  on  $K_m$ ), we have

$$0 = \int_{K_l} v_{b,l} \cdot \nabla \xi_h dV = \int_{K_m} \left( B_{m,l}^{-\top} \hat{v} \right) \cdot \left( B_{m,l}^{-\top} \hat{\nabla} \tilde{\xi} \right) \left| \det(B_{m,l}) \right| dV.$$

Thus,  $\hat{u}$  is discrete divergence free on  $K_m$  but with the weight matrix

$$\hat{\epsilon} = \left| \det(B_{m,l}) \right| B_{m,l}^{-1} B_{m,l}^{-\top}.$$

This matrix is symmetric and positive definite so, provided  $b$  is small enough, Corollary 7.22 shows that

$$\int_{K_m} \left( \hat{\epsilon} \hat{v} \right) \cdot \hat{v} dV \leq C_{\rho}(\hat{\epsilon}) \| \nabla \times \hat{v} \|^2_{\left( L^2(K_m) \right)^3},$$

where  $\rho(\hat{\epsilon})$  is the spectral radius of  $\hat{\epsilon}$ . Using (13.9) and (13.10) to map back to the domain  $K$ , we obtain

$$\| v_{h,l} \|^2_{\left( L^2(K_l) \right)^3} \leq C_{\rho}(\hat{\epsilon}) \left| \det(B_{m,l}) \right| \| B_{m,l} \|^{-2} \| \nabla \times v_{h,l} \|^2_{\left( L^2(K_l) \right)^3}.$$

From the bounds in the definition of  $H$ -independent uniformity, we have

$$\rho(\hat{\epsilon}) \left| \det(B_{m,l}) \right| \| B_{m,l} \|^{-2} \leq CH^2$$

and hence the lemma is proved.  $\square$

Using the previous lemma in the same way as in [147], we can then prove the following *a priori* estimate for  $T_{h,l}$ .

**Lemma 13.5** *For all  $H$  sufficiently small, the operators  $T_{h,l}$ ,  $l = 1, \dots, M_h$ , are well defined and*

$$\| T_{h,l} u_h \|_{H(\text{curl}; K_l)} \leq C \| u_h \|_{H(\text{curl}; K_l)}$$

for all  $u_h \in X_h$  and  $C$  independent of  $h$ ,  $H$ ,  $l$  and  $u_h$ ,

**Remark 13.6** *This lemma is not surprising. If  $H$  is small enough, we expect that  $\kappa$  will become smaller than the smallest positive Maxwell eigenvalue of  $K$  for each  $l$ . In this case the Maxwell operator is coercive provided gradients are factored out of the function space.*

**Proof of Lemma 13.5** Using the discrete Helmholtz decomposition (7.8), we have (13.11)

$$T_{h,l} u_h = \omega_{h,l} + \nabla p_{h,l},$$

where  $w_{h,l} \in X_{0,h,l}$  and  $p_{h,l} \in S_{h,l}$ . By multiplying the above equation by  $\nabla p_{h,l}$  and integrating over  $\Omega$ , we see that  $p_{h,l}$  satisfies  $(\nabla p_{h,l}, \nabla p_{h,l}) = (u_h, \nabla p_{h,l})$ . Hence, (13.12)

$$\| \nabla p_{h,l} \|_{\left( L^2(K_l) \right)^3} \leq \| u_h \|_{\left( L^2(K_l) \right)^3}.$$

To estimate the other function in (13.11), we choose  $\varphi_{h,l} = w_{h,l}$  in (13.7) so that

$$(\nabla \times \omega_{h,l}, \nabla \times \omega_{h,l}) - \kappa^2 (\omega_{h,l}, \omega_{h,l}) = (\nabla \times u_h, \nabla \times \omega_{h,l}) - \kappa^2 (u_h, \omega_{h,l}).$$

Using the previous lemma, we obtain

$$\left(1 - CH^2\right) \|\nabla \times \omega_{h,l}\|_{L^2(\kappa_l)}^3 \leq C_1 \|u_h\|_{H(\text{curl}; \kappa_l)} \|\omega_{h,l}\|_{H(\text{curl}; \kappa_l)}. \quad (13.13)$$

Choosing  $H$  small enough shows that (13.14)

$$\|\nabla \times \omega_{h,l}\|_{L^2(\kappa_l)}^2 \leq C \|u_h\|_{H(\text{curl}; \kappa_l)} \|\omega_{h,l}\|_{H(\text{curl}; \kappa_l)}$$

and using (13.8) we obtain (13.15)

$$\|\omega_{h,l}\|_{L^2(\kappa_l)}^3 \leq C \|\nabla \times \omega_{h,l}\|_{L^2(\kappa_l)}^3.$$

Putting together (13.13)–(13.15) proves that  $\|w_{h,l}\|_{H(\text{curl}; \kappa_l)} \leq C \|u_h\|_{H(\text{curl}; \kappa_l)}$ . Note that  $\|T_{h,l} U_h\|_{H(\text{curl}; \Omega)} \leq C(\|W_{h,l}\|_{H(\text{curl}; \Omega)} + \|\nabla p_{h,l}\|_{L^2(\Omega)})$ . Using (13.12) and the above estimate for the norm of  $w_{h,l}$  completes the proof.  $\square$

**Lemma 13.7** *For  $H$  sufficiently small, we have (13.16)*

$$\begin{aligned} (u_h - T_H u_h, v_H) &\leq CH^{1/2+\delta} \|u_h - T_H u_h\|_{H(\text{curl}; \kappa_l)} \\ &\quad \times \|v_H\|_{H(\text{curl}; \kappa_l)}, \\ (u_h - T_{h,l} u_h, v_{h,l}) &\leq CH \|u_h - T_H u_h\|_{H(\text{curl}; \kappa_l)} \|v_{h,l}\|_{H(\text{curl}; \kappa_l)} \end{aligned} \quad (13.17)$$

for all  $v_H \in X_H$  and  $v_{h,l} \in X_{h,l}$ ,  $1 \leq l \leq M_H$ . Here  $\delta$  is a constant depending on the domain with  $1/2 \geq \delta > 0$  (see Lemma 7.7).

**Proof** Estimate (13.16) is the result of Lemma 7.7. Estimate (13.17) follows from Lemma 13.4 in the following way. We use the discrete Helmholtz decomposition (7.8) to write  $v_{h,l} = w_{h,l} + \nabla p_{h,l}$  for some  $w_{h,l} \in X_{0,h,l}$  and  $p_{h,l} \in S_{h,l}$ . Then using the test function  $\varphi_{h,l} = \nabla \xi_{h,l}$  for some  $\xi_{h,l} \in S_{h,l}$  in (13.7) shows that  $(u_h - T_{h,l} \mu_h, \nabla \xi_{h,l}) = 0$  so that

$$\begin{aligned} (u_h - T_{h,l} u_h, v_{h,l}) &= (u_h - T_{h,l} u_h, \omega_{h,l}) \\ &\leq \|u_h - T_{h,l} u_h\|_{L^2(\kappa_l)}^3 \|\omega_{h,l}\|_{L^2(\kappa_l)}^3, \end{aligned}$$

and use of the Friedrichs inequality (13.8) completes the proof.  $\square$

Recall that the  $H(\text{curl}; \Omega)$  inner product is defined by

$$(u, v)_{H(\text{curl}; \Omega)} = (\nabla \times u, \nabla \times v) + (u, v),$$

and that the  $H_0(\text{curl}; \Omega)$  orthogonal projection  $P_H : H(\text{curl}; \Omega) \rightarrow X_H$  by

$$(P_H u_h - u_h, \varphi_H)_{H(\text{curl}; \Omega)} = 0 \quad \text{for all } \varphi_H \in X_H.$$

Similarly,  $P_{h,l} : H(\text{curl}; \Omega) \rightarrow X_{h,l}$  is defined by

$$(P_{h,l} \mu_h - u_h, \varphi_{h,l})_{H(\text{curl}; \Omega)} = 0 \quad \text{for all } \varphi_{h,l} \in X_{h,l}.$$

We now state the following theorem from Toselli [281] concerning the relationship between the  $H(\text{curl}; \Omega)$  norm of  $u_h$

$\in X_b$  and its projections into the local subspaces. This is quite elaborate to prove, and we direct the interested reader to Toselli's work.

**Lemma 13.8** *There exists a constant  $C$ , independent of  $b$  and  $H$ , such that for all  $u_b \in X_b$ ,*

$$(u_h, u_h)_{H(\text{curl}; \Omega)} \leq C \left( \sum_{l=1}^{M_H} (P_{h,l} u_h, u_h)_{H(\text{curl}; \Omega)} + (P_H u_h, u_h)_{H(\text{curl}; \Omega)} \right).$$

The final lemma is related to the additive version of the Schwarz preconditioner in [147], and will be used in the proof of our main result.

**Lemma 13.9** *Let  $\tilde{u}_b \in X_b$  be such that  $T_H \tilde{u}_b = 0$ . Then there is a constant  $C$ , independent of  $H$ , such that*

$$(\bar{u}_h, \bar{T}_h, \bar{u}_h)_{H(\text{curl}; \Omega)} \geq C (\bar{u}_h, \bar{u}_h)_{H(\text{curl}; \Omega)}.$$

**Proof** Using Lemma 13.8, the definition of  $T_{h,l}$  and of  $a(\cdot, \cdot)$ , and rearranging terms we obtain

$$\begin{aligned} C(\tilde{u}_h, \tilde{u}_h)_{H(\text{curl}; \Omega)} &\leq \sum_{l=1}^{M_H} (P_{h,l} \tilde{u}_h, \tilde{u}_h)_{H(\text{curl}; \Omega)} + (P_H \tilde{u}_h, \tilde{u}_h)_{H(\text{curl}; \Omega)} \\ &= \sum_{l=1}^{M_H} \left\{ a(T_{h,l} \bar{u}_h, P_{h,l} \tilde{u}_h) + (1 + \kappa^2)(\tilde{u}_h, P_{h,l} \tilde{u}_h) \right\} \\ &\quad + a(T_H \bar{u}_h, P_H \bar{u}_h) + (1 + \kappa^2)(\bar{u}_H, P_H \bar{u}_h). \end{aligned}$$

Using the assumption on  $\tilde{u}_b$ , and rewriting  $a(\cdot, \cdot)$  in terms of the  $(\cdot, \cdot)_{H(\text{curl}; \Omega)}$  inner product, this may be rewritten as

$$\begin{aligned} C(\tilde{u}_h, \tilde{u}_h)_{H(\text{curl}; \Omega)} &\leq \sum_{l=1}^{M_H} \left\{ (T_{h,l} \bar{u}_h, P_{h,l} \bar{u}_h)_{H(\text{curl}; \Omega)} \right. \\ &\quad \left. + (1 + \kappa^2)(\tilde{u}_h - T_{h,l} \bar{u}_h, P_{h,l} \bar{u}_h) \right\} \\ &\quad + (1 + \kappa^2)(\tilde{u}_h - T_H \bar{u}_h, P_H \bar{u}_h). \end{aligned}$$

Using Lemma 13.7 and the definition of the projection  $P_{h,l}$  we obtain

$$\begin{aligned} C(\bar{u}_h, \bar{u}_h)_{H(\text{curl}; \Omega)} &\leq \sum_{l=1}^{M_H} (T_{h,l} \bar{u}_h, \bar{u}_h)_{H(\text{curl}; \Omega)} \\ &\quad + C_1 H^{1/2+\delta} \left\{ \sum_{l=1}^{M_H} \|\tilde{u}_h - T_{h,l} \bar{u}_h\|_{H(\text{curl}; \kappa_l)} \|P_{h,l} \tilde{u}_h\|_{H(\text{curl}; \kappa_l)} \right. \\ &\quad \left. + \|\tilde{u}_h - T_H \bar{u}_h\|_{H(\text{curl}; \kappa_l)} \|P_H \bar{u}_h\|_{H(\text{curl}; \kappa_l)} \right\}, \end{aligned}$$

where  $\delta > 0$  is the parameter in Lemma 13.7. Then by Lemma 13.5 and Theorem 7.1, we can estimate  $\tilde{u}_b - T_{h,l} \tilde{u}_b$  and  $\tilde{u}_b - T_H \tilde{u}_b$ . In particular, using the

boundedness of the projection we obtain, via the finite covering property, the following estimate holds.

$$\begin{aligned} & \sum_{l=1}^{M_H} \| \tilde{u}_h - T_h \tilde{u}_h \|_{H(\text{curl}; \kappa_l)} \| P_{h,l} u_h \|_{H(\text{curl}; \kappa_l)} \\ & + \| \tilde{u}_h - T_H \tilde{u}_h \|_{H(\text{curl}; \Omega)} \| P_H u_h \|_{H(\text{curl}; \Omega)} \leq C_2(\tilde{u}_h, \tilde{u}_h)_{H(\text{curl}; \Omega)}. \end{aligned}$$

So there is a constant  $C_3$  such that

$$C(\tilde{u}_h, \tilde{u}_h)_{H(\text{curl}; \Omega)} \leq \sum_{l=1}^{M_H} (T_h \bar{u}_h, \tilde{u}_h)_{H(\text{curl}; \Omega)} + C_3 H^{1/2+\delta} (\tilde{u}_h, \tilde{u}_h)_{H(\text{curl}; \Omega)}.$$

If  $H$  is chosen so that  $C - C_3 H^{1/2+\delta}$  is positive, we get the desired result.  $\square$

Now we can prove the main result of this section.

**Proof of Theorem 13.2** For  $u_b \in X_b$ , set  $\tilde{u}_b = (I - T_b)u_b$ . Then (13.18)

$$\begin{aligned} \| \tilde{u}_b - a \tilde{T}_b \tilde{u}_b \|_{H(\text{curl}; \Omega)}^2 &= \| \tilde{u}_b \|_{H(\text{curl}; \Omega)}^2 - 2\alpha (\tilde{u}_b, \tilde{T}_b \tilde{u}_b)_{H(\text{curl}; \Omega)} \\ &+ \alpha^2 \| \tilde{T}_b \tilde{u}_b \|_{H(\text{curl}; \Omega)}^2. \end{aligned}$$

By Lemma 13.5 and the finite covering property, there is a constant  $C_1$  such that

$$\| \tilde{T}_b \tilde{u}_b \|_{H(\text{curl}; \Omega)}^2 \leq \sum_{l=1}^{M_H} C \| \tilde{u}_b \|_{H(\text{curl}; \kappa_l)}^2 \leq C_1 \| \tilde{u}_b \|_{H(\text{curl}; \Omega)}^2.$$

Lemma 13.7 estimates the second term in (13.18). We obtain

$$\| \tilde{u}_b - \tilde{T}_b \tilde{u}_b \|_{H(\text{curl}; \Omega)}^2 \leq (1 - 2\alpha C_2 + C_1 \alpha^2) (\tilde{u}_b, \tilde{u}_b)_{H(\text{curl}; \Omega)}.$$

Now we choose  $a$  small enough that  $1 - 2\alpha C_2 + \alpha_1 \alpha^2 = \gamma_1^2 < 1$ . Then using the boundedness of  $I - T_H$  guaranteed by Theorem 7.1, we have

$$\begin{aligned} \| \tilde{u}_b - \tilde{T}_b \tilde{u}_b \|_{H(\text{curl}; \Omega)}^2 &\leq \gamma_1 \| (I - T_H) u_b \|_{H(\text{curl}; \Omega)} \\ &\leq \frac{\gamma_1}{1 - CH^{1/2+\delta}} \| u_b \|_{H(\text{curl}; \Omega)}, \end{aligned}$$

where  $\delta > 0$  is the exponent in Lemma 7.6. Choosing  $H$  small enough shows that  $\gamma_1 / (1 - CH^{1/2+\delta}) = \gamma < 1$ . This completes the proof.  $\square$

We have now verified that  $I - B_b A_b$  has norm less than one independent of  $b$  and thus the iteration scheme (13.4) will converge at a rate independent of  $b$ . Of course, the Richardson scheme is not the preferred method for solving this type, but was presented for ease of exposition. Many researchers use instead the GMRES method. In [147] it is pointed out that the multiplicative preconditioner (and an additive Schwarz preconditioner also analyzed in that paper)

can be used in conjunction with GMRES. In their numerical tests of the above theory, Gopalkrishnan and Pasciak [147] found that the multiplicative method performed slightly better than the additive method (hence why we present only the multiplicative method). These numerical experiments also reveal that the constraint that  $H$  be “small enough” is necessary. An obvious question is just how small must  $H$  be chosen. A glance at the proof of the main theorem suggests that  $H$  has to be small enough to provide a quasi-optimal error estimate for the Maxwell problem on the coarse mesh with a constant sufficiently close to one. Thus, the coarse mesh must already be fine enough to provide some approximation to the solution of the scattering problem. In the next section we shall examine the question of how fine the mesh must be for this to happen in more detail. Concluding this section, we see that an important open problem is how to avoid the “coarse” grid and still maintain optimal convergence rates (if, indeed, this is possible).

### 13.3 Phase error in finite element methods

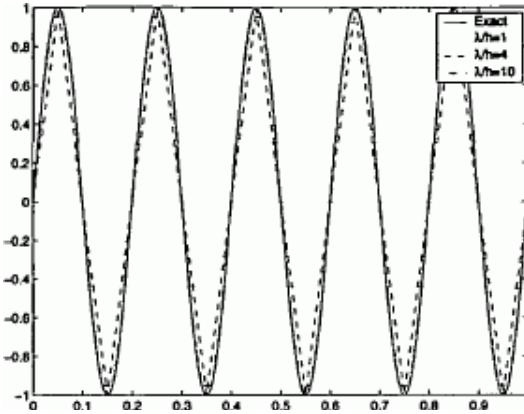
The error estimates we have proved so far guarantee quasi-optimal accuracy for a given problem provided all data are fixed and the mesh size  $h$  is sufficiently small (e.g. Theorem 7.1). In practice, we often want to solve scattering problems for a given geometry and incoming wave, but a variety of wavenumbers  $\kappa$ . Our error estimates do not include the effect of changing wavenumber. Generally, as the wavenumber increases, using a fixed grid, the error in the computed field increases [28, 249]. There seems to be very few papers that rigorously analyze the effect of changing wavenumber on error. In this section we shall provide some analysis motivated by the work of Ihlenburg and Babuška [169, 170] for a special model problem. For a more detailed discussion, see the book of Ihlenburg [168]. Another goal for this section is to provide some heuristics for choosing the mesh size in a time-harmonic electromagnetic computation.

Let us first recall that the Maxwell system

$$\nabla \times \nabla \times E - \kappa^2 E = 0 \text{ in } \mathbb{R}^3$$

has the following plane wave solution  $E = p \exp(i\kappa x \cdot d)$ , where  $|d| = |p| = 1$  and  $d \cdot p = 1$ . The wavelength of this plane wave is denoted by  $\lambda$  and given by  $\lambda = 2\pi/\kappa$ . Considerations of approximating sinusoidal waveforms by piecewise linear functions suggests that in order to approximate a plane wave the mesh must be sufficiently fine compared to the wavelength. Typical engineering rules of thumb suggest that for piecewise linear functions,  $h$  should be chosen so that  $\lambda \approx 10h$  or  $b\kappa \approx 2\pi/10$  to provide “reasonable” accuracy. In Fig. 13.1 we show a graph of the piecewise linear interpolant of  $\sin(\kappa x)$ ,  $0 \leq x \leq 1$ , for  $\kappa = 10\pi$  using equally spaced interpolation points for  $\lambda/b = 1, 4, 10$ . It turns out that, at least for oscillatory functions such as this, choosing  $h$  such that  $\kappa h$  is a fixed value provides a roughly uniformly accurate interpolant as  $\kappa$  changes. Thus as  $\kappa$  increases,  $h$  must correspondingly decrease, and  $h = O(1/\kappa)$ .

Fig. 13.1. The interpolant of  $f(x) = \sin(\kappa x)$ ,  $\kappa = 10\pi$  using  $\lambda/b = 1, 4$  and  $10$ . When  $\lambda/b = 1$  the approximation is very poor (100% error — see the dottedline), while for  $\lambda/b = 10$  the error is barely visible.



Of course for higher-order methods, we expect to be able to use a larger value of  $b$  and if we have a degree- $p$  polynomial space, then it is only necessary to control  $b\kappa/p$  in order to obtain satisfactory accuracy from the interpolant.

Unfortunately the approximation of Maxwell's equations by finite elements does not compute the interpolant. It turns out that in order to control the error in the finite element solution as  $\kappa$  increases, it is necessary to decrease  $b$  faster than  $O(1/\kappa)$ . Simply keeping  $\kappa b$  fixed as  $\kappa$  increases does not give a sequence of solutions of fixed accuracy. We examine this in detail in the next section using a one dimensional model problem. The second section provides some further insight into multidimensional problems.

### 13.3.1 Wavenumber dependent error estimates

To understand the wavenumber dependence of the error in the finite element solution, Ihlenburg and Babuška [169, 170] have examined a one-dimensional time-harmonic problem. We shall follow their lead by analyzing the problem of computing  $u \in H^1(0, 1)$  such that(13.19a)

$$u'' + \kappa^2 u = 0 \text{ on } (0,1) \quad (13.19b)$$

$$u(0) = 1, \quad (13.19c)$$

$$u'(1) - i\kappa u(1) = 0,$$

This problem has the advantage that the solution is easy to compute (just  $u = \exp(i\kappa x)$ ) and there are no singularities due to the geometry of the domain. Thus, the accuracy of the finite element method is only due to approximating the smooth time-harmonic wave  $u$ . As part of the analysis we shall present, we also need to consider the solution  $z \in H^1(0, 1)$  of an adjoint problem given by(13.20a)

$$z'' + \kappa^2 z = f \text{ on } (0,1),$$

$$z(0) = 0, \quad (13.20b)$$

$$z'(1) + i\kappa z(1) = 0, \quad (13.20c)$$

for a special choice of  $f \in L^2(0, 1)$ . The solution of this problem can be written using a Green's function [169] and it is possible to show that (13.21a)

$$\|z\|_{L^2[0,1]} \leq \kappa^{-1} \|f\|_{L^2[0,1]},$$

$$|z|_{H^2[0,1]} \leq (1 + \kappa) \|f\|_{L^2[0,1]}, \quad (13.21b)$$

where we recall that  $|\cdot|_{H^2[0,1]}$  is the semi-norm of order 2 on  $H^2(0, 1)$ .

Using the usual Galerkin strategy, we can derive a variational problem associated to (13.19). If  $\varphi$  is a smooth test function such that  $\varphi(0) = 0$ , then

$$0 = \int_0^1 (u'' + \kappa^2 u) \bar{\varphi} dx = \int_0^1 (\kappa^2 u \bar{\varphi} - u' \bar{\varphi}') dx + u'(1) \bar{\varphi}(1).$$

Hence using the boundary condition (13.19c),

$$\int_0^1 (u' \bar{\varphi}' - \kappa^2 u \bar{\varphi}) dx - i\kappa u(1) \bar{\varphi}(1) = 0.$$

So if  $a_{1D}(u, \varphi)$  denotes the left-hand side of this equation and

$$S = \left\{ \varphi \in H^1(0, 1) \mid \varphi(0) = 0 \right\},$$

we can see that  $u \in H^1(0, 1)$  satisfies  $u(0) = 1$  and (13.22)

$$a_{1D}(u, \varphi) = 0 \text{ for all } \varphi \in S.$$

To construct the usual piecewise linear approximation to  $u$ , we introduce a mesh

$$0 = x_0 < x_1 < \dots < x_N = 1,$$

where  $|x_l - x_{l-1}| = b$ ,  $l = 1, \dots, N$ . Then let (13.23)

$$S_h = \left\{ u_h \in H^1(0, 1) \mid u_h \Big|_{[x_{l-1}, x_l]} \in P_1 \text{ for } 1 \leq l \leq N \right\}.$$

and define (13.24)

$$S_{0,h} = \left\{ u_h \in S_h \mid u_h(0) = 0 \right\}.$$

Thus, the standard continuous piecewise linear approximation  $u_h \in S_h$  satisfies  $u_h(0) = 1$  and (13.25)

$$a_{1D}(u_h, \varphi_h) = 0 \text{ for all } \varphi_h \in S_{0,h}.$$

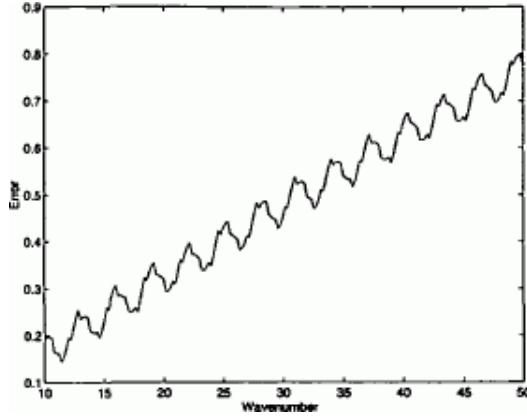
A unique solution to (13.25) exists. This follows from the same argument as in the proof of Theorem 7.1 and Corollary 7.3, and was shown by Babuška and Ihlenburg [169]. Furthermore, by the one-dimensional analogue of

Theorem 7.1,

$$\|u - u_h\|_{H^1(0,1)} \leq Ch \|u\|_{H^2(0,1)}.$$

In Fig. 13.2 we plot the relative error  $\|u - u_{hH}^{-1}(0,1)\| / \|u\|_{H^1(0,1)}$  against  $\kappa$  keeping  $\kappa h$  roughly fixed. Of course, keeping  $\kappa h$  fixed is not possible precisely, since  $1/h$

Fig. 13.2. Plot of relative  $H^1(0,1)$  norm error against wavenumber  $\kappa$  using a mesh such that  $\pi/\kappa h \geq 4$  (i.e. at least eight grid calls per wavelength). The error grows even though  $\kappa h$  is essentially fixed.



must be an integer. We choose the largest  $h$  such that there are at least eight grid intervals per wavelength (thus  $\kappa h \leq \pi/4$ ). Clearly the error is not controlled by keeping  $\kappa h$  bounded from above.

Heuristically, we can understand this effect as follows. At the left end of the interval  $(0, 1)$ , we have  $u_h(0) = u(0) = 1$ , but the wavelength of the numerical solution  $u_h(x)$  differs from the exact wavelength  $2\pi/\kappa$ . In fact, it is slightly too short. Thus, after one oscillation, the peak of the wave is displaced. After one more oscillation the peak is further displaced (almost double the previous displacement) and so on. So with each oscillation of the solution, the waves become progressively more and more out of phase and the error grows with  $\kappa$ . This can be seen in Fig. 13.3 . The phase error buildup depends on the number of wavelengths in the domain which increases as  $\kappa$  increases. Hence,  $h$  must be decreased faster than  $O(1/\kappa)$  to control this error.

Mathematically, this error buildup is proved in the following theorem [169] . We derive the error in the following  $\kappa$ -dependent norm:(13.26)

$$\|u\|^2 = \|u'\|_{L^2(0,1)}^2 + \kappa^2 \|u\|_{L^2(0,1)}^2.$$

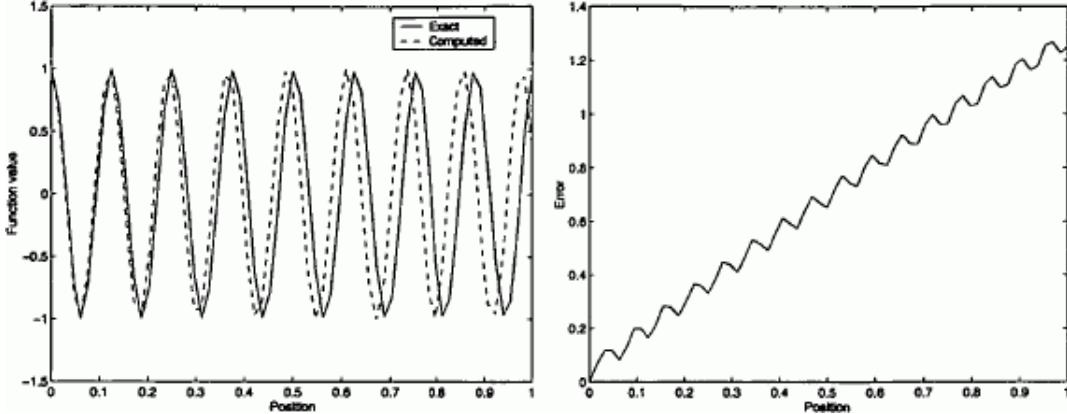
**Theorem 13.10** Let  $u_h \in S_h$  satisfy (13.25) and  $u \in S$  satisfy (13.22). Then, provided  $h^2\kappa^3$  is small enough, and  $\kappa \geq \kappa_0 > 0$  for some constant  $\kappa_0$  there is a constant  $C$  independent of  $h$ ,  $\kappa$ ,  $u$  and  $u_h$  (but depending on  $\kappa_0$ ) such that(13.27)

$$\|u - u_h\| \leq C(\kappa h + h^2\kappa^3)$$

**Remark 13.11** In (13.27) the term  $h\kappa$  corresponds to polynomial interpolation error and gives the asymptotic  $\mathcal{O}(h)$  rate of convergence. The term  $h^2\kappa^3$  must also be controlled and this accounts for phase error. For fixed  $h$   $\kappa$ , if  $\kappa$  is increased, this term will eventually cause a blowup in the error. We need

$$(h\kappa)^2\kappa \ll 1 \text{ and } h\kappa \ll 1.$$

Fig. 13.3. The real part of the exact solution  $u$  and the approximate solution  $u_h$  of the one-dimensional problem (*left panel*) and pointwise error (*right panel*). Here there are eight grid points per wavelength and  $\kappa = 50$ . Clearly the error grows from left to right as the numerical solution becomes progressively more and more out of phase with the exact solution.



So the number of grid points per wavelength needed to control phase error increases  $O(\sqrt{\kappa})$  as  $\kappa \rightarrow \infty$ .

Phase error causes linear finite element methods to become extremely expensive (particularly in three dimensions) when the domain of calculation spans many wavelengths.

For a fixed  $\kappa$ , as  $h$  is decreased, the error is first dominated by phase error governed by the term  $h^2\kappa^3$ . In this “pre-asymptotic” phase, the convergence rate can be quite different (here  $O(h^2)$ ) compared to the asymptotic convergence rate (here  $O(h)$ ). As  $h$  is decreased still further the dominant term in the error becomes  $O(h)$ , and we see the true asymptotic rate of convergence.

**Proof of Theorem 13.10** The proof is similar to that of the convergence of edge finite elements for Maxwell's equations in Section 7.2. First we derive a Gårding inequality:

$$\begin{aligned} \left\| u - u_h \right\|^2 &= \| (u - u_h)' \|_{L^2(0,1)}^2 + \kappa^2 \| (u - u_h) \|_{L^2(0,1)}^2 \\ &\leq a_{1D}(u - u_h, u - u_h) + 2\kappa^2 \| u - u_h \|_{L^2(0,1)}^2 \\ &\leq a_{1D}(u - u_h, u - \pi_h u) + 2\kappa^2 \| u - u_h \|_{L^2(0,1)}^2, \end{aligned}$$

where  $\pi_h u \in S_h$  is the interpolant of  $u$ . Here we have used the definition of  $a_{1D}(u, \varphi)$  and the fact that if (13.25) is subtracted from (13.22), then (13.28)

$$a_{1D}(u - u_h, \varphi_h) = 0 \quad \text{for all } \varphi_h \in S_{0,h}.$$

Using the interpolant  $\pi_h u$  of  $u$  in  $S_h$  removes the boundary term at  $x = 1$  from  $a_{1D}(u - \pi_h u)$  and using the Cauchy–Schwarz inequality, we have

$$\|u - u_h\|^2 \leq \|u - u_h\| \|u - \pi_h u\| + 2\kappa^2 \|u - u_h\|_{L^2(0,1)}^2. \quad (13.29)$$

It now remains to estimate  $\|u - u_h\|_{L^2(0,1)}$  using a duality argument. Let  $\zeta \in H^1(0, 1)$  satisfy

$$\begin{aligned} z'' + \kappa^2 z &= u - u_h \text{ in } (0,1), \\ z(0) &= 0, \\ z'(1) + i\kappa z(1) &= 0. \end{aligned}$$

This solution also exists and satisfies the *a priori* bound (13.21b) so that (13.30)

$$|z|_{H^2(0,1)} \leq (\kappa + 1) \|u - u_h\|_{L^2(0,1)}.$$

Now using  $\zeta$ , integration by parts and the boundary conditions on  $\zeta$  we have (13.31)

$$\begin{aligned} \|u - u_h\|_{L^2(0,1)}^2 &= \int_0^1 (u - u_h) \overline{(z'' + \kappa^2 z)} dx \\ &= -a_{1D}(u - u_h, z) \\ &= -a_{1D}(u - u_h, z - \pi_h z), \end{aligned}$$

where we have used the Galerkin orthogonality condition (13.28).

To obtain the correct estimate it is now necessary to use a trick from page 127 of [168] that relies on the fact that we are working in one dimension. By the definition of  $a_{1D}(\cdot, \cdot)$  and using integration by parts we may write

$$\begin{aligned} a_{1D}(\pi_h u - u_h, z - \pi_h z) &= \sum_{l=1}^{N-1} \int_{x_l}^{x_{l+1}} \left( (\pi_h u - u_h)' \overline{(z - \pi_h z)}' \right. \\ &\quad \left. - \kappa^2 (\pi_h u - u_h) \overline{(z - \pi_h z)} \right) dz \\ &= \sum_{l=1}^{N-1} \left. (\pi_h u - u_h)' \overline{(z - \pi_h z)} \right| - \int_{x_l}^{x_{l+1}} \left( (\pi_h u - u_h)'' \overline{(z - \pi_h z)} \right. \\ &\quad \left. + \kappa^2 (\pi_h u - u_h) \overline{(z - \pi_h z)} \right) dz \\ &\quad - \kappa^2 \int_0^1 (\pi_h u - u_h) \overline{(z - \pi_h z)} dx \end{aligned}$$

where we have used the fact that, since we are using a piecewise linear finite element space,  $(\pi_h u - u_h)'' = 0$  on each element, and, since  $\pi_h \zeta$  interpolates  $\zeta$  we also have  $(\pi_h u - u_h)'(x - \pi_h z)|_{x_l}^{x_{l+1}} = 0$  for each  $l$ .

Now using the above equality we have (13.32)

$$\begin{aligned} a_{1D}(u - u_h, z - \pi_h z) &= a_{1D}(u - \pi_h z) + a_{1D}(\pi_h u - u_h, z - \pi_h z) \\ a_{1D}(u - u_h, z - \pi_h z) &= \int_0^1 \left( (u - \pi_h u)' \overline{(z - \pi_h z)}' - \kappa^2 (u - u_h) \overline{(z - \pi_h z)} \right) dx \\ &\leq \|u - \pi_h u\|_{L^2(0,1)} \|z - \pi_h z\|_{L^2(0,1)} \\ &\quad + \kappa^2 \|u - u_h\|_{L^2(0,1)} \|z - \pi_h z\|_{L^2(0,1)}. \end{aligned}$$

The usual error estimate for piecewise linear interpolation (5.4) and the *a priori* estimate (13.30) now implies that

$$\|z - \pi_h z\|_{L^2(0,1)} \leq Ch^2 |z|_{H^2(0,1)} \leq Ch^2 (\kappa + 1) \|u - u_h\|_{L^2(0,1)}.$$

Using this estimate in (13.32) shows that

$$\begin{aligned} |a(u - u_h, z - \pi_h z)| &\leq \left\| (u - \pi_h u)' \right\|_{L^2(0,1)} \left\| (z - \pi_h z)' \right\|_{L^2(0,1)} \\ &\quad + Ch^2 \kappa^2 (\kappa + 1) \|u - u_h\|_{L^2(0,1)}^2 \end{aligned}$$

Note that  $C$  is just the constant from the interpolation estimate and therefore independent of  $\kappa$ ,  $h$  and  $u$ . Provided  $h$  is small enough that  $Ch^2(\kappa+1) \leq 1/2$  we have, using equality (13.31),

$$\|u - u_h\|_{L^2(0,1)}^2 \leq 2 \left\| (u - \pi_h u)' \right\|_{L^2(0,1)} \left\| (z - \pi_h z)' \right\|_{L^2(0,1)}.$$

Using this estimate in (13.31) we have

$$\begin{aligned} \left\| u - u_h \right\|^2 &\leq \left\| u - u_h \right\| \left\| u - \pi_h u \right\| \\ &\quad + C \kappa^2 \left\| (u - \pi_h u)' \right\|_{L^2(0,1)} \left\| (z - \pi_h z)' \right\|_{L^2(0,1)}. \end{aligned}$$

Again using the standard estimate for the error in piecewise linear interpolation and (13.30) this estimate may be written as

$$\begin{aligned} \left\| u - u_h \right\|^2 &\leq \left\| u - u_h \right\| \left\| u - \pi_h u \right\| \\ &\quad + C \kappa^2 h (\kappa + 1) \left\| (u - \pi_h u)' \right\|_{L^2(0,1)} \|u - u_h\|_{L^2(0,1)}. \end{aligned}$$

Via the arithmetic geometric mean inequality, and using the definition of the triple-bar norm, we obtain that there is a constant  $C$  independent of  $h$ ,  $\kappa$  and  $u$  such that

$$\left\| u - u_h \right\|^2 \leq C \left( \left\| u - \pi_h u \right\|^2 + \kappa^2 h^2 (\kappa + 1)^2 \left\| (u - \pi_h u)' \right\|_{L^2(0,1)}^2 \right).$$

Once more using error estimates for the piecewise linear interpolant and (13.21b) (this applies to  $u$  once the boundary condition is lifted to  $(0,1)$  and the problem is rewritten as a source problem with homogeneous boundary data), we obtain for all  $\kappa$  bounded away from zero,

$$\left\| u - u_h \right\|^2 \leq C \left( h^2 \kappa^2 + h^4 \kappa^4 + h^4 \kappa^6 \right)$$

which completes the proof.

One way of controlling phase error is to use a higher order method. This is well recognized in the engineering community [149]. If continuous degree  $p$

elements are used to approximate the one-dimensional problem (13.19), and if  $u$  is smooth, Ihlenburg and Babuška [170] prove that if  $b\kappa \leq a < \pi$

$$\|u - u_h\|_{H^1(0,1)} \leq C_1(p) \left(1 + C_2 K\left(\frac{\kappa h}{p}\right)\right) \left(\frac{h}{2p}\right)^p \|u\|_{H^{p+1}(0,1)},$$

where  $C = (\sqrt{2})^p (\pi p)^{\nu_2}$  and  $C_2$  is independent of  $b$ ,  $\kappa$  and  $p$ . Hence for oscillatory solutions, by which we mean that  $\|u\|_{H^{p+1}(0,1)} \leq C(1 + \kappa)^p \|f\|_{L^2(0,1)}$ , we have

$$\|u - u_h\|_{H^1(0,1)} \leq C_1(p) \left(\frac{h\kappa}{2p}\right)^p + c_2(p) \kappa \left(\frac{h\kappa}{2p}\right)^{p+1}.$$

It is thus clear that for higher-order methods the phase error term can be reduced both due to a larger  $p$  in  $(h\kappa/2p)$  and due to a higher power of  $(h\kappa/2p)$ . Note, however, that we still cannot decrease below a small integer number of interpolation points per wavelength (including interior nodes). For example, figures of four to five degrees of freedom per wavelength are mentioned for very high order methods. The goal of reducing the number of grid points per wavelength is one motivation for developing  $hp$  finite element codes [122, 286], spectral element methods in the time-domain [300, 34] and for using higher-order fixed degree edge elements in engineering codes (e.g. [149]).

### 13.3.2 Phase error in three dimensional edge elements

The previous discussion focussed in detail on continuous piecewise linear elements in one dimension, so the results are just suggestive of what happens to edge elements in three dimensions. Nevertheless, we do see phase error in finite element calculations for Maxwell's equations. This error is more serious for linear elements than quadratic elements. To my knowledge there has been little or no study of phase error for edge finite elements applied to time-harmonic problems. For time-dependent problems there has been more progress, and some of these results can be reinterpreted in the context of time-harmonic problems. Numerical evidence suggests that the phase error is  $O(h^r)$  for linear tetrahedral edge elements, and similarly for linear hexahedral edge elements (see [83, 84, 299]). Interestingly it seems that a uniform grid of cubes converted to a tetrahedral grid by subdividing into six tetrahedra is particularly poor from the point of view of dispersion error [299, 224]. For right hexahedral elements, it is possible to analyze the phase error for any order of polynomial subspace by relating the three-dimensional phase error to phase error for appropriate one-dimensional problems such as the one analyzed previously [83].

To illuminate this discussion further, let us now derive the phase error for linear hexahedral elements. For a more complete discussion, see [83]. We consider

$$\nabla \times \nabla \times E - \kappa^2 E = 0 \quad \text{in } \mathbb{R}^3,$$

and in particular the plane wave solution  $E = p \exp(i\kappa \cdot d)$ , for  $|d| = |p| = 1$  and  $d \cdot p = 0$ . Suppose  $\mathbb{R}^3$  is covered by an infinite mesh of cubes with sides of

length  $b$  and with edges parallel to the coordinate axes as in Chapter 6 . Denote the mesh by  $\tau_b$ , and let

$$X_h = \left\{ u_h \in H_{\text{loc}}(\text{curl}; \mathbb{R}^3) \mid u_h|_K \in Q_{0,1,1} \times Q_{1,0,1} \times Q_{1,1,0} \right. \\ \left. \text{for all elements } K \in \tau_h \right\} .$$

In other words  $X_b$  is the standard linear edge finite element space on  $\mathbb{R}^3$ . We seek  $E_b \in X_b$  such that(13.33)

$$\int_{\mathbb{R}^3} \nabla \times E_b \cdot \nabla \times \bar{\varphi}_h - \kappa^2 E_b \cdot \bar{\varphi}_h dV = 0 \quad \text{forall } \varphi_h \in X_h .$$

In particular we are interested in discrete solutions that interpolate a plane wave.

In general, the numerical wavenumber  $\kappa_b$  is such that  $\kappa_b \neq \kappa$  and so the wavelength of the numerical wave, denoted by  $\lambda_b$ , is such that  $\lambda_b = 2\pi/\kappa_b \neq 2\pi/\kappa = \lambda$ . This accounts for phase error in the solution. We define a measure of phase error to be  $|\kappa_b - \kappa|$  and note that  $\kappa_b$  depends on  $d$ . Thus, the numerical wavenumber is anisotropic in that it depends on the direction of propagation of the wave. Waves propagating in different directions on the mesh have slightly different wavelengths further complicating the phase error problem in three dimensions.

To simplify the presentation, we are going to analyze phase error for a masslumped approximation to (13.33). For an analysis of (13.33) with a full mass matrix, see [83] . In this mass-lumped approximation we replace the integral over  $\mathbb{R}^3$  by a quadrature. Let  $a_1^K, \dots, a_8^K$  denote the eight vertices of the parallelepiped  $K$ . For any sufficiently smooth function of compact support (continuous on each element), we write

$$\int_{\mathbb{R}^3} f(x) dV = \sum_{K \in \tau_h} \int_K f(x) dV \approx \sum_{K \in \tau_h} Q_K(f),$$

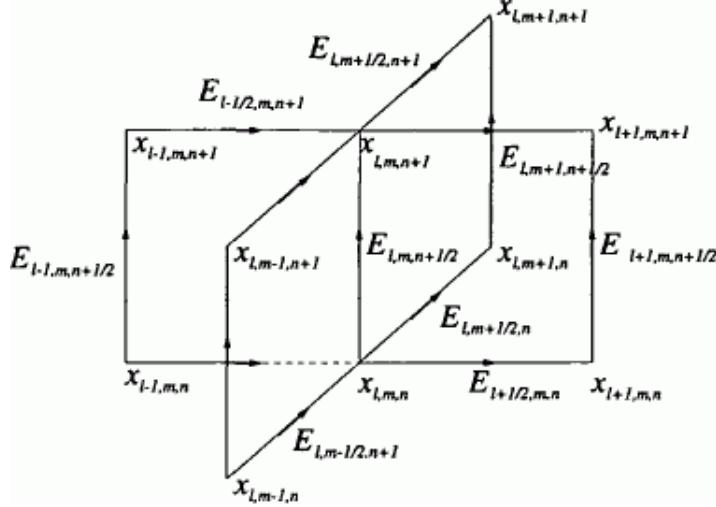
where  $Q_K$  is given by the three-dimensional trapezoidal rule

$$Q_K(f) = \frac{\text{vol}(K)}{8} \sum_{j=1}^8 f(a_j^K) .$$

This quadrature has the effect that the mass matrix corresponding to the term  $-\kappa^2 \int E_b \cdot \varphi_b dV$  is diagonal and certain couplings in the curl–curl matrix are also absent. In fact, evaluating the integrals in (13.33) by this quadrature results in the standard Yee finite difference scheme for this problem [301] .

By choosing  $\varphi_b$  in (13.33) to be a basis function for  $X_b$  with degrees of freedom that vanish except on one edge, we can write down the finite difference equation satisfied by the degrees of freedom of  $E_b$ . We suppose the grid to be a tensor product of grids  $x_{1,l} = lb$ ,  $l \in \mathbb{Z}$ ,  $x_{2,m} = mb$ ,  $m \in \mathbb{Z}$  and  $x_{3,n} = nb$ ,  $n \in \mathbb{Z}$ . Thus, nodes in the grid have coordinates  $(x_{1,l}, x_{2,m}, x_{3,n})$  for some integer  $l, m, n$

Fig. 13.4. Labeling of the vertices and degrees of freedom for elements surrounding the edge connecting node  $x_{l,m,n} = (x_{1,l}, x_{2,m}, x_{3,n})$ , to  $x_{l,m,n+1} = (x_{1,l}, x_{2,m}, x_{3,n+1})$ . Here the bold arrow marks the position of the degree of freedom for each edge, and we have marked some of the labels for the degrees of freedom (some are left off so as not to overly clutter the figure).



and we can index this node by  $(l, m, n)$ . Edges in the mesh can be indexed by the midpoint of the edge  $(l + 1/2, m, n)$  for edges in the  $x_1$ -direction (oriented in the positive  $x$  direction),  $(l, m+1/2, n)$  in the  $x_2$ -direction and  $(l, m, n+1/2)$  in the  $x_3$ -direction. With each edge is associated a degree of freedom which we take to be the value of the tangential component of the field at the midpoint of the edge (i.e. not scaled by the edge length although this is possible). So  $E_{l+1/2,m,n}$  denotes the value of  $(E)_1$  at  $(x_{1,l+1/2}, y_m, z_n)$ . See Fig. 13.4 for a summary of this convection for the edge connecting  $(x_{1,l}, x_{2,m}, x_{3,n})$  to  $(x_{1,l}, x_{2,m}, x_{3,n+1})$ . Choosing  $\varphi_b$  to interpolate 1 on this edge and zero elsewhere gives an equation relating  $E_{l,m,n+1/2}$  to surrounding values.

Tedious calculation then shows that the degrees of freedom satisfy the following difference equation:(13.34)

$$\begin{aligned} & \left(4 - \kappa^2 h^2\right) E_{1,m,n+1/2} + E_{l,m+1/2,n+1} + E_{l-1/2,m,n} + E_{l+1/2,m,n+1} \\ & + E_{l,m-1/2,n+1} - E_{l,m+1,n+1/2} - E_{l,m+1/2,n} - E_{l-1/2,m,n+1} - E_{l-1,m,n+1/2} \\ & - E_{l,m-1/2,n+1} - E_{l,m-1,n+1/2} - E_{l+1,m,n+1/2} - E_{l+1/2,m,n} = 0 \end{aligned}$$

Similar equations hold for the  $x_1$ -directed edge between  $(x_{1,l}, x_{2,m+1}, x_{3,n})$  and  $(x_{1,l+1}, x_{2,m}, x_{3,n})$  relating  $E_{l+1/2,m,n}$  to degrees on surrounding edges and on the edge connecting  $(x_{1,l}, x_{2,m}, x_{3,n})$  to  $(x_{1,l}, x_{2,m+1}, x_{3,n})$  relating  $E_{l,m+1/2,n}$  to degrees on surrounding edges. Since the mesh is translation invariant in the  $x_1$ -,  $x_2$ - and  $x_3$ -directions, these equations describe the entire, infinite, set of equations satisfied by the degrees of freedom.

We now seek plane wave solutions by substituting for the degrees of freedom, using (13.34) such that(13.35)

$$\begin{aligned} E_{l+\frac{1}{2},m,n} &= p_1 \exp(i\kappa_h(x_{1,l+\frac{1}{2}}d_1 + x_{2,m}d_2 + x_{3,n}d_3)), \\ E_{l,m+\frac{1}{2},n} &= p_2 \exp(i\kappa_h(x_{1,l}d_1 + x_{2,m+\frac{1}{2}}d_2 + x_{3,n}d_3)), \\ E_{l,m,n+\frac{1}{2}} &= p_3 \exp(i\kappa_h(x_{1,l}d_1 + x_{2,m}d_2 + x_{3,n+\frac{1}{2}}d_3)). \end{aligned}$$

We repeat this for the equation on the other two edges and obtain a  $3 \times 3$  matrix problem in which the coefficient matrix  $\mathcal{A}$  depends on  $\kappa_b$ ,  $\kappa$ ,  $b$ ,  $d$  and there is an unknown vector  $p_b = (p_1, b, p_2, b, p_3, b)^T$  related by

$$\mathcal{A}(\kappa_h, \kappa, h, d)p_b = 0.$$

In order to have non-trivial solutions,  $\mathcal{A}$  must have zero as an eigenvalue and MAPLE tells us that this implies

$$\kappa^2 h^2 = 4 \left( \sin^2 \left( \frac{\kappa_h d_1 h}{2} \right) + \sin^2 \left( \frac{\kappa_h d_2 h}{2} \right) + \sin^2 \left( \frac{\kappa_h d_3 h}{2} \right) \right).$$

Letting  $\mu = \kappa b$  we obtain(13.36)

$$\mu^2 = 4 \left( \sin^2 \left( \frac{\kappa_h \mu d_1}{2} \right) + \sin^2 \left( \frac{\kappa_h \mu d_2}{2} \right) + \sin^2 \left( \frac{\kappa_h \mu d_3}{2} \right) \right),$$

which emphasizes the role played by the product  $\kappa b$ . Consider the case  $d = (1, 0, 0)^T$  corresponding to one dimensional wave propagation along the edges in the mesh. Then

$$\mu^2 = 4 \sin^2 \left( \frac{\mu \kappa_h}{2\kappa} \right).$$

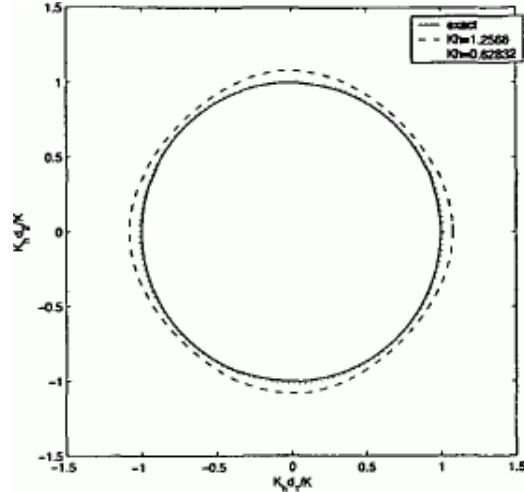
If  $\mu^2 > 4$ , we see that  $\kappa_b$  must be complex. This corresponds to exponential decay (or growth) of the discrete plane wave given by (13.35) and shows that the discrete scheme has become a very poor approximation to the true solution. This occurs when  $\kappa b > 2$  or  $\lambda/b < \pi$ , so we need at least four grid points per wavelength to maintain an oscillatory solution. This gives some idea of the number of grid cells per wavelength needed for the finite element method to maintain a realistic wave solution (and hence the size of “ $b$  sufficiently small” in our error estimates).

A Taylor series expansion of (13.36) shows that (13.34) implies(13.37)

$$\frac{\kappa_h}{\kappa} = 1 + \frac{1}{12} \left( d_1^4 + d_2^4 + d_3^4 \right) (\kappa h)^2 + O((h\kappa)^4).$$

Thus, the numerical wavenumber  $\kappa_b$  is larger than the true number and so the numerical wavelength is smaller. The numerical wave will slightly lag the true solution in phase. Fig. 13.5 shows a plot of  $\lambda_b$  as a function of  $d$  for  $\lambda/b = 5$  and  $\lambda/b = 10$ . The propagation anisotropy is obvious. However, it should

Fig. 13.5. A plot of  $\kappa_b d / \kappa$  when  $d = (\cos(\theta), \sin(\theta), 0)^T$  as  $\theta$  varies. We show results for five or ten grid cells per wavelength (and the exact result of a circle). For coarse grids there is noticeable grid anisotropy for this low-order scheme.



be mentioned that usually anisotropy will not cause problems if phase error is controlled.

A more detailed analysis of the finite element method using exact integration, and hence a standard mass matrix, shows that the corresponding error  $\kappa_b / \kappa - 1$  is the opposite sign to that in (13.36). In fact by taking a method that is a weighted average of the finite element difference equations with and without quadrature (with weight  $1/2$ ), we obtain a new scheme whose phase error  $\kappa_b / \kappa - 1 = O(h^4)$ . We shall not investigate this further since this result only applies to uniform grids.

For higher order methods on uniform parallelepiped grids, we can also compute the relation between  $\kappa$  and  $\kappa_b$  as in (13.36). There are then parasitic solutions due to internal nodes with the element [83]. For tetrahedra only computational results are available but the behavior of the phase velocity is even more complex. In some directions  $\kappa_b > \kappa$  and in other  $\kappa_b < \kappa$  (see [224]).

## 13.4 *A posteriori* error estimation

Using finite elements on a practical level, one is faced with the problem of determining the actual error in a given calculation. This is difficult to ascertain. Of course, one should perform a mesh convergence study in which the solution on a sequence of finer and finer meshes is compared. From the change in the solution as the mesh is refined, an indication of the overall accuracy can be obtained. Unfortunately we are often faced with the problem that we cannot afford to refine a three-dimensional mesh, say by halving the mesh size, many times before running out of computing resources. In practical computations, after producing

a grid, I usually check the phase accuracy of the solution by adjusting boundary conditions until a plane wave is the exact solution. The accuracy of the finite element method for approximating this problem is then easy to assess. If necessary the grid can be refined to produce a good approximation to a variety of plane waves in different directions. It is then likely that phase error is under control.

If we want to improve a solution computed on a given mesh, it would also be useful to refine the mesh in selected areas, or non-uniformly, to decrease the error by refining in those parts of the domain that are currently causing the majority of the error.

In attempting to solve problems of the type raised in the previous paragraphs, we are led to questions of *a posteriori* error analysis (estimation of the error in a given solution after computing the solution) and adaptivity (mesh modification to produce a more accurate solution). Obviously the goal is to adapt the mesh to produce a solution of a desired accuracy with close to minimum work. We are still far from the goal for Maxwell's equations.

This area is not very well developed for computational electromagnetics for high frequencies, although there have been some notable successes in the work of Demkowicz and co-workers [256]. This work uses a different *a posteriori* indicator to the one proposed here. It has been particularly important in pointing out that singularities do crop up in electromagnetic problems, and they can have a profound effect on the solution in the near field. The effect on the far field is not so clear. For low frequency problems, the error estimation problem is relatively well studied [31, 33, 32].

The fact that, due to limits on computing resources, many computational electromagnetic calculations are performed on a coarse grid in the *pre-asymptotic* convergence regime, where the error is dominated by phase error (see Section 13.3), implies that adaptive methods based on asymptotic estimates may be unreliable. In fact, the wavenumber dependence of *a posteriori* error estimates is an unsolved problem. In the next subsection we shall derive a residual based error estimator, and then perform a few numerical experiments to illuminate the problem.

### 13.4.1 A residual-based error estimator

In this section we derive an *a posteriori* estimator of Eriksson-Johnson type based on local residuals [138]. For simplicity, we restrict ourselves to the model problem from Chapter 4 rather than the full scattering problem in Chapter 12. The reader interested in *a posteriori* error indication for the full scattering problem can consult [256, 221]. The material for this section is essentially from [221].

The variational problem to be approximated is (4.4) where the coefficients satisfy the conditions of Section 4.2. The discrete edge element approximation to this problem is given by problem (7.2) using edge finite elements. We recall that the mesh is regular and that discontinuities of the coefficients fall on faces of the mesh only. It is useful to recall the bilinear form  $a(\cdot, \cdot)$  defined in (4.5) and to define the linear functional

$$F(u) = (F, u) + \langle g, \varphi_T \rangle.$$

In contrast to Chapter 7, in this section we shall assume that  $F \in H(\text{div}; \Omega)$ . We wish to estimate the  $(L^2(\Omega))^3$  norm of  $e_b = E - E_b$  where  $E_b \in X_b$  satisfies (7.2) and  $E \in X$  satisfies (4.4). An estimate for the error in the  $H(\text{curl}; \Omega)$  norm is then possible via the  $\square$ -inequality.

We start by deriving a general lemma concerning the  $(L^2(\Omega))^3$  error. To this end, we define the function  $\zeta \in X$  by

$$a(\varphi, z) = (\epsilon_r \varphi, e_h) \quad \text{for all } \varphi \in X,$$

where  $X$  is given in (4.3). This solution exists since it is the weak solution of the adjoint problem

$$\begin{aligned} \nabla \times \mu_r^{-1} \nabla \times z - \kappa^2 z &= \epsilon_r e_h \quad \text{in } \Omega, \\ v \times z &= 0 \text{ on } \Gamma \\ (\nabla \times z) \times v + i\kappa \lambda z_T &= 0 \text{ on } \Sigma. \end{aligned}$$

This problem can be analyzed via the Fredholm theory in exactly the same way as the standard problem for Maxwell's equations analyzed in Chapter 4 (or note that  $Z^-$  satisfies (4.4) with a suitable choice of data functions). Thus,  $\zeta \in X$  exists uniquely since we have assumed that  $\Sigma \neq \emptyset$ , and

$$\|z\|_{H(\text{curl}; \Omega)} \leq C \|\epsilon_r e_h\|_{(L^2(\Omega))^3}.$$

Using the Helmholtz decomposition (4.7) we have (13.38)

$$z = z_0 + \nabla p, \quad \text{for some } z_0 \in X_0 \text{ and } p \in S$$

where  $X_0$  is defined in (4.8) and  $S$  in (4.6). Of course, (13.39)

$$\|\nabla p\|_{(L^2(\Omega))^3} \leq C \|z\|_{(L^2(\Omega))^3} \leq C \|e_h\|_{(L^2(\Omega))^3}.$$

Choosing  $\varphi = e_b$  we have  $\alpha(e_b, z_0) + \alpha(e_b, \nabla p) = (\epsilon_r e_b, e_b)$ . Using the Galerkin property that  $\alpha(e_b, \varphi_b) = 0$  for all  $\varphi_b \in X_b$ , we obtain (13.40)

$$\begin{aligned} a(e_h, z_0 - \varphi_h) + a(e_h, \nabla(p - \xi_h)) &= (\epsilon_r e_h, e_h) \\ \text{for all } \xi_h \in S_h \text{ and } \varphi_h \in X_h, \end{aligned}$$

where  $S_b$  and  $X_b$  are defined in (7.23) and (7.1), respectively.

Now using the notation that for an element  $K$  and face  $f$ ,

$$(u, v)_K = \int_K u \cdot \bar{v} dV \quad \text{and} \quad \langle u_T, v_T \rangle_f = \int_f u_T \cdot \bar{v}_T dA,$$

we have, after integrating by parts over each tetrahedron  $K$  in the mesh  $\tau_b$ ,

$$\begin{aligned}
& a(e_h, z - \varphi_h) \\
&= \sum_{K \in T_h} \left\{ \left( \mu_r^{-1} \nabla \times e_h, \nabla \times (z_0 - \varphi_h) \right)_K - \kappa^2 (\in_r e_h, z_0 - \varphi_h)_K \right\} \\
&\quad - i \kappa \langle \lambda e_h, (z_0, \varphi_h)_T \rangle \\
&= \sum_{K \in T_h} \left\{ \left( \nabla \times \mu_r^{-1} \nabla \times e_h, z_0 - \varphi_h \right)_K - \kappa^2 (\in_r e_h, z_0 - \varphi_h)_K \right\} \\
&\quad + \sum_{K \in T_h} \left\{ \left( \mu_r^{-1} \nabla \times e_h, v_K \times (z_0 - \varphi_h) \right)_{\partial K} - i \kappa \langle \lambda e_h, (z_0, \varphi_h)_T \rangle \right\},
\end{aligned}$$

where  $v_K$  is the outward normal to  $K$ .

For a piecewise smooth function  $v$ , we now define the jump in the  $\mu_r^{-1}(\nabla \times v) \times v$  across a face  $f$  as follows. Suppose  $f$  is a face between two tetrahedra  $K_1$  and  $K_2$  in the mesh. Let  $v|_{K_1} = v_1$  and  $v|_{K_2} = v_2$ . Similarly, let  $\mu_r|_{K_l} = |K_l|^{-1/2}$ ,  $l = 1, 2$ . In addition, let the outward normal to  $K_l$  be denoted by  $v_l$ ,  $l = 1, 2$ . Then

$$[\nabla \times v]_T = \mu_{r,1}^{-1}(\nabla \times v_1) \times v_1 + \mu_{r,2}^{-1}(\nabla \times v_2) \times v_2.$$

For the true solution  $E$  of (4.4),  $[\nabla \times E]_T = 0$ , and thus, using, in addition, the fact that  $\nabla \times \mu_r^{-1} \nabla \times E - \kappa^2 \in_r E = F$  in  $\Omega$ , we have

$$\begin{aligned}
a(e_h, z_0 - \varphi_h) &= \sum_{K \in T_h} \left( F - \nabla \times \mu_r^{-1} \nabla \times E_h + \kappa^2 \in_r E_h, z_0 - \varphi_h \right)_K \\
&\quad + \sum_{f \in F_I} ([\nabla \times E_h]_T, (z_0 - \varphi_h)_T)_f \\
&\quad + \sum_{f \in F \setminus F_I} \left[ \mu_r^{-1} \nabla \times e_h \times v - i \kappa \lambda e_h, (z_0 - \varphi_h)_T \right],
\end{aligned}$$

where  $F_I$  is the set of all faces in the interior of  $\Omega$  and  $F_\Sigma$  is the set of all faces of the mesh on  $\Sigma$ . We have also used the fact that  $v \times (z - \varphi) = 0$  on  $\Gamma$  to eliminate these faces.

Using the fact that on  $\Sigma$ , the impedance trace  $\mu_r^{-1}(\nabla \times E) \times v - i \kappa E_T = g$ , we obtain, having also used the Cauchy–Schwarz inequality (13.41)

$$\begin{aligned}
& a(e_h, z_0 - \varphi_h) \\
&= \sum_{K \in T_h} \left\| F - \nabla \times \mu_r^{-1} \nabla \times E_h + \kappa^2 \in_r E_h \right\|_{L^2(\Omega)}^3 \|z_0 - \varphi_h\|_{L^2(\Omega)}^3 \\
&\quad + \sum_{f \in F_I} \|[\nabla \times E_h]_T\|_{L_t^2(f)} \| (z_0 - \varphi_h)_T \|_{L_t^2(f)} \\
&\quad + \sum_{f \in F \setminus F_I} \|g - \mu_r^{-1} \nabla \times E_h \times v + i \kappa \lambda E_h, (z_0 - \varphi_h)_T\|_{L_t^2(f)} \| (z_0 - \varphi_h)_T \|_{L_t^2(\Sigma)}.
\end{aligned}$$

The second term on the left-hand side of (13.40) is estimated similarly, using the fact that  $-i \kappa \nabla \cdot (e_h) = \nabla \cdot F$ ,

$$\begin{aligned}
& a(e_h, \nabla(p - \xi_h)) \\
&= -\kappa^2 \sum_{K \in T_h} (\in_r e_h, \nabla(p - \xi_h))_K \\
&= -\sum_{K \in T_h} \left( \nabla \cdot F + \kappa^2 \nabla \cdot \left( \in_r E_h, (p - \xi_h) \right)_K + \kappa^2 \left( \in_r e_h \cdot v_K, (p - \xi_h) \right) \right) \partial_K.
\end{aligned}$$

We now define, for a face  $f$  separating  $K_1$  and  $K_2$  and a piecewise smooth function  $v$ , the normal jump as follows

$$[v]_N = v_{r,1} \mathbf{v}_1 \cdot \mathbf{v}_1 + v_{r,2} \mathbf{v}_2 \cdot \mathbf{v}_2,$$

where  $\varepsilon_{r,l} = \varepsilon_r|_{\kappa_l}$ ,  $l = 1, 2$ . Obviously  $[\varepsilon_r E]_N = 0$  on each face interior to  $\Omega$ , and we may choose  $\xi_b = p$  on  $\sum$  since both functions are constants there. Hence, (13.42)

$$\begin{aligned} a(e_h, \nabla(p - \xi_h)) &\leq \left\{ \sum_{K \in \tau_h} \left\| \nabla \cdot F + \kappa^2 \nabla \cdot (\in_r E_h) \right\|_{L^2(K)} \|p - \xi_h\|_{L^2(K)} \right. \\ &\quad \left. + \kappa^2 \sum_{f \in FI} \| [E_h]_N \|_{L^2(f)} \|p - \xi_h\|_{L^2(f)} \right\}. \end{aligned}$$

Using (13.41) and (13.42) in (13.40), we have proved the following basic lemma.

**Lemma 13.12** Assume that jumps in  $\varepsilon_r$  or  $\mu_r$  lie on faces of the regular mesh  $\tau_b$ . Then for any  $\varphi_b \in X_b$  and  $\xi_b \in S_b$  with  $\xi_b = p$  on  $\sum$ , we have

$$\begin{aligned} |\in_r e_h, e_h| &\leq \sum_{K \in \tau_h} \left\| F - \nabla \times \mu_r^{-1} \nabla \times E_h + \kappa^2 \in_r E_h \right\|_{\left(L^2(K)\right)^3} \|z_0 - \varphi_h\|_{\left(L^2(K)\right)^3} \\ &\quad + \sum_{K \in \tau_h} \left\| \nabla \cdot F + \kappa^2 \nabla \cdot (\in_r E_h) \right\|_{L^2(K)} \|p - \xi_h\|_{L^2(K)} \\ &\quad + \sum_{f \in Fr} \left\{ \| [\nabla \times E_h]_T \|_{L_t^2(f)} \| (z_0 - \varphi_h)_T \|_{L_t^2(f)} \right. \\ &\quad \left. + \kappa^2 \| [E_h]_N \|_{L^2(f)} \|p - \xi_h\|_{L^2(f)} \right\} \\ &\quad + \sum_{f \in F_\Sigma} \left\| g - \mu_r^{-1} \nabla \times E_h \times \mathbf{v} + i\kappa\lambda E_{hr} \right\|_{L_t^2(f)} \| (z_0 - \varphi_h)_T \|_{L_t^2(f)}. \end{aligned}$$

Here  $z_0$  and  $p$  are given by the Helmholtz decomposition of the solution  $z$  of the adjoint problem in (13.38).

This estimate can be used directly. For example an approximation to  $z_0$  and  $p$  could be computed on a finer mesh, and the estimate evaluated directly (see [141]).

To derive a classical residual based error estimator, we now need to make some specific choices of  $\xi_b$  and  $\varphi_b$ . The choice of  $\xi_b$  is easiest. In general, we cannot expect  $p$  to be smoother than  $p \in H^1(\Omega)$  and so we choose  $\xi_b = \Pi_{\text{clem}} \xi$  where  $\Pi_{\text{clem}}$  is the Clément interpolant defined in Section 5.6.1. Recalling that

$b_K$  denotes the diameter of an element  $K$  and letting  $K_f$  denote the element of largest diameter having  $f$  as a face, we obtain

$$\begin{aligned} & \sum_{K \in \mathcal{T}_h} \left\| \nabla \cdot F + \kappa^2 \nabla \cdot (\in {}_r E_h) \right\|_{L^2(K)} \| p - \prod_{\text{Clem}} p \|_{L^2(K)} \\ & + \sum_{f \in \mathcal{F}_1} \| [E_h]_N \|_{L^2(f)} \| p - \prod_{\text{Clem}} p \|_{L^2(f)} \\ & \leq \left( \sum_{K \in \mathcal{T}_h} h_K^2 \| \nabla \cdot F + \kappa^2 \nabla \cdot (\in {}_r E_h) \|_{L^2(K)}^2 \right)^{1/2} \\ & \times \left( \sum_{K \in \mathcal{T}_h} \frac{1}{h_K^2} \| p - \prod_{\text{Clem}} p \|_{L^2(K)}^2 \right)^{1/2} \\ & + \left( \sum_{f \in \mathcal{F}_1} h_{K_f} \| [E_h]_N \|_{L^2(K)}^2 \right)^{1/2} \\ & \times \left( \sum_{f \in \mathcal{F}_1} \frac{1}{h_{K_f}} \| p - \prod_{\text{Clem}} p \|_{L^2(f)}^2 \right)^{1/2}. \end{aligned}$$

But because of the regularity of the mesh, there is a constant  $C$  such that

$$\sum_{f \in \mathcal{F}_I} \frac{1}{h_{K_f}} \| p - \prod_{\text{Clem}} p \|_{L^2(f)}^2 \leq C \sum_{K \in \mathcal{T}_h} \frac{1}{h_K} \| p - \prod_{\text{Clem}} p \|_{L^2(\partial K)}^2.$$

Thus, using estimate (5.51) for the Clément interpolant, and the *a priori* estimate for  $p$  in (13.39), we have

$$\begin{aligned} & \sum_{K \in \mathcal{T}_h} \| \nabla \cdot F + \kappa^2 \nabla \cdot (\in {}_r E_h) \|_{L^2(K)} \| p - \prod_{\text{Clem}} p \|_{L^2(K)} \\ & + \sum_{f \in \mathcal{F}_1} \| [E_h]_N \|_{L^2(f)} \| p - \prod_{\text{Clem}} p \|_{L^2(f)} \\ & \leq C \left( \sum_{K \in \mathcal{T}_h} h_K^2 \| \nabla \cdot F + \kappa^2 \nabla \cdot (\in {}_r E_h) \|_{L^2(K)}^2 \right)^{1/2} \| e_h \|_{(L^2(\Omega))^3} \\ & + \left( \sum_{f \in \mathcal{F}_1} h_{K_f} \| [E_h]_N \|_{L^2(f)}^2 \right)^{1/2} \| e_h \|_{(L^2(\Omega))^3}. \end{aligned}$$

The estimation of the terms involving  $\zeta_0 - \varphi_b$  is a little more tricky. It depends on the regularity of  $\zeta$  (because we do not currently have an analogue of the Clément interpolant for edge elements). Let  $\delta_K > 0$  and  $\delta_f > 0$  be exponents associated with edges and faces to be given shortly. Then, proceeding as before,

$$\begin{aligned}
& \sum_{K \in \tau_h} \|F - \nabla \times \mu_r^{-1} \nabla \times E_h - \kappa^2 \in_r E_h\|_{(L^2(K))^3} \|z_0 - \varphi_h\|_{(L^2(K))^3} \\
& + \sum_{f \in F_1} \|[\nabla \times E_h]_T\|_{L_t^2(f)} \|(z_0 - \varphi_h)_T\|_{L_t^2(f)} \\
& + \sum_{f \in F_1} \|g - \nabla \times E_h \times v + i\kappa\lambda E_h T\|_{L_t^2(f)} \|(z_0 - \varphi_h)_T\|_{L_t^2(f)} \\
& \leq C \left\{ \sum_{K \in \tau_h} h_K^{2\delta_K} \|F - \nabla \times \mu_r^{-1} \nabla \times E_h - \kappa^2 \in_r E_h\|_{(L^2(K))^3}^2 \right. \\
& + \sum_{f \in F_1} h_{K_f}^{2\delta_f} \|[\nabla \times E_h]_T\|_{L_t^2(f)}^2 \\
& \left. + \sum_{f \in F} h_{K_f}^{2\delta_f} \|g - \nabla \times E_h \times v + i\kappa\lambda E_h T\|_{L_t^2(f)}^2 \right\}^{1/2} \\
& \times \left\{ \sum_{K \in \tau_h} h_K^{-2\delta_K} \|z_0 - \varphi_h\|_{(L_t^2(f))^3}^2 \right. \\
& \left. + \sum_{f \in F} h_K^{-2\delta_K} \|(z_0 - \varphi_h)_T\|_{(L^2(\partial K))^3}^2 \right\}^{1/2}.
\end{aligned}$$

Away from boundaries and jumps in the coefficients  $\epsilon_r$  and  $\mu_r$  we expect  $z_0$  to be smooth, but near singularities caused by the boundary or by jumps in the coefficients it will be less smooth [106, 107]. To simplify the presentation, we suppose  $\epsilon_r$  and  $\mu_r$  are smooth enough (continuously differentiable is enough!) that  $z_0 \in H^s(\text{curl}; K)$ , for some  $s$  with  $\frac{1}{2} < s \leq 1$ , for all  $K \in \tau_b$ . Then using the estimates in Theorem 5.41 and Lemma 5.53 for the interpolant  $r_h$  and choosing  $\delta_K = s$  and  $\delta_f = s - \frac{1}{2}$  we have

$$\begin{aligned}
& \sum_{K \in \tau_h} h_K^{-2s} \|z_0 - r_h z_0\|_{(L^2(K))^3}^2 + h_K^{-2(s-1/2)} \|z_0 - r_h z_0\|_{(L^2(K))^3}^2 \\
& \leq C \sum_{K \in \tau_h} \|z_0\|_{H^s(\text{curl}; K)}^2,
\end{aligned}$$

and we also need to assume that the following *a priori* estimate holds

$$\sum_{K \in \tau_h} \|z_0\|_{H^s(\text{curl}; K)}^2 \leq C \|e_h\|_{(L^2(\Omega))^3}^2.$$

Note that this estimate holds (with  $s = 1$ ), if  $\Omega$  is convex and the coefficients are smooth for example.

We thus we have the following theorem.

**Theorem 13.13** *Assume that  $\tau_b$  is a regular mesh and that  $\Omega, \epsilon_r$  and  $\mu_r$  satisfy the conditions of Section 4.2. In addition suppose that the mesh is such that  $\epsilon_r$  and  $\mu_r$  are differentiable functions on each tetrahedron (jumps occur at faces of*

the mesh) and that  $F \in H(\operatorname{div}; \Omega)$ . Assume that there is an exponent  $s$  with  $\frac{1}{2} < s \leq 1$  such that  $\tilde{\zeta}_0 \in H(\operatorname{curl}; K)$  for all  $K \in \tau_h$ . In addition, we assume

$$\sum_{K \in \tau_h} \|z_0\|_{H^s(\operatorname{curl}; K)}^2 \leq C \|e_h\|_{(L^2(\Omega))^3}^2.$$

Then there is a constant  $C$  independent of  $E$ ,  $E_h$  and  $b$  such that

$$\begin{aligned} \|E - E_h\|_{(L^2(\Omega))^3} &\leq C \left\{ \sum_{K \in \tau_h} \left( h_K^{2s} \|F - \nabla \times \mu_r^{-1} \nabla \times E_h + \kappa^2 \in_r E_h\|_{(L^2(K))^3}^2 \right. \right. \\ &\quad + h_K^2 \|\nabla \cdot F - \nabla \cdot (\in_r E_h)\|_{L^2(K)}^2 \Big) \\ &\quad + \sum_{f \in F_1} \left( h_{K_f}^{2(s-1/2)} \|[\nabla \times E_h]_T\|_{L^2(f)}^2 + h_{K_f} \|[E_h]_N\|_{L^2(f)}^2 \right) \\ &\quad \left. \left. + \sum_{f \in F} h_{K_f}^{2(s-1/2)} \|g - \nabla \times E_h \times v + i \kappa \lambda E_h\|_{L^2(f)}^2 \right) \right\}^{1/2}. \end{aligned}$$

**Remark 13.14** If  $\mu_r$  and  $\kappa$  are smooth, and  $\Omega$  is convex, the estimate holds with  $s = 1$ .

The above estimate can be criticized for three things:

- First of course is the appearance of  $s$ . It seems likely that even for non-convex domains  $s = 1$  is the correct choice, but without a better interpolant (i.e. one defined on function with less regularity) this is as much as we can say without a detailed analysis of the singularities of  $E$ .
- Second, the constant  $C$  is dependent on the wavenumber  $\kappa$ . It is expected (from experience in two dimensions) that  $C$  will grow with  $\kappa$ . Better results might be obtained using the dimensionless quantity  $bx$  in place of  $b$  throughout the estimate. Of course,  $C$  is not known even for a posteriori estimates of this type for simpler problems like the Dirichlet problem for Poisson's equation, but in that case  $C$  is at least constant.
- Third, the constant  $C$  will depend on the size of jumps in the coefficients. A more detailed analysis would include weights to take care of regions where the coefficients are poorly behaved (see, e.g., [76] for the case of the Laplace equation with discontinuous coefficients).

All these improvements have yet to be performed, and significant work would be needed to make them possible.

### 13.4.2 Numerical experiments

To get some ideas of the issues involved with error estimation for time-harmonic problems, we can consider a simple one-dimensional model problem. This problem has no singularities so that we can focus on the problem that distinguishes the boundary value problems for time-harmonic waves from simple uniformly elliptic problems namely propagation error. We seek  $u \in H^1(0, 1)$  such that

$$\begin{aligned} u'' + \kappa^2 u &= 0(0, 1), \\ u(0) &= 1, \\ u'(1) - i \kappa u(1) &= 0. \end{aligned}$$

The solution to this problem is  $u(x) = \exp(i\kappa x)$ . Now we seek a numerical approximation  $u_h \in S_b$  where  $S_b$  is the standard space of piecewise linear functions on a uniform mesh of size  $h$  given by (13.23) (thus the mesh points are  $x_j = (j - 1)h, j = 1, 2, \dots, N + 1$ , where  $N = 1/h$ ). The discrete solution  $u_h \in S_b$  satisfies

$$\int_0^1 u'_h \varphi'_h - \kappa^2 u_h \varphi_h dx - i \kappa u_h(1) \varphi_h(1) = 0 \quad \text{forall } \varphi_h \in S_{0,h}$$

and  $u_h(0) = 1$ , where  $S_{0,h}$  is given by (13.24).

Applying the error estimation philosophy developed in the proof of the previous theorem to this problem shows that we should define  $\zeta$  by

$$\begin{aligned} z'' + \kappa^2 z &= u - u_h && \text{in } (0, 1), \\ z(0) &= 0, \\ z'(1) + i \kappa z(1) &= 0. \end{aligned}$$

Here  $\zeta \in H^2(0, 1)$  and  $\|\zeta\|_{H^2(0, 1)} \leq C \|u - u_h\|_{L^2(0, 1)}$ . The *a posteriori* error estimator derived by taking  $\zeta_h = \pi_h \zeta$  (i.e. the piecewise linear interpolant) is

$$\|u - u_h\|_{L^2(0, 1)} \leq C \left( \sum_{j=1}^{N-1} \kappa^4 h^4 \int_{x_j}^{x_{j+1}} |u_h|^2 dx \right)^{1/2}.$$

Figure 13.6 shows that for fixed  $\kappa = 50$ , the indicator

$$E_1 = \left( \sum_{j=1}^{N-1} \kappa^4 h^4 \int_{x_j}^{x_{j+1}} |u_h|^2 dx \right)^{1/2}$$

parallels  $\|u - u_h\|_{L^2(0, 1)}$  as  $N$  increases. This is the result claimed in Theorem 13.13.

As we have commented previously, the constant of proportionality in Theorem 13.13 depends on the wavenumber  $\kappa$ . Figure 13.7 shows how the approximate constant of proportionality changes with  $\kappa$ . We choose a very fine discretization,  $N = 1000$ , and compute the ratio  $\|u - u_h\|_{L^2(0, 1)}^2 / E_1$ . On average, the constant of proportionality increases with  $\kappa$ .

Figure 13.8 (a) shows the pointwise error  $|u - u_h(x)|$  against  $x$  for  $\kappa = 50$  and  $N = 1000$ . As is to be expected, the error grows from left to right due to the accumulation of phase error, across the domain. A plot of the local error indicator (13.43)

$$h^4 \kappa^4 \int_{x_j}^{x_{j+1}} |u_h|^2 dx$$

at  $(x_{j+1} + x_j)/2$  is shown in Fig. 13.8 (b). The error indicator is roughly constant across the domain and does not have any relation to the local error in the method.

Fig. 13.6. *Left:* plot of the error indicator  $E_1$  against the number of grid points  $N$  when  $\varkappa = 50$ . *Right:* plot of the ratio  $\|u - u_b\|_{L^2(0,1)}^2/E_1$  against the number of grid points  $N$ . As expected this ratio approaches a constant value as  $N$  increases.

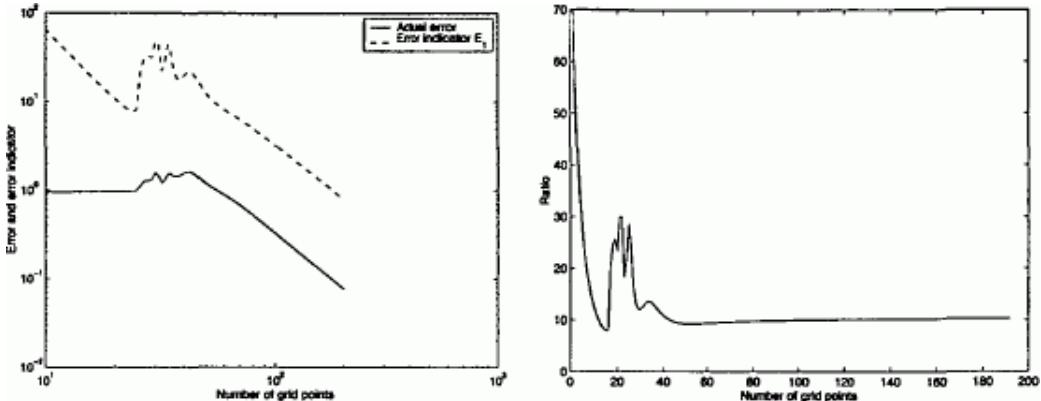
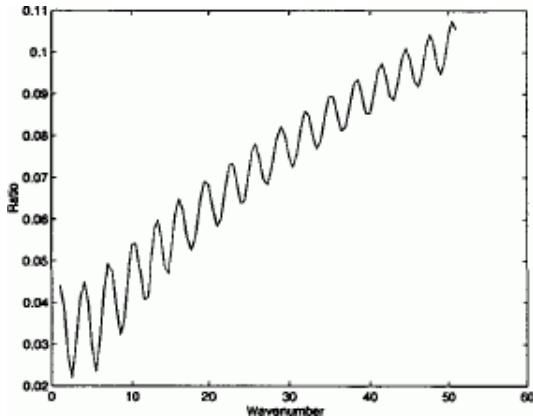


Fig. 13.7. The ratio  $\|u - u_b\|_{L^2(0,1)}^2/E_1$  as a function of wavenumber  $\varkappa$  using  $N = 1000$ . This shows that the constant of proportionality  $C$  in Theorem 13.13 is  $\varkappa$  dependent (generally increasing with  $\varkappa$ ).

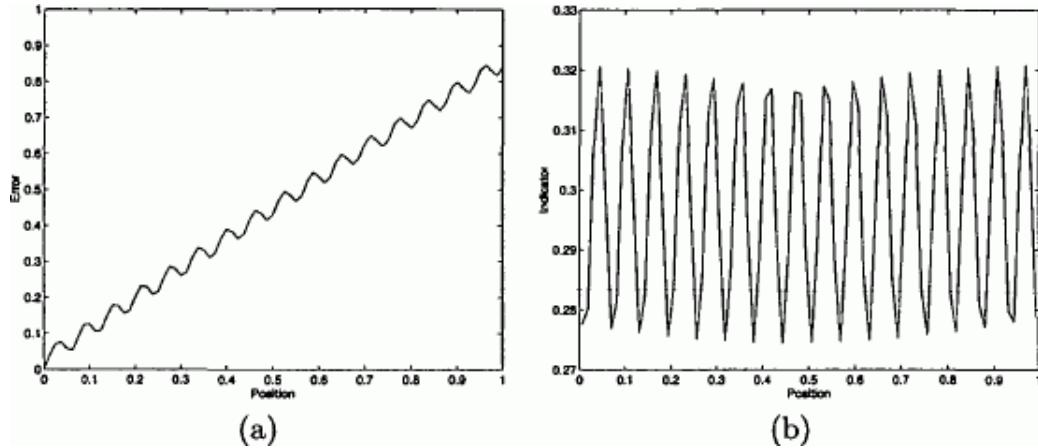


In this case, uniform refinement seems like a good strategy (as suggested by the roughly constant indicator). We see that just refining those parts of the grid with a large local indicator is not sufficient for the model problem. How to decide when to refine globally (as would be the correct strategy for this problem) and when to refine locally (e.g. due to a singularity in the field) needs further examination.

## 13.5 Absorbing boundary conditions

A theme throughout this book has been the development of numerical methods, based on finite elements, that can approximate scattering problems posed on unbounded domains. The first idea, proposed heuristically in Section 1.3, is to

Fig. 13.8. *Left:* pointwise error  $|(\boldsymbol{u} - \boldsymbol{u}_h)(x)|$  plotted against  $x$  for  $\kappa = 50$  and  $N = 1000$ . Due to the accumulation of phase error this increases from left to right. *Right:* the local error indicator  $E_i$  in (13.43) against  $x$ . This is approximately constant, and is consistent with the fact that uniform refinement would be a good strategy for this problem.



observe that the scattered wave satisfies the Silver–Müller radiation condition to progressively better accuracy moving away from the scatterer (as  $\varrho \rightarrow \infty$  in (1.22)). Thus, if an auxiliary boundary  $\Sigma$  is introduced sufficiently far from the scatterer, we expect that solving a boundary value problem on a bounded domain using the Silver–Müller radiation condition as a boundary condition on  $\Sigma$  will give rise to an approximate electromagnetic field that is close to the true solution. Furthermore, moving the boundary  $\Sigma$  further from the scatterer should result in an improved, or more accurate, solution.

We start this section by justifying this approach, but only for a special choice of auxiliary boundary  $\Sigma$ . The second part of this section is devoted to a description and some preliminary analysis of the infinite element method [72]. The final part of this section introduces the famous Bérenger PML.

### 13.5.1 Silver–Müller absorbing boundary condition

In this section we assume that  $\Sigma = \Sigma_R = \partial B_R$ , that is the surface of a sphere of radius  $R$  centered at the origin (here we include a subscript  $R$  to underline that  $\Sigma$  will change as  $R$  changes). Let  $R_0$  be chosen big enough such that  $D \subset B_{R_0}$ . Then we assume  $R \geq R_0$  and let  $\Omega_R$  denote the region exterior to  $D$  and interior to  $B_R$  ( $\Omega_R = B_R \setminus D$ ).

Let  $E \in H_{\text{loc}}(\text{curl}; \mathbb{R}^3 \setminus \bar{D})$  denote the weak solution of (12.4) when  $F = 0$  (e.g. using a plane-wave as the incoming wave). We denote by  $E_R$  the approximation to  $E$  obtained by finding the weak solution of  $E_R \in X$  of the following problem (recall  $X$  is given by (4.3)): (13.44)

$$\left. \begin{aligned} \nabla \times \nabla \times E_R - \kappa^2 E_R &= 0 \\ E_R &= E^i + E_R^s \end{aligned} \right\} \text{in } \Omega_R$$

$$E_R \times u = 0 \quad \text{on } \Gamma, \quad (13.45)$$

$$\nabla \times E_R^s \times \hat{x} - i \kappa E_{R,T}^s = 0 \quad \text{on } \Sigma_R. \quad (13.46)$$

Here  $E^i$  is a given incident field, assumed to be an analytic solution of Maxwell's equations in  $\mathbb{R}^3$ , (e.g. (1.20)). As proved in Chapter 4, this problem has a unique solution in  $H(\text{curl}; \Omega)$ . We have the following estimate.

**Theorem 13.15** *Let  $B$  be a fixed bounded domain contained in  $\Omega_{R_0}$ . Then there exists a constant  $C$  independent of  $R$  such that for all  $R$  large enough*

$$\|E - E_R\|_{H(\text{curl}; B)} \leq CR^{-2}.$$

*The constant  $C$  depends on  $B$ .*

**Remark 13.16** *The rather slow convergence in  $R$  shown in the above theorem accounts for the fact that, if high accuracy is needed, it is necessary to take  $\Sigma_R$  far from the scatterer. Since  $\Omega_R$  must be filled with finite elements, this makes the method time consuming.*

*Note that if  $E_{R,h}$  is a finite element approximation of  $E_R$ , then we have*

$$\|E - E_{R,h}\|_{H(\text{curl}; B)} \leq \|E - E_R\|_{H(\text{curl}; B)} + \|E_R - E_{R,h}\|_{H(\text{curl}; B)}.$$

*The first term on the right-hand side converges to zero as  $R \rightarrow \infty$ , while the second term converges as  $h \rightarrow 0$ . For a fixed mesh size  $h$ , increasing  $R$  can actually cause the total error on the left-hand side to increase if the increase in phase error in the second term offsets the decrease due to larger  $R$  in the first term. In general, when increasing  $R$  it is necessary to decrease  $h$  in order to keep the error  $\|E_R - E_{R,h}\|_{H(\text{curl}; B)}$  from growing (see our discussion of phase error in Section 13.3 of this chapter).*

The theorem is proved by an argument due to Goldstein [145], who analyzed the corresponding method for the Helmholtz equation. Our proof follows his with the necessary modifications for handling Maxwell's equations. First we need a lemma concerning the radial dependence of solutions of Maxwell's equations.

**Lemma 13.17** *Suppose  $E$  satisfies (12.4) with  $F = 0$  and let  $E^s = E - E^i$ . Define  $u = \exp(-i\kappa|x|)E^s$  for  $Q = |x| \geq R_0$ . Then for all  $Q > R_0$ ,* (13.47)

$$\begin{aligned} \left( \frac{\partial}{\partial \rho} \right)^\alpha u &= \sum_{n=1}^{\infty} \alpha_n^\alpha(\hat{x}) \rho^{-n-\alpha}, \\ \left( \frac{\partial}{\partial \rho} \right)^\alpha u &\leq C_\rho^{-1-\alpha} \|E^s\|_{H(\text{curl}; \Omega_{R_0})}, \end{aligned} \quad (13.48)$$

$$|(\nabla \times u) \times \hat{x}| \leq C_\rho^{-3} \|E^s\|_{H(\text{curl}; \Omega_{R_0})}. \quad (13.49)$$

**Remark 13.18** *A direct computation shows that  $\nabla \times u = \exp(-i\kappa|x|)\nabla \times E^s - i\kappa \exp(-i\kappa|x|)\hat{x} \times E^s$ , so  $|\nabla \times u| \times |\hat{x}| = |\nabla \times E^s \times \hat{x} - i\kappa E^s|$  where as usual  $\hat{x}$  is a unit radial vector. Hence, (13.49) is an estimate of how well the Silver–Müller absorbing boundary condition is satisfied by the solution of the problem (12.4).*

**Proof of Lemma 13.17** Estimate (13.47) follows from the well-known Wilcox expansion of the solution [296]. A simple but indirect proof is given in Theorem 4.8 of [93]. A direct proof is possible by using the Stratton–Chu formula (9.15), and expanding the fundamental solution  $\Phi$  of the Helmholtz equation as a series in  $|x|$  as in the proof of Corollary 9.5.

The second estimate (13.48) also follows from the Stratton–Chu formula (9.15), using the regularized form in (12.8). Again we use a series approximation of  $\Phi$  valid for large  $\rho$ .

To prove the last inequality we again use the Stratton–Chu formula together with the boundary condition to obtain

$$E^s(x) = -\frac{1}{i\kappa} \nabla \times \nabla \times \int_{\Gamma} u \times H \cdot \Phi(x, y) dA(y)$$

and the asymptotic estimate (9.25) shows that (13.50)

$$\begin{aligned} E^s(x) = & i\kappa \int_{\Gamma} \frac{\exp(i\kappa|x|)}{4\pi|x|} \left\{ \exp(-i\kappa\hat{x} \cdot y) \hat{x} \times ((u \times H) \times \hat{x}) \right. \\ & \left. + O\left(\frac{|u \times H|}{\kappa}\right) \right\} dA(y). \end{aligned}$$

We see that the first term above is just the far field pattern. Hence, using the notation of (13.47), we have

$$\exp(-i\kappa|x|) E^s(x) = \frac{a_1^0(\hat{x})}{\rho} + O\left(\frac{1}{\rho^2}\right),$$

where  $a_1^0(\hat{x})$  is the far field pattern and is thus tangential to the unit sphere (*i.e.*  $\hat{x} \cdot a_1^0(\hat{x}) = 0$ ). Hence, a direct calculation in spherical polar coordinates shows that

$$\left( \nabla \times \left( \frac{a_1^0(\hat{x})}{\rho} \right) \right) \times \hat{x} = 0$$

and so

$$|(\nabla \times u) \times \hat{x}| = O\left(\frac{1}{\rho^3}\right) \text{ as } \rho \rightarrow \infty.$$

The norm estimate follows by an examination of the remainder term in (13.50) via the regularized Stratton–Chu formula (12.8).  $\square$

Now we can prove our main theorem. As usual, a fundamental tool is a suitable adjoint problem. Let  $\varphi \in (C_0^\infty(\mathbb{R}^3 \setminus \bar{D}))^3$  and extend  $\varphi$  by zero to  $\mathbb{R}^3$ . Then we define  $\tilde{z} \in H_{\text{loc}}(\text{curl}; \mathbb{R}^3 \setminus \bar{D})$  to be the weak solution of the adjoint problem (13.51)

$$\begin{aligned} \nabla \times \nabla \times z - \kappa^2 z &= \varphi \text{ in } \mathbb{R}^3 \setminus \bar{D}, \\ u \times z &= 0 \text{ on } \Gamma, \end{aligned} \tag{13.52}$$

$$\lim_{\rho \rightarrow \infty} \rho(\nabla \times z \times \hat{x} + i \kappa z) = 0. \quad (13.53)$$

Using the methods of Chapter 12, it is easy to see that this problem has a unique solution. Furthermore, there is a constant  $C$  such that

$$\|z\|_{H(\text{curl}; \Omega_{R_0})} \leq C \|\varphi\|_{(L^2(B))^3}.$$

**Proof of Theorem 13.15** The proof follows the proof of Theorem 3.1 of [145]. Let  $e = E - E_k$ . Then, using Lemma 3.4, we have

$$\|e\|_{(L^2(B))^3} = \sup_{\varphi \in (C_0^\infty(B))^3} \frac{|(e, \varphi)|}{\|\varphi\|_{(L^2(B))^3}}$$

where  $(e, \varphi) = \int_{\Omega_R} e \cdot \varphi dV$ . Hence, using (13.51)(13.54)

$$\|e\|_{(L^2(B))^3} = \sup_{\varphi \in (C_0^\infty(B))^3} \frac{(e, \nabla \times \nabla \times z - \kappa^2 z)}{\|\varphi\|_{(L^2(B))^3}}.$$

Using integration by parts (3.51), we have

$$(e, \nabla \times \nabla \times z - \kappa^2 z) = (\nabla \times e, \nabla \times z) - \kappa^2 (e, z) + \langle e, u \times \nabla \times z \rangle,$$

where we have used the boundary condition that  $e \times v = 0$  on  $\Gamma$ . Expanding the above expression gives

$$\begin{aligned} (e, \nabla \times \nabla \times z - \kappa^2 z) &= (\nabla \times e, \nabla \times z) - \kappa^2 (e, z) \\ &\quad - i \kappa \langle e_T, z_T \rangle + \langle e_T, u \times \nabla \times z - i \kappa z_T \rangle. \end{aligned}$$

Integrating by parts once more, and using the fact that  $e$  satisfies the homogeneous Maxwell's equations in  $\Omega_k$ ,

$$(e, \nabla \times \nabla \times z - \kappa^2 z) = \langle (\nabla \times e) \times u - i \kappa e_T, z_T \rangle - \langle e_T, (\nabla \times z) \times u + i \kappa z_T \rangle.$$

Using (13.54), this implies that (13.55)

$$\begin{aligned} \|e\|_{(L^2(B))^3} &\leq \left\| (\nabla \times e) \times u - i \kappa e_T \right\|_{L_t^2(\Sigma_R)} \|z_T\|_{L_t^2(\Sigma_R)} \\ &\quad + \|e_T\|_{L_t^2(\Sigma_R)} \|(\nabla \times z) \times u + i \kappa z_T\|_{L_t^2(\Sigma_R)}. \end{aligned}$$

The first term on the left-hand side is estimated by Lemma 13.17, parts (13.48) and (13.49). We see that  $\hat{z}$  satisfies a standard scattering problem, so the estimates of Lemma 13.17 also apply to  $\hat{z}$ . Hence,

$$\|z_T\|_{L_t^2(\Sigma)} \leq C \|z\|_{H(\text{curl}; \Omega_{R_0})} \leq C \|\varphi\|_{(L^2(B))^3}$$

and

$$\begin{aligned} \|(\nabla \times e) \times u - i \cdot e_T\|_{L_t^2(\Sigma)} &= \|(\nabla \times E^s) \times u - i \cdot E_T^s\|_{L_t^2(\Sigma)} \\ &\leq C R^{-2} \|E^s\|_{H(\text{curl}; \Omega_{R_0})}. \end{aligned}$$

Similarly, the second term on the left-hand side of (13.55) can be estimated. Using these estimates in (13.55) proves the theorem.  $\square$

An obvious question which has received considerable attention in the mathematics and engineering literature is whether it is possible to obtain better absorbing boundary conditions than the Silver–Müller condition on  $\Sigma_R$ . In particular, boundary operators  $B$  are sought such that

$$B(E^s) = O\left(\frac{1}{R^\alpha}\right) \text{ for } \alpha > 3.$$

Using the Wilcox expansion (13.47), Webb and Kanellopoulos [291] have derived a family of operators of increasing order. For their family, denoted by  $B_N$ ,  $N = 1, 2, \dots$ , the following estimate holds

$$B_N(E^s) = O\left(\frac{1}{R^{2N+1}}\right), \quad N = 1, 2, \dots$$

The lowest-order operator is exactly the Silver–Müller condition. Higher-order operators require extra regularity for the finite element space on  $\Sigma_R$ . This can be achieved for some edge spaces by modifying the degrees of freedom on  $\Sigma$ , but seems difficult in general. Nevertheless, a computational test of this boundary condition [180, 75] shows that it can provide considerable improvement.

An alternative approach, first proposed by Mur [229], is to recall that each component of  $E^s$  satisfies the Helmholtz equation in free space. Furthermore, from the Stratton–Chu formula (9.15) we see that each component also satisfies the radiation condition appropriate for the Helmholtz equation. Hence, it is possible to use absorbing boundary conditions appropriate for the Helmholtz equation on each component of the scattered field. This is difficult to implement for general edge finite element spaces.

Rather than pursue these approaches here we instead point out that the method used in Chapter 10 gives a natural way to implement progressively higher-order absorbing boundary conditions. Truncating the series expansion (9.71) to  $N$  terms results in a boundary operator that annihilates the first  $N$  terms of the Fourier series for  $E^s$ . Using this fact it should be possible to derive an estimate like the one in estimate (13.49) of Lemma 13.17. Hence we might prove a convergence result for fixed  $N$  as  $R \rightarrow \infty$ . This is essentially the approach of Grote and Keller [152, 154]. They construct absorbing boundary conditions of arbitrary order for the time domain problem by a careful analysis of the series expansion (9.71).

The major criticism of the approach of Webb and Kanellopoulos [291], Grote and Keller [154] or our suggestion in the previous paragraph is that the auxiliary boundary  $\Sigma$  must be spherical. Of course in the lowest-order case, we can apply the Silver–Müller absorbing boundary condition (1.22) on non-spherical boundaries by interpreting  $\hat{x}$  as the unit outward normal to  $\Sigma$ . Typical auxiliary boundaries in practice include parallelepipeds and ellipsoids. Again if the auxiliary boundary is sufficiently far from the scatterer, we can expect reasonable accuracy (although not necessarily a convergence rate of  $R^{-2}$  as guaranteed in

Theorem 13.15). Modifications to include curvature information are also possible (see, e.g., [156]).

### 13.5.2 Infinite element method

Another approach to infinite domain problems deserves mention. This is related to the Calderon operator approach in Chapter 10. The idea is to use infinite elements to fill all of  $\mathbb{R}^3 \setminus D^-$  [65]. Cecot *et al.* [72] suggest basing this approach on the variational approach to scattering problems due to Leis [207] (see also [236, 121]). We now present the method of Leis and then describe an infinite element scheme based on this variational formulation.

The variational formulation is easiest when applied to the scattered field. Recall that the total field and scattered field are related by

$$E = E^i + E^s \text{ in } \mathbb{R}^3 \setminus D.$$

Let  $\chi$  denote a cutoff function such that  $\chi = 1$  on  $\Gamma$  and  $\chi = 0$  outside a ball  $B_{R_0}$  of radius  $R_0$  such that  $D^- \subset B_{R_0}$ . For convenience, we shall assume the origin is in  $D$  and hence not in  $\mathbb{R}^3 \setminus D^-$ .

Now let

$$\tilde{E}^s = E^s + \chi E^i.$$

Obviously,(13.56)

$$\nabla \times \nabla \times \tilde{E}^s - \kappa^2 \tilde{E}^s = \nabla \times \nabla \times (\chi E^i) - \kappa^2 (\chi E^i) \text{ in } \mathbb{R}^3 \setminus \bar{D}$$

and we define

$$F = \nabla \times \nabla \times (\chi E^i) - \kappa^2 (\chi E^i).$$

Note that  $F$  is of compact support and obviously in  $(L^2(\mathbb{R}^3 \setminus D^-))^3$  (in fact, it is in  $\varphi \in (C_0^\infty(\mathbb{R}^3 \setminus \bar{D}))^3$ ). Thus, there is an  $R_0 > 0$  such that  $D^- \subset B_{R_0}$  and  $F = 0$  in  $\mathbb{R}^3 \setminus B_{R_0}$ .

On the boundary of  $D$ ,  $\chi = 1$  and so(13.57)

$$\tilde{E}^s \times u = (E^s + \chi E^i) \times u = 0 \text{ on } \Gamma$$

and at infinity  $\tilde{E}^s = E^s$ , so(13.58)

$$\lim_{\rho \rightarrow \infty} \rho (\nabla \times \tilde{E}^s \times \hat{x} - i \kappa \tilde{E}^s) = 0.$$

To motivate the choice of space for  $\tilde{E}^s$ , note that the Wilcox expansion (13.47) implies that  $\tilde{E}^s = O(1/\rho)$  for large  $\rho$  (and the same estimate holds for  $\nabla \times \tilde{E}^s$ ). Thus, we know that  $\tilde{E}^s$  is contained in the space(13.59)

$$X_L = \left\{ u \in H_{loc}(\text{curl}; \mathbb{R}^3 \setminus \bar{D}) \mid \rho^{-2}(|u|^2 + |\nabla \times u|^2) \in L^1(\mathbb{R}^3 \setminus \bar{D}), \right. \\ \left. v \times u = 0 \text{ on } \Gamma \text{ and } \nabla \times u \times \hat{x} - i \kappa u \in L^2(\mathbb{R}^3 \setminus \bar{D}) \right\}.$$

Note that by (13.49) this space contains  $\tilde{E}^s$ . Furthermore, this space is a Hilbert space when equipped with the norm

$$\begin{aligned}\|u\|_{X_L}^2 &= \left\|\rho^{-1}u\right\|_{L^2(\mathbb{R}^3 \setminus \bar{D})}^2 + \left\|\rho^{-1}\nabla \times u\right\|_{L^2(\mathbb{R}^3 \setminus \bar{D})}^2 \\ &\quad + \left\|\nabla \times u \times \hat{x} - i\kappa u\right\|_{L^2(\mathbb{R}^3 \setminus \bar{D})}^2.\end{aligned}$$

The space includes the radiation condition in its definition in an integral sense.

Now we wish to provide a variational characterization of  $\tilde{E}^s \in X_L$ . Multiplying (13.56) by a test function  $\varphi \in C_0^\infty(\mathbb{R}^3 \setminus \bar{D})$  and integrating by parts, we arrive at the equation

$$\int_{\mathbb{R}^3 \setminus \bar{D}} \nabla \times \tilde{E}^s \cdot \nabla \times \bar{\varphi} - \kappa^2 \tilde{E}^s \cdot \bar{\varphi} dV = \int_{\mathbb{R}^3 \setminus \bar{D}} F \cdot \bar{\varphi} dV.$$

We note that

$$\begin{aligned}&\left| \int_{\mathbb{R}^3 \setminus \bar{D}} \nabla \times \tilde{E}^s \cdot \nabla \times \bar{\varphi} - \kappa^2 \tilde{E}^s \cdot \bar{\varphi} dV \right| \\ &\leq \|\tilde{E}^s\|_{X_L} \left( \|\rho\varphi\|_{L^2(\mathbb{R}^3 \setminus \bar{D})}^2 + \|\rho \nabla \times \varphi\|_{L^2(\mathbb{R}^3 \setminus \bar{D})}^2 \right)^{1/2}.\end{aligned}$$

Let  $\|\varphi\|_{X_L^*}^2 = \|\rho\varphi\|_{L^2(\mathbb{R}^3 \setminus \bar{D})}^2 + \|\rho \nabla \times \varphi\|_{L^2(\mathbb{R}^3 \setminus \bar{D})}^2$ . Then a suitable space for test functions is

$$X_L^* = \text{closure of } C_0^\infty(\mathbb{R}^3 \setminus \bar{D})^3 \text{ in the norm } \|\cdot\|_{X_L^*}.$$

Thus, a possible variational characterization of  $\tilde{E}^s \in X_L$  is (13.60)

$$\begin{aligned}\int_{\mathbb{R}^3 \setminus \bar{D}} \nabla \times \tilde{E}^s \cdot \nabla \times \bar{\varphi} - \kappa^2 \tilde{E}^s \cdot \bar{\varphi} dV \\ = \int_{\mathbb{R}^3 \setminus \bar{D}} F \cdot \bar{\varphi} dV \text{ for all } \varphi \in X_L^*.\end{aligned}$$

We know that this problem has a variational solution, since we know the original infinite domain problem has a solution (see Chapter 12). All that remains is to show that (13.60) has only one solution.

**Lemma 13.19** Problem (13.60) has at most one solution.

**Proof** The proof mimics the proof of Theorem 4.22 of [207]. Let  $F = 0$ , and let  $\varphi = \chi_R E^s$  where  $\chi_R$  is a cutoff function such that  $\chi = 1$  in  $B_R$  and  $\chi = 0$  outside  $B_{2R}$ . In proving this theorem we use the fact that  $\tilde{E}^s$  satisfies the weak homogeneous Maxwell system outside  $B_{R_0}$  and hence  $\tilde{E}^s$  is a classical solution

(this follows from the Stratton–Chu formula, see [94]). Thus, the necessary integration by parts and trace results hold. In particular, considering (13.60) with  $F = 0$  and using the integration by parts result in (3.51), we see that

$$\begin{aligned} 0 &= \int_{\mathbb{R}^3 \setminus \bar{D}} \nabla \times \tilde{\mathbf{E}}^s \cdot \nabla \times (\chi_R \bar{\tilde{\mathbf{E}}^s}) - \kappa^2 \tilde{\mathbf{E}}^s \cdot (\chi_R \bar{\tilde{\mathbf{E}}^s}) dV \\ &= \int_{B_R \setminus \bar{D}} |\nabla \times \tilde{\mathbf{E}}^s|^2 - \kappa^2 |\tilde{\mathbf{E}}^s|^2 dV \\ &\quad + \int_{\mathbb{R}^3 \setminus \bar{B}_R} \nabla \times \tilde{\mathbf{E}}^s \cdot \nabla \times (\chi_R \bar{\tilde{\mathbf{E}}^s}) - \kappa^2 \chi_R |\tilde{\mathbf{E}}^s|^2 dV \\ &= \int_{B_R \setminus \bar{D}} |\nabla \times \tilde{\mathbf{E}}^s|^2 - \kappa^2 |\tilde{\mathbf{E}}^s|^2 dV \\ &\quad + \int_{\mathbb{R}^3 \setminus \bar{B}_R} \chi_R (\nabla \times \nabla \times \tilde{\mathbf{E}}^s - \kappa^2 \tilde{\mathbf{E}}^s) \cdot \bar{\tilde{\mathbf{E}}^s} dV \\ &\quad - \int_{\partial B_R} \nabla \times \tilde{\mathbf{E}}^s \cdot \hat{\mathbf{x}} \times \bar{\tilde{\mathbf{E}}^s} dA. \end{aligned}$$

Thus, taking imaginary parts

$$0 = \Im \left( \int_{\partial B_R} \nabla \times \tilde{\mathbf{E}}^s \cdot (\hat{\mathbf{x}} \times \bar{\tilde{\mathbf{E}}^s}) dA \right) \text{ for } R > R_0$$

where  $B_{R_0}$  contains  $D^-$ . Hence Rellich's Lemma 9.28 and the unique continuation principle (Theorem 4.13) imply that have  $\tilde{\mathbf{E}}^s = 0$ .  $\square$

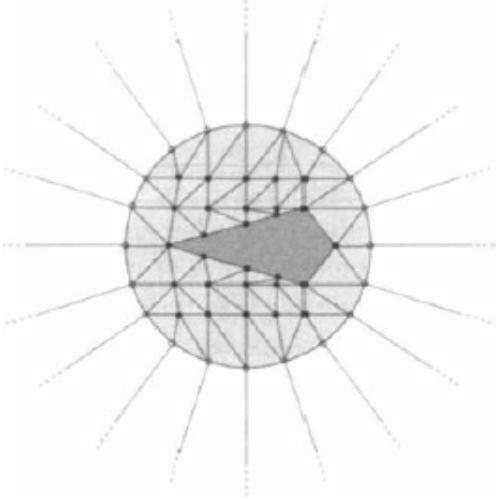
Now that we have a good variational formulation for the exterior Maxwell system, we can discretize it using suitable elements. Cecot *et al.* [72] suggest constructing the elements using mapped hexahedral  $hp$  elements. We shall suggest the analogue for  $b$ -version tetrahedral elements in the following way.

We fix a radius  $R > R_0$  and divide  $\mathbb{R}^3 \setminus D^-$  into  $\mathbb{R}^3 \setminus B_R$  and  $B_R \setminus D^-$ . We mesh  $B_R \setminus D^-$  in the normal way using a tetrahedral mesh except near  $\partial B_R$  where we use mapped tetrahedra (see also Section 8.3). The mesh on  $B_R \setminus D^-$  induces a curvilinear triangular mesh on  $\partial B_R$ . The domain  $\mathbb{R}^3 \setminus B_R$  is then meshed by infinite radial prismatic elements by extending mesh lines radially outward from each node on  $\partial B_R$  (see Fig. 13.9 for a two-dimensional example). On elements in  $B_R \setminus D^-$ , we can use standard edge elements, with the appropriate mapping on curvilinear tetrahedra. On an infinite prismatic element  $K$ , we obtain the infinite element basic functions via mapping. In particular let

$$\widehat{\mathcal{K}}_\infty = \left\{ (\xi, \eta, \rho) \in \mathbb{R}^3 \mid (\xi, \eta) \in \hat{T} \text{ and } R < \rho < \infty \right\}.$$

Here  $\hat{T}$  is the standard two-dimensional triangular reference element with vertices  $(0, 0)$ ,  $(0, 1)$  and  $(1, 0)$ . Then  $K$  can be obtained by a map of the form, for  $\hat{\mathbf{x}} = (\xi, \eta, \rho)^\top \in \mathcal{K}_\infty$ ,

Fig. 13.9. A sketch of two-dimensional infinite radial elements. The scatterer occupies the darker region containing the origin. In the lighter shaded region standard finite elements are used (with curved boundaries near the circle). Outside the circle are infinite radial elements (their infinite extent is denoted by a dashed line).



(13.61)

$$F(\hat{x}) = \frac{\rho}{R} \cdot F_T(\xi, \eta),$$

where  $F_T(\xi, \eta)$  is a parametric representation of  $\partial B_R$  mapping  $T$  onto the base of the prism on  $\partial B_R$  (see Section 8.3 for a discussion of how the triangulation on  $\partial B_R$  is obtained). Now using the usual curl conforming transformation, we can obtain finite element functions on  $K$  from functions defined on the reference element  $K$ .

We shall now construct a subspace  $X_{L_b} \subset X_L$ . Let us assume for completeness that we are using mapped linear ( $k = 1$ ) curvilinear elements as in Section 8.3. Then the basis functions on an infinite element  $K$  are obtained by mapping the following functions on  $K_\infty$  to an infinite element  $K_\infty$  using (13.61):

$$\hat{u}(\xi, \eta, \rho) = \sum_{n=0}^N \sum_{m=1}^3 \hat{a}_{nm} \chi_n(\rho) \hat{\varphi}_m(\xi, \eta) + \sum_{n=0}^N \sum_{m=1}^3 \hat{b}_{n,m} \alpha_n(\rho) \hat{\lambda}_m(\xi, \eta) \hat{e}_\rho,$$

where  $\hat{e}_\rho$  is a unit radial vector in the direction of increasing  $\rho \geq R$  and

$$\begin{aligned} \chi_n(\rho) &= \begin{cases} \left(\frac{R}{\rho}\right) \exp(iK(\rho - R)) & \text{for } n = 0, \\ \left(\left(\frac{R}{\rho}\right)^{n+1} - \left(\frac{R}{\rho}\right)\right) \exp(iK(\rho - R)) & \text{for } n > 0, \end{cases} \\ \alpha_n(\rho) &= \left(\frac{R}{\rho}\right)^{n+2} \exp(iK(\rho - a)) \quad \text{for } n \geq 0. \end{aligned}$$

Here the vector functions  $\hat{\varphi}_m$  are the tangential basis functions obtained for the two-dimensional curl conforming edge finite element space. These functions are defined on the reference triangle and are just the tangential trace of edge finite element functions defined on the three-dimensional reference element (see Section 5.8). In addition, the functions  $\hat{\lambda}_m$  are the usual barycentric coordinate functions for piecewise linear scalar finite elements in two dimensions.

Some comments are needed on the choice of basis functions which follows the philosophy of [72]. The lowest-order  $O(1/\rho)$  term in the expansion is a tangential vector field. This is suggested by the expansion in (13.47), where  $a_0^0$  is the far field pattern and hence a tangential vector field. This ensures (see (13.49)) that the resulting expansion satisfies the Silver–Müller radiation condition. The choice of functions for  $\chi_n(\rho)$ ,  $n > 0$ , is motivated by the need to provide continuity of the tangential component of the discrete field across  $\partial B_R$ .

A corresponding basis for a subspace  $X_L^*$  is also needed. We denote this subspace by  $X_{L,h}^*$ . Again following Cecot *et al.* [72], we choose the elements in  $B_R \setminus D^-$  to be standard edge finite element functions (in our discussion  $k = 1$  edge elements). On an infinite element  $K$ , the basic functions are obtained using mapping from the functions (of course, using curl conforming mapping)

$$\hat{\varphi}(\xi, \eta, \rho) = \sum_{n=0}^N \sum_{m=1}^3 A_{n,m} \varphi_n(\rho) \hat{\varphi}_m(\xi, \eta) + \sum_{n=0}^N \sum_{m=1}^3 B_{n,m} \beta_n(\rho) \hat{\lambda}_m(\xi, \eta) e_\rho.$$

Here

$$\varphi_n(\rho) = \begin{cases} \left(\frac{R}{\rho}\right)^3 \exp(i \kappa (\rho - R)) & n = 0, \\ \left(\left(\frac{R}{\rho}\right)^{n+3} - \left(\frac{R}{\rho}\right)^3\right) \exp(i \kappa (\rho - R)) & n > 0, \end{cases}$$

$$\beta_n(\rho) = (\rho \varphi_n)'.$$

This choice of functions satisfies the decay conditions for functions in  $X_L^*$  (functions in  $X_{L,h}^*$  differ from those in  $X_L$  by a factor  $\rho^3$ ) and is chosen so that the transformed gradient of a scalar function given on  $K$  by

$$\hat{p} = \sum_{n=0}^N \sum_{m=1}^3 C_{nm} \rho \varphi_n(\rho) \hat{\lambda}_m(\xi, \eta)$$

is contained in  $X_{L,h}^*$ . This allows a large set of gradient test functions in the discrete analogue of (13.60), namely to find  $\tilde{E}_h \in X_{L,h}$  such that

$$\int_{\mathbb{R}^3 \setminus \bar{D}} \nabla \times \tilde{E}_h^s \cdot \nabla \times \overline{\varphi_h} - \kappa^2 \tilde{E}_h^s \cdot \overline{\varphi_h} dV = \int_{\mathbb{R}^3 \setminus \bar{D}} F \cdot \overline{\varphi_h} dV \text{ for all } \varphi_h \in X_{L,h}^*.$$

This is not a standard Galerkin method since  $X_{L,h}^* \neq \tilde{X}_h$ . An error analysis has yet to be performed, but computational results given in [72] show good

performance for the  $b\beta$  method on practical problems. The tetrahedral  $b$ -version we have outlined here has not been tried in practice yet!

Let us remark that the basic idea behind the infinite element method is to use a truncated Wilcox expansion

$$\tilde{E}^s \Big|_{\mathbb{R}^3 \setminus \tilde{B}_R} \equiv \sum_{n=1}^N a_n^0(\hat{x}) \left( \frac{1}{\rho} \right)^n \quad \text{with } a_1^0 \cdot \hat{x} = 0.$$

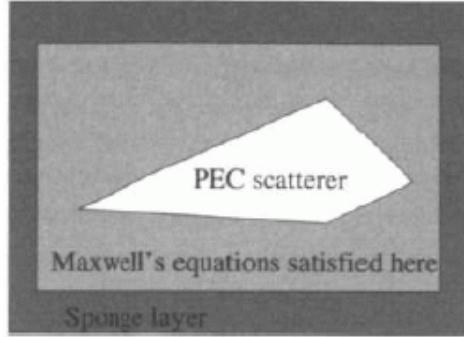
The functions  $a_n^0(\hat{x})$  are then discretized by finite elements that depend on  $\hat{x}$ . Thus, the infinite element method is similar to the method from Chapter 10 which consists in using an expansion of  $E^s$  in terms of the spherical Bessel functions. However, the infinite element method computes the coefficients  $a_n^0(\hat{x})$  rather than using a spherical harmonic expansion. Hence, an indication of the likely error behavior as a function of  $b$  and  $N$  can be found in Corollary 11.19. Indeed, this is roughly the viewpoint of Demkowicz and Pal [121].

One potential advantage of the infinite element is that it can be used with more general coordinate systems than simple spherical coordinates. For example if ellipsoidal coordinates are used, we can take the outer boundary of the finite element region to be the surface of an ellipsoid. The requirement from the point of view of theory is to prove a Wilcox-like expansion in the ellipsoidal coordinate system and adequately characterizes the first few terms in the expansion so that the radiation condition is satisfied. The infinite element scheme is then applicable without the use of ellipsoidal special functions [66].

### 13.5.3 The perfectly matched layer

As we have seen above and in Chapters 10 - 12, the truncation of the infinite domain of a scattering problem to obtain a finite computational domain requires some suitable boundary condition be imposed on the artificial boundary. We have, to a greater or lesser degree, examined the use of artificial boundary conditions, the Calderon operator, Lagrange multiplier methods, infinite element methods and the overlapping Schwarz method of Hazard and Lenoir. In this section, we discuss another truncation procedure motivated from considerations of the scattering problem in the time domain. In this method we view the scattered wave as propagating out from the scatterer (the incident field hits the scatterer creating a scattered field). If a convex auxiliary boundary is used, the scattered field must leave the computational domain and, in the time domain, there should be no reflection of the scattered wave at the auxiliary boundary. Thus, the auxiliary boundary should “absorb” the scattered wave. One way to obtain such an absorber is to surround the computational domain by an artificial “sponge” layer in which an artificial conducting medium is placed. Figure 13.10 shows a schematic of the situation. The problem with this classical approach is that the conducting sponge layer will produce a significant reflected wave at the interface between vacuum and conductor unless the conductivity is small at the interface. The need to limit reflection, while at the same time increasing the conductivity

Fig. 13.10. Geometry of a typical sponge layer. The concave (viewed from the scatterer) sponge layer “absorbs” waves scattered by the perfect conductor.



into the layer to provide sufficient absorption, implies that the classical sponge layer must be thick compared to the wavelength of the radiation and hence inefficient (for a sophisticated version of the sponge layer, see, e.g., [259]).

In 1994 Bérenger suggested a revolutionary approach to sponge layers [36, 37]. He suggested modifying the Maxwell system in the sponge layer so that the resulting Bérenger-Maxwell system has the following properties before discretization:

- (1) exponential absorption of the scattered wave into the sponge layer;
- (2) no reflection at any wavenumber  $\kappa > 0$  and any angle of incidence of the scattered wave at the vacuum/sponge layer interface.

The second property explains the usual name for the Bérenger layer: the *Perfectly Matched Layer* or PML. These properties are carried over approximately to the discrete Bérenger-Maxwell system and provide a very attractive method for implementing a sponge layer. Not surprisingly, this pioneering work has spawned an enormous number of papers investigating various aspects of the layer, modifications and alternative derivations (e.g. [264, 276, 277]). Although originally intended for time domain problems, the layer can be applied in the frequency domain. In order to do this with edge elements it is important that the resulting equations still fit into the curl-curl framework we have used throughout this book.

We shall now provide a formal derivation of the Bérenger-Maxwell system in the form suggested by Zhao and Cangellaris [302]. In this derivation we use the idea that the Bérenger layer can be viewed as a continuation of the solution of the Maxwell system into the upper half complex plane. This view is due to Chew and Wheadon [79] (see also [86]).

The problem we wish to approximate is the model scattering problem of computing the electromagnetic field scattered from a bounded perfectly conducting scatterer in an infinite homogeneous background. Let the scatterer occupy a bounded domain  $D \subset \mathbb{R}^3$  with a Lipschitz polyhedral boundary and connected complement. It is easiest to describe the PML if we work with the scattered

field. The scattered electric field  $E^s$  satisfies the time harmonic Maxwell system in  $\mathbb{R}^3 \setminus D^-$  together with an inhomogeneous perfectly conducting boundary condition on  $\Gamma = \Gamma$ : (13.62)

$$\begin{aligned} \nabla \times \nabla \times E^s - \kappa^2 E^s &= 0 \quad \text{in } \mathbb{R}^3 \setminus \bar{D}, \\ (13.63) \end{aligned}$$

$$\boldsymbol{\nu} \times E^s = -\boldsymbol{\nu} \times E^i \text{ on } \Gamma,$$

where  $\boldsymbol{\nu}$  is the unit outward normal to  $D$  and  $E^i$  is a given incident field. In addition, the scattered field must satisfy the usual Silver–Müller radiation condition.

As we have seen in Theorem 10.8, this problem has a unique solution  $E^s \in H_{loc}(\text{curl}; \Omega)$ . We wish to derive a truncated problem on a bounded domain with the goal that the solution of this modified problem approximates  $E$  well near  $D$ . As a first step we now derive the equations that hold in the sponge layer when the PML is used (see Fig. 13.10).

### 13.5.3.1 The PML in rectilinear coordinates

We start by using the change of variables approach to derive the Maxwell–Bérenger equations in Cartesian coordinates (i.e. the inner boundary of the sponge layer is along coordinate planes as shown in Fig. 13.10). We allow for general coordinate stretching in the  $x_1$ -,  $x_2$ - and  $x_3$ -direction. Let  $\boldsymbol{x} = (x_1, x_2, x_3)^\top$  and define, for  $i = 1, 2, 3$ , (13.64)

$$\tilde{x}_i = \begin{cases} x_i + \frac{i}{\kappa} \int_{a_i}^{x_i} \sigma_i(s) ds & \text{if } x_i \geq a_i, \\ x_i & \text{otherwise.} \end{cases}$$

Here, for each  $i$ ,  $a_i$  is the point at which the PML begins and  $\sigma_i$  is the function that governs the absorption in the PML such that  $\sigma_i(s) > 0$  for  $s > a_i$ . Of course, we also stretch in negative coordinate directions as well to provide a PML in those directions. Other changes of variables could also be considered in place of (13.64). However, the one in (13.64) gives a PML like the original Bérenger layer [86]. For low-frequency work it may be that  $i/\kappa$  should be replaced by  $1/(\beta - ix)$  for some  $\beta > 0$  on the right-hand side above [252].

To obtain the Maxwell–Bérenger equations in the layer, let us suppose that  $E$  can be continued into the upper half of the complex plane in each of  $x_1$ ,  $x_2$  and  $x_3$  as a solution of Maxwell's equations so that (13.65)

$$\widetilde{\nabla} \times \widetilde{\nabla} \times \tilde{E} - \kappa^2 \tilde{E} = 0,$$

where  $\widetilde{\nabla} \times$  denotes the curl with respect to the complex vector variable  $\tilde{\boldsymbol{x}} = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3)^\top$ . We now wish to change variables back to real coordinates using the expressions for  $\tilde{\boldsymbol{x}}$  from (13.64). We define  $d_j(\boldsymbol{x}) = 1 + i\sigma_j(\boldsymbol{x})/\kappa$ ,  $j = 1, 2, 3$ , and obtain by direct calculation that

$$\widetilde{\nabla} \times \widetilde{E} = \begin{pmatrix} \frac{1}{d_2} \frac{\partial \widetilde{E}_3}{\partial x_2} - \frac{1}{d_3} \frac{\partial \widetilde{E}_2}{\partial x_3} \\ - \left( \frac{1}{d_1} \frac{\partial \widetilde{E}_3}{\partial x_1} - \frac{1}{d_3} \frac{\partial \widetilde{E}_1}{\partial x_3} \right) \\ \frac{1}{d_1} \frac{\partial \widetilde{E}_2}{\partial x_1} - \frac{1}{d_2} \frac{\partial \widetilde{E}_1}{\partial x_2} \end{pmatrix}.$$

Rearranging this expression using the fact that  $d_j = d(x_j)$ ,  $j = 1, 2, 3$ , we obtain

$$\widetilde{\nabla} \times \widetilde{E} = \begin{pmatrix} \frac{1}{d_2 d_3} \left( \frac{\partial(d_3 \widetilde{E}_3)}{\partial x_2} - \frac{\partial(d_2 \widetilde{E}_2)}{\partial x_3} \right) \\ - \frac{1}{d_1 d_3} \left( \frac{\partial(d_3 \widetilde{E}_3)}{\partial x_1} - \frac{\partial(d_1 \widetilde{E}_1)}{\partial x_3} \right) \\ \frac{1}{d_1 d_2} \left( \frac{\partial(d_2 \widetilde{E}_2)}{\partial x_1} - \frac{\partial(d_1 \widetilde{E}_1)}{\partial x_2} \right) \end{pmatrix}.$$

Thus, if we define the matrices  $A$  and  $B$  by

$$A = \begin{pmatrix} \frac{1}{d_2 d_3} & 0 & 0 \\ 0 & \frac{1}{d_1 d_3} & 0 \\ 0 & 0 & \frac{1}{d_1 d_2} \end{pmatrix}, \quad B = \begin{pmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{pmatrix},$$

we may write  $\nabla^* \times \tilde{u} = A \nabla^* (B \tilde{u})$ . Returning to (13.65), we can now use the above expression to write the Maxwell–Bérenger equation (13.65) for the field  $\tilde{E}$  in  $\tilde{x}$ -coordinates in terms of  $x$  as follows:

$$A \nabla^* (B \tilde{E}) - \kappa^2 B \tilde{E} = 0.$$

Defining  $\mu_B = \epsilon_B = B^1 A^1$  and defining a new variable  $\hat{E} = B \tilde{E}$ , we conclude that  $\hat{E}$  satisfies (13.66)

$$\nabla^* \times \mu_B^{-1} \nabla^* \hat{E} - \kappa^2 \epsilon_B \hat{E} = 0,$$

where

$$\mu_B = \epsilon_B = \begin{pmatrix} \frac{d_2 d_3}{d_1} & 0 & 0 \\ 0 & \frac{d_1 d_3}{d_2} & 0 \\ 0 & 0 & \frac{d_1 d_2}{d_3} \end{pmatrix}.$$

### 13.5.3.2 The PML in spherical coordinates

We now repeat the derivation of the previous section in spherical coordinates (see Section A.2). The Maxwell–;renger system is again obtained by continuing the standard solution into the upper half plane, this time as a function of radius. To define the continuation variables, let  $R_0 > 0$  be the radius of a sphere containing the scatterer in its interior. We choose  $R > R_0$  and a real valued absorption function  $\sigma = \sigma(\rho)$  parametrized by  $\rho$  and satisfying  $\sigma(\rho) = 0$  for  $\rho < R$  and  $\sigma(\rho) > 0$  for  $\rho > R$ . Then

$$\bar{\rho} = \begin{cases} \rho + \int_R^\rho \frac{i\sigma(s)}{\kappa} ds & \text{if } \rho \geq R, \\ \rho & \text{if } \rho < R. \end{cases}$$

Note that

$$\frac{d\tilde{\rho}}{d\rho} = d = \begin{cases} 1 + \frac{i\sigma(\rho)}{\kappa} & \text{if } \rho \geq R, \\ 1 & \text{if } \rho < R. \end{cases}$$

It will prove convenient to define  $\bar{\sigma}$  and  $D$  by

$$\bar{\sigma} = \begin{cases} \frac{1}{\rho} \int_R^\rho \sigma(s) ds & \text{if } \rho < R, \\ 0 & \text{if } \rho > R, \end{cases} \quad \text{and } D = 1 + \frac{i\bar{\sigma}}{\kappa}.$$

Then  $\bar{\rho}$  and  $\rho$  are related by  $d$  as follows(13.67)

$$\bar{\rho} = \rho \left( 1 + \frac{i\bar{\sigma}}{\kappa} \right) = \rho D.$$

Now suppose that  $\tilde{E}$  satisfies Maxwell's equations (13.65) using spherical polar coordinates  $(\rho, \theta, \varphi)$ . We wish to rewrite (13.65) using standard spherical coordinates. If

$$\tilde{E} = \tilde{E}_\rho e_\rho + \tilde{E}_\theta e_\theta + \tilde{E}_\varphi e_\varphi,$$

we can write

$$\begin{aligned} \nabla \times \tilde{E} &= \frac{1}{\rho \sin \theta} \left( \frac{\partial}{\partial \theta} (\sin \theta \tilde{E}_\varphi) - \frac{\partial}{\partial \varphi} \tilde{E}_\theta \right) e_\rho + \frac{1}{\rho} \left( \frac{1}{\sin \theta} \frac{\partial}{\partial \varphi} \tilde{E}_\rho - \frac{\partial}{\partial \rho} (\tilde{p} \tilde{E}_\varphi) \right) e_\theta \\ &\quad + \frac{1}{\rho} \left( \frac{\partial}{\partial \rho} (\tilde{\rho} \tilde{E}_\theta) - \frac{\partial}{\partial \theta} \tilde{E}_\rho \right) e_\varphi. \end{aligned}$$

Then changing back to the real coordinate  $\rho$  using the fact that  $\bar{\rho} = d\rho$  and  $\partial/\partial\bar{\rho} = (1/d)\partial/\partial\rho$ , we obtain

$$\begin{aligned} \nabla \times \tilde{E} &= \frac{1}{\frac{d^2}{d\rho^2} \rho \sin \theta} \left( \frac{\partial}{\partial \theta} (\sin \theta d \tilde{E}_\varphi) - \frac{\partial}{\partial \varphi} (d \tilde{E}_\theta) \right) e_\rho \\ &\quad + \frac{1}{dd\rho} \left( \frac{1}{\sin \theta} \frac{\partial}{\partial \varphi} (d \tilde{E}_\rho) - \frac{\partial}{\partial \rho} (\rho d \tilde{E}_\varphi) \right) e_\theta \\ &\quad + \frac{1}{dd\rho} \left( \frac{\partial}{\partial \rho} (d \tilde{\rho} \tilde{E}_\theta) - \frac{\partial}{\partial \theta} (d \tilde{E}_\rho) \right) e_\varphi. \end{aligned}$$

We define the operators  $A : \mathbf{R}^3 \rightarrow \mathbf{R}^3$  and  $B : \mathbf{R}^3 \rightarrow \mathbf{R}^3$  so that, if  $\mathbf{v} = v_\rho e_\rho + v_\theta e_\theta + v_\varphi e_\varphi$ , then

$$\begin{aligned} Au &= \frac{1}{\frac{-2}{d}} u_\rho e_\rho + \frac{1}{d\bar{d}} u_\theta e_\theta + \frac{1}{d\bar{d}} u_\varphi e_\varphi, \\ Bu &= du_\rho e_\rho + \bar{d}u_\theta e_\theta + \bar{d}u_\varphi e_\varphi. \end{aligned}$$

The curl relation above can be written as  $\nabla \times \tilde{\mathbf{E}} = A \nabla \times (B\tilde{\mathbf{E}})$ . Using this expression in (13.65), we obtain

$$\nabla \times BA \nabla \times (B\tilde{\mathbf{E}}) - \kappa^2 A^{-1} B^{-1} B\tilde{\mathbf{E}} = 0.$$

Hence, if we define  $\epsilon_B = \mu_B = A^{-1}B^{-1}$  and  $\hat{\mathbf{E}} = B\tilde{\mathbf{E}}$ , we can write the above equations as (13.68)

$$\nabla \times (\mu_B^{-1} \nabla \times \hat{\mathbf{E}}) - \kappa^2 \epsilon_B \hat{\mathbf{E}} = 0,$$

where (13.69)

$$\epsilon_B u = \mu_B u = \frac{-2}{d} u_\rho e_\rho + du_\theta e_\theta + du_\varphi e_\varphi.$$

Note that in spherical coordinates it is easy to see how the Bérenger PML absorbs outgoing waves using the coordinate stretching derivation. From Theorem 9.17, outgoing solutions of (13.65) can be expressed as a series involving the spherical Hankel function  $h_n^{(0)}(\kappa r)$ . This function decays exponentially as the imaginary part of  $\kappa r$  increases (see the asymptotic estimate (9.44)).

### 13.5.3.3 A bounded Bérenger medium

We now return to the problem of approximating the solution  $E$  of (13.62) and (13.63) subject to the Silver–Müller radiation condition. This will be easiest to describe in polar coordinates. Having seen this case, the reader can write down the appropriate problem in the rectilinear case (in addition, the rectilinear case is standard in the literature, whereas the spherical case is less well known).

Let  $B_b$  be a ball of radius  $b > R$  containing  $D$  in its interior and let the computational domain be denoted by  $\Omega = B_b \setminus D^\perp$ . In the sponge layer ( $b > \rho > R$ ), the Maxwell–Bérenger equation (13.68) is satisfied. In the remainder of  $\Omega$ , the standard Maxwell system is satisfied. Define  $\mu_r^\wedge$  and  $\epsilon_r^\wedge$  as follows, using (13.69):

$$\mu_r^\wedge \begin{cases} \mu_B(x) & \text{if } |x| > R, \\ 1 & \text{otherwise,} \end{cases} \quad \epsilon_r^\wedge(x) = \begin{cases} \epsilon_B(x) & \text{if } |x| > R, \\ 1 & \text{otherwise.} \end{cases}$$

Let  $E_b$  denote the solution of the PML truncated problem (both in the PML and standard medium). Then using the above definition, we see that (13.70)

$$\nabla \times \hat{\mu}_r^{-1} \nabla \times E_b - \kappa^2 \epsilon_r E_b = 0 \text{ in } \Omega.$$

On the surface of the scatterer we just require that  $E_b$  satisfy (13.63).

It remains to give a boundary condition on  $\Gamma_b$ . Most authors advocate the use of a simple perfectly conducting boundary condition there [251]. However in [86] we showed that there is an advantage in using a simple absorbing boundary condition (ABC). We now derive a Silver–Müller type ABC suitable for the outer boundary. The standard Silver–Müller ABC is (assuming  $\epsilon_r = \mu_r = 1$  near this boundary)

$$\mathbf{u} \times \nabla \times \mathbf{E} + i\kappa E_T = 0 \quad \text{on } \partial B_b,$$

where  $\mathbf{v}$  is the unit outward normal to  $B_b$  and  $E_T$  denotes the tangential component of  $\mathbf{E}$  on the boundary. If we now apply this equation to  $\tilde{\mathbf{E}}$  using  $\tilde{q}$  as the radial variable we obtain

$$\mathbf{u} \times \widetilde{\nabla} \times \tilde{\mathbf{E}} + i\kappa \tilde{E}_T = 0 \quad \text{on } \partial B_b$$

and changing variables back to the real radius  $q$  we obtain

$$\mathbf{u} \times A \nabla \times B \tilde{\mathbf{E}} + i\kappa B^{-1} B \tilde{E}_T = 0 \quad \text{on } \partial B_b.$$

Now replacing  $B\tilde{\mathbf{E}}$  with  $\mathbf{E}_B$  we have  $\mathbf{v} \times A \nabla \times \mathbf{E}_B + i\kappa B^{-1} E_{B,T} = 0$  on  $\partial B_b$ . But  $B^{-1} E_{B,T}$  involves only tangential components of  $\mathbf{E}$ , so  $\mathbf{v} \times A \nabla \times \mathbf{E}_B + i\kappa D^{-1} E_{B,T} = 0$  on  $\partial B_b$ . Hence, multiplying through by  $D^{-1}$  and using the definition of  $\mu_B$ , we obtain the modified Silver–Müller condition(13.71)

$$\mathbf{u} \times \mu_B^{-1} \nabla \times \mathbf{E}_B + i\kappa E_{B,T} = 0 \quad \text{on } \partial B_b.$$

Now we can derive the weak form of the problem. We multiply eqn (13.70) by a function  $\varphi \in X$  (see (4.3) for the definition of  $X$ ) and proceed as in the derivation of (4.4). This results in the problem of finding  $\mathbf{E}_B \in H(\text{curl}; \Omega)$  such that  $\mathbf{E}_{B,T} \in L^2(\Gamma_b)$  and(13.72a)

$$\mathbf{u} \times \mathbf{E}_B = -\mathbf{u} \times \mathbf{E}^i \quad \text{on } \Gamma, \tag{13.72b}$$

$$\left( \hat{\mu}_r^{-1} \nabla \times \mathbf{E}_B, \nabla \times \varphi \right) - \kappa^2 (\widehat{\epsilon}_r \mathbf{E}_B, \varphi) - i\kappa \langle \mathbf{E}_{B,T}, \varphi_T \rangle = 0 \tag{13.72c}$$

for all  $\varphi \in X$ .

Superficially this problem is very close to the standard cavity problem (4.4), but because the coefficients  $\mu_r^\wedge$  and  $\epsilon_r^\wedge$  are derived from the PML, they do not obey the conditions for our theorems regarding existence and uniqueness of the solution of (4.4). In particular,  $\Im(\epsilon_r^\wedge)$  is not semi-definite. Thus, even the existence of a unique solution of the variational problem (13.72) is not known. For the related problem of time-harmonic acoustic scattering modeled by the Helmholtz equation in curvilinear coordinates, Lassas and Sommersalo [201, 200] have verified existence and uniqueness. In addition, they prove convergence of the continuous Helmholtz–Bérenger solution to the solution of the scattering problem as the total absorption in the layer increases. We do not give their proof here, since it does not pertain directly to the Maxwell equations and is rather technical. A

proof of convergence of the discrete finite element solution to the solution of the scattering problem is an open problem, as is any result about (13.72).

Assuming that the Lassas and Sommersalo result holds for Maxwell's equations, we would expect that  $E_b$  approaches  $E$  in the  $H(\text{curl}; \Omega)$  norm exponentially fast as  $\int_R^\infty \sigma ds$  increases. This suggests that either  $\sigma$  can be taken large or  $b$  large. Taking  $\sigma$  large cuts down the volume of the PML to be meshed, but care must be taken as we shall discuss next. Note that Teixeira *et al.* [277] argue that any PML must be concave as viewed from inside the computational domain.

Problem (13.72) can be approximated by standard edge finite elements as discussed in Chapter 7. Indeed one attraction of the PML approach to truncation is that, once a finite element code is written for general matrix-valued  $\epsilon_r$  and  $\mu_r$ , implementation of the PML just involves a careful definition of these coefficients. Care must be taken, however, in the following ways:

- (1) *Linear system solver* The Maxwell-Bérenger  $\mu_r^+$  and  $\epsilon_r^+$  do not satisfy standard assumptions on the coefficients for Maxwell's equations, and hence may adversely affect matrix conditioning or the performance of the iterative solver used to solve the linear system.
- (2) *Choice of  $\sigma$*  It is critical to choose the layer function  $\sigma$  and the thickness of the layer correctly. In particular, for the discrete problem, there will be some reflection at the layer interface  $Q = R$  with magnitude depending on the size of  $b\sigma(R_+)$ , where  $\sigma(R_+)$  denotes the limiting value of  $\sigma$  at  $R$  from the layer (see [85] for more details in the case of the Helmholtz equation). Thus, it is usual to choose  $\sigma(R_+)$  small (perhaps even zero) and increase  $\sigma$  into the layer. Many functional forms for  $\sigma$  have been suggested (see, e.g., [36, 37]). In [85], we used a numerical approach to predict optimal absorption functions  $\sigma$  for the Helmholtz equation. This approach has been extended to Maxwell's equations by Petropoulis [250].

### 13.5.3.4 Numerical results in two dimensions

The purpose of this section is to show that the Bérenger PML layers can be used to compute near-field solutions of Maxwell's equations in two dimensions. Suppose that the solution of (13.62) is independent of  $x_3$  and that the magnetic field can be written as  $H = (0, 0, H)^\top$ . This implies that  $E = (E_1, E_2, 0)^\top$ . The function  $H$  then satisfies the Helmholtz equation,

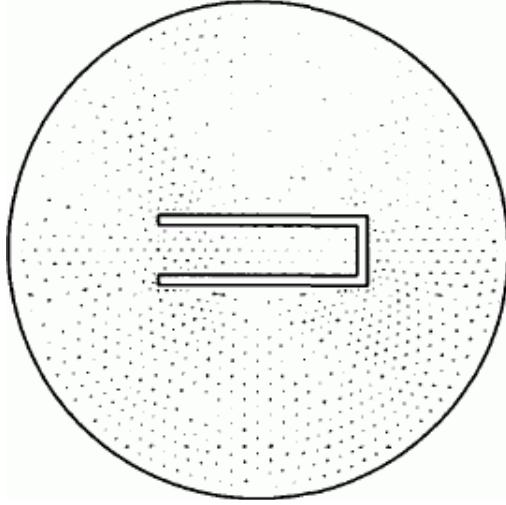
$$\Delta H + \kappa^2 H = 0,$$

outside the scatterer which is now a two-dimensional Lipschitz polygon, still denoted by  $D$ . The perfectly conducting boundary condition reduces to a Neumann boundary condition for  $H$  on  $\Gamma$ . Thus, the scattering problem becomes the problem of computing the weak solution  $H \in H_{\text{loc}}^1(\mathbb{R}^2 \setminus \bar{D})$  of

$$\Delta H + \kappa^2 H = 0 \quad \text{in } \mathbb{R}^2 \setminus \bar{D},$$

$$\frac{\partial H}{\partial u} = -\frac{\partial H^i}{\partial u} \quad \text{on } \Gamma = \partial D,$$

Fig. 13.11. The mesh used in the numerical experiments reported on the time-harmonic wave equation (Helmholtz equation) with PML.<sup>4</sup>



$$\lim_{\rho \rightarrow \infty} \rho^{1/2} \left( \frac{\partial H}{\partial \rho} - i \kappa H \right) = 0,$$

where the last equation is the Sommerfeld radiation condition. We take the incident field to be  $H^i = \exp(i\kappa x_1)$ .

We can now apply the coordinate stretching philosophy to the Helmholtz equation to obtain a PML (see [86]). Using the same notation as the previous section, we see that, within the PML, the Helmholtz–Bérenger solution  $H_B$  satisfies

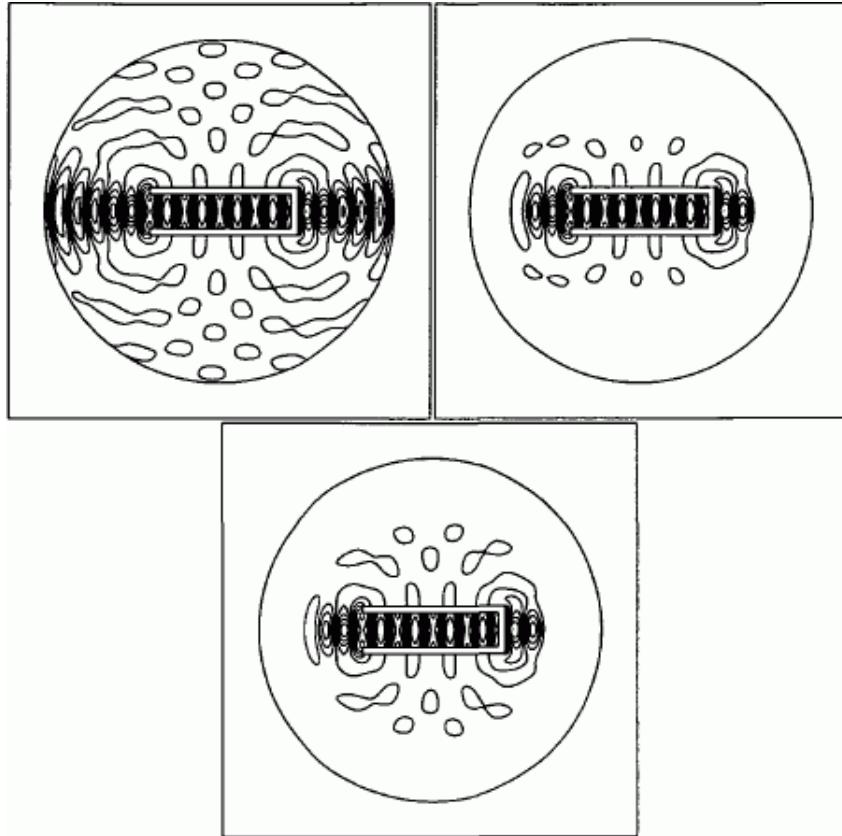
$$\frac{\partial}{\partial x_1} \left( \frac{d_2}{d_1} \frac{\partial H_B}{\partial x_1} \right) + \frac{\partial}{\partial x_2} \left( \frac{d_1}{d_2} \frac{\partial H_B}{\partial x_2} \right) + \kappa^2 d_1 d_2 H_B = 0,$$

which is a generalized Helmholtz equation.

We shall now present examples of the use of the PML in the frequency domain [86]. We consider two cases: the rectilinear PML just described and a cylindrical PML derived by change of variables in the same way. For this test, we compute the field scattered from a perfectly conducting metal obstacle in transverse polarization. As we discussed above, this corresponds to solving the Helmholtz equation with Neumann data imposed on the metal wall. The scatterer is contained in a  $4.2 \times 1.4$  box and  $\kappa = 6.2832$ . The grid used is shown in Fig. 13.11, where the outer radius of the circle is  $\rho = 5$  and the maximum diameter of elements in the mesh is  $h = 0.279$ . Note that the grid is not aligned with either the Cartesian or radial coordinate system. Cubic isoparametric  $H^1$  conforming finite elements are used to discretize the problem [80]. For the Cartesian case we use a simple Neumann boundary condition on the outer boundary. For the

<sup>4</sup> Reprinted from *SIAM J. Sd. Comput.*, 19, The perfectly matched layer in curvilinear coordinates, 2061–2090, F. Collino and P. Monk, Copyright 1998, with permission from SIAM Publications.

Fig. 13.12. Contours of the real part of the scattered field computed using the two PMLs outlined in the text. *Upper left:* the field is computed using a capacitance matrix code that handles the infinite domain exactly. *Upper right:* The real part of the field computed using the Cartesian PML. *Bottom:* The cylindrical PML is used. Near the scatterer the fields are similar, decaying away to zero in both PML layers.<sup>5</sup>



cylindrical PML, we use a modified Sommerfeld radiation condition analogous to (13.71).

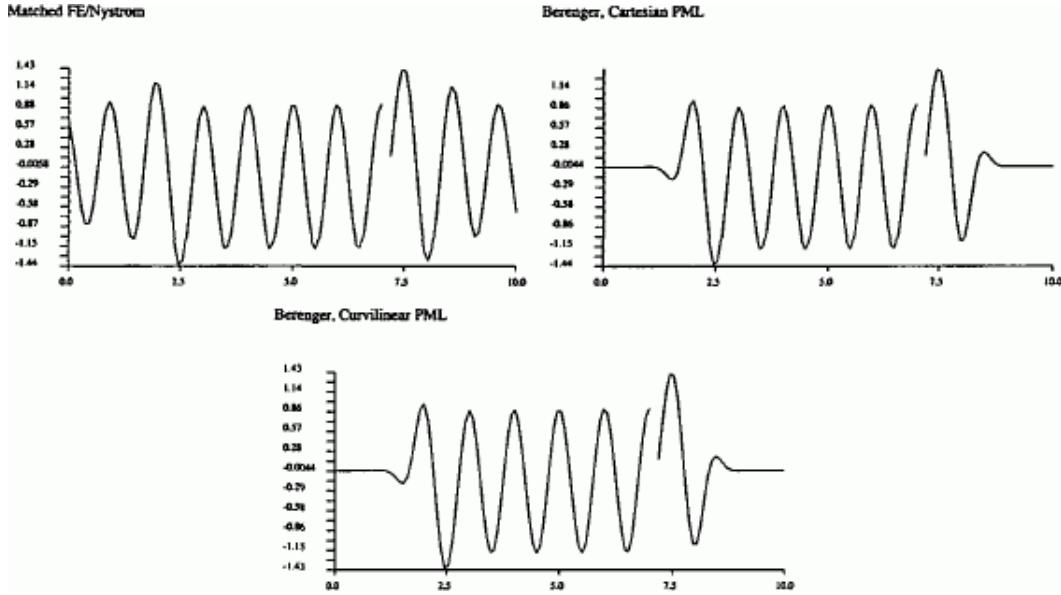
We choose parameters for the Bérenger layer along the lines of those discussed in [37]. In particular, we use a parabolic layer,  $\sigma(s, a) = \sigma_0(|s| - a)^2$  for  $|s| > a$  and  $\sigma = 0$  for  $|s| \leq a$ . In all cases we choose  $\sigma_0 = 5$ . In the Cartesian case, we use

$$d_1(x_1) = 1 + \frac{i}{\kappa} \sigma(x_1 - 5, 2.5) \quad \text{and} \quad d_2(x_2) = 1 + \frac{i}{\kappa} \sigma(x_2, 1.2) .$$

This implies that the layer is 0.5 units from the scatterer. For the cylindrical layer, we use

<sup>5</sup> Reprinted from *SIAM J. Sci. Comput.*, 19, The perfectly matched layer in curvilinear coordinates, 2061–2090, F. Collino and P. Monk, Copyright 1998, with permission from SIAM Publications.

Fig. 13.13. The real part of the scattered field along the line  $x_2 = 0$  for the three solutions in Fig. 13.13. Note that in these plots the point  $x_1 = 0$  corresponds to the left edge of the computational domain. The break in the solution is due to the perfectly conducting scatterer. In the top left panel, we use the capacitance matrix approach. In the top right panel, we use the Cartesian PML. In the bottom panel, we plot the solution for the cylindrical Bérenger medium.<sup>6</sup>



$$d(\rho) = 1 + \frac{i}{\kappa} \sigma(\rho, 2.5).$$

The upper left panel of Fig. 13.12 shows contours of the real part of the scattered field  $H$  computed by a capacitance matrix technique that matches the finite element solution to an integral equation solution outside the grid (thus handling the infinite domain accurately) [189]. In Fig. 13.12, the upper right panel shows the real part of the Helmholtz–Bérenger solution  $H_B$  using the Cartesian layer, and the bottom panel shows the real part of the Helmholtz–Bérenger solution  $H_B$  using the cylindrical layer. Clearly, the near field (e.g. in the mouth of the scatterer) computed using both Bérenger layers is similar to the capacitance matrix solution. But, as expected, the Bérenger solution dies away rapidly in the absorbing layer. Furthermore, in both cases, the contours of the solution show that no abrupt curvature changes when the Bérenger layer is entered which indicate that the layer is “perfectly matched”. In this case the layer is too large, since  $H_B$  has vanished well before the outer boundary.

<sup>6</sup> Reprinted from *SIAM J. Sci. Comput.*, 19, The perfectly matched layer in curvilinear coordinates, 2061–2090, F. Collino and P. Monk, Copyright 1998, with permission from SIAM Publications.

Figure 13.13 shows the real part of the scattered field along the  $x_1$ -axis computed using the different schemes. In the top left panel, we use the capacitance matrix approach. The remainder of Fig. 13.13 shows the corresponding result for the Cartesian and cylindrical Helmholtz–Bérenger media. The solutions are similar for  $|x - 5| < 2.5$ , but the Helmholtz–Bérenger result dies away quickly once the PML is reached. These results give numerical support to the claim that the PML can be used to compute time-harmonic solutions.

## 13.6 Far field recovery

Frequently, the goal of solving Maxwell's equations is to compute the electric far field pattern of the scattered field (a similar problem arises in computing the field at points outside the computational domain). For the methods of Chapters 10 - 12, the truncation scheme allows a direct computation of the far field pattern via the series or integral equation used in the truncation procedure. However, if the field has been computed using an absorbing boundary condition or perfectly matched layer to truncate the domain (as discussed in Sections 13.5 and 13.5.3), we need a procedure to compute the far field pattern from the near field. Here we present the method used in [224] which is an extension of the method in [226] to Maxwell's equations.

To fix ideas, suppose we want to compute an approximation to the far field pattern due to a perfectly conducting scatterer occupying a bounded Lipschitz smooth polyhedral domain  $D \subset \mathbb{R}^3$ . As we have seen in Chapter 1, this implies computing an approximation to the solution  $E$  of(13.73a)

$$\nabla \times \nabla \times E - \kappa^2 E = 0 \quad \text{in } \mathbb{R}^3 \setminus \bar{D},$$

$$E = E^i + E^s \quad \text{in } \mathbb{R}^3 \setminus D, \tag{13.73b}$$

$$\boldsymbol{\nu} \times E = 0 \quad \text{on } \Gamma, \tag{13.73c}$$

$$\lim_{\rho \rightarrow \infty} \rho (\nabla \times E^s \times \hat{x} - i \kappa E^s) = 0, \tag{13.73d}$$

where  $E^i$  is a given incident wave, and the unit outward normal to  $D$  is denoted by  $\boldsymbol{\nu}$ . We then have, according to Theorem 9.5,(13.74)

$$E^s(x) = \frac{\exp(i \kappa |x|)}{|x|} \left\{ E_\infty(\hat{x}) + O\left(\frac{1}{|x|}\right) \right\} \quad \text{as } |x| \rightarrow \infty,$$

where  $\hat{x} = x / |x|$  and(13.75)

$$E_\infty(\hat{x}) = \frac{i \kappa}{4\pi} \hat{x} \times \int_{\Gamma} \left\{ \boldsymbol{\nu}_y \times E(y) + \left( \boldsymbol{\nu}_y \times \frac{\nabla \times E}{i \kappa} \right) \times \hat{x} \right\} \exp(-i \kappa \hat{x} \cdot y) dA(y).$$

In our model problem  $\boldsymbol{\nu} \times E = 0$  on  $\Gamma$  so this term can be dropped from (13.75), but in general for a penetrable scatterer or a surface away from  $\Gamma$  we will need to include this term.

The first step in computing  $E_\infty$  is to truncate the scattering problem. Here we assume the use of the simplest absorbing boundary condition as described in Chapter 1 and Section 13.5. Of course, for methods like those in Chapters 10 - 12 there is no need to use the approach here, since the field in the exterior domain is explicitly modeled. We introduce an artificial boundary  $\Sigma$  far enough from  $\Gamma$  such that  $E^i$  is smooth inside  $\Sigma$ . The solution  $\tilde{E}$  of the truncated problem satisfies

$$\begin{aligned} \nabla \times \nabla \times \tilde{E} - \kappa^2 \tilde{E} &= 0 \quad \text{in } \Omega, \\ \mathbf{u} \times \tilde{E} &= 0 \quad \text{on } \Gamma, \\ \nabla \times \tilde{E} \times \mathbf{u} - i\kappa \tilde{E}_T &= (\nabla \times E^i) \times \mathbf{u} - i\kappa E_T^i \quad \text{on } \Sigma, \end{aligned}$$

where  $\Omega$  is the region exterior to  $\Gamma$  and interior to  $\Sigma$ . We define  $g = \nabla \times E^i \times \mathbf{v} - i\kappa E_T^i$  on  $\Sigma$ .

On  $\Sigma$  the normal  $\mathbf{v}$  is oriented outward to  $\Omega$ . Hence,  $\tilde{E} \in X$  satisfies(13.76)

$$(\nabla \times \tilde{E}, \nabla \times \varphi) - \kappa^2 (\tilde{E}, \varphi) - i\kappa \langle \tilde{E}_T, \varphi_T \rangle = \langle g, \varphi_T \rangle$$

for all  $\varphi \in X$ . Existence and uniqueness for this problem is verified in Chapter 4 . If  $\Sigma$  is far enough from  $\Gamma$ , Theorem 13.15 assures us that we can assume that  $\tilde{E}$  is a good approximation of  $E$  (actually verifying the accuracy for a given computation is not easy!).

In practice, we compute  $E_b \in X_b$  where  $X_b$  is defined in (7.1) such that(13.77)

$$\begin{aligned} (\nabla \times E_b, \nabla \times \varphi_b) - \kappa^2 (E_b, \varphi_b) - i\kappa \langle E_{b,T}, \varphi_{b,T} \rangle &= \langle g, \varphi_{b,T} \rangle \\ \text{forall } \varphi_b \in X_b. \end{aligned}$$

In Section 7.3 we studied the convergence of  $E_b$  to  $\tilde{E}$ . The main problem in using (13.75) with  $E$  replaced by  $E_b$  is that  $\mathbf{v} \times (\nabla \times E_b)$  is not well defined since  $\nabla \times E_b \notin H(\text{curl}; \Omega)$ .

Thus, we must extend (13.75) so that we can apply it to finite element functions. To do this, we recall  $a(\cdot, \cdot)$  defined in (4.5) with the choice  $\epsilon_r = \mu_r = \lambda = 1$  so that

$$a(u, \omega) = (\nabla \times u, \nabla \times \omega) - \kappa^2 (u, \omega) - i\kappa \langle u_T, \omega_T \rangle.$$

Now let  $\mathbf{v} \in H(\text{curl}; \Omega)$  be any function of  $y$  such that,(13.78a)

$$\begin{aligned} \mathbf{u}_T &\in L_t^2(\Sigma), \\ (13.78b) \end{aligned}$$

$$\mathbf{u}_T(y) = ((\hat{\mathbf{x}} \times \mathbf{e}) \times \hat{\mathbf{x}})_T \exp(-i\kappa \hat{\mathbf{x}} \cdot y) \quad \text{on } \Gamma,$$

where  $\mathbf{e}$  is any unit vector. Then(13.79)

$$\mathbf{e} \cdot E_\infty = \frac{1}{4\pi} \int_{\Gamma} (\mathbf{u} \times \nabla \times E) \cdot \mathbf{u}_T dA.$$

For each choice of  $\hat{\mathbf{x}}$  and  $\mathbf{e}$ , there is a different function  $\mathbf{v}$  satisfying (13.78). Taking  $\mathbf{e}$  to be each of the orthogonal unit vectors  $e_1, e_2$  and  $e_3$  in (13.79), we

can compute the three components of  $E_\infty$ . In fact since  $E_\infty \cdot \hat{x} = 0$ , we need only use two tangential unit vectors for each  $\hat{x}$ .

We see that for a fixed  $\hat{x}$ , the desired quantity  $e \cdot E_\infty$  is a functional of the field  $E$ . Thus, we can use the ideas of super-convergent flux recovery [294, 21–23] in the same way as was used in [226] for the Helmholtz equation to approximate  $e \cdot E_\infty$  (see also Chapter 12). To start we use the definition of  $v$  and integration by parts to obtain

$$\begin{aligned} e \cdot E_\infty &= -\frac{1}{4\pi} \int_{\partial\Omega} (\mathbf{v} \times \nabla \times E) \cdot \mathbf{v}_T dA + \frac{1}{4\pi} \sum (\mathbf{v} \times (\nabla \times E)) \cdot \mathbf{v}_T dA \\ &= -\frac{1}{4\pi} \int (\nabla \times \nabla \times E \cdot \mathbf{v} - \nabla \times E \cdot \nabla \times \mathbf{v}) dV \\ &\quad + \frac{1}{4\pi} \sum (\mathbf{v} \times \nabla \times E) \cdot \mathbf{v}_T dA \\ &= \frac{1}{4\pi} \left\{ a(E, \bar{\mathbf{v}}) - \sum (\nabla \times E \times \mathbf{v} - i \kappa E_T) \cdot \mathbf{v}_T dA \right\}, \end{aligned}$$

where  $\mathbf{v}$  is the unit outward normal to  $\Omega$  and we recall that the definition of  $a(\cdot, \cdot)$  includes complex conjugation of its second argument.

Replacing  $E$  by  $\tilde{E}$  in the above equality, we define  $\tilde{E}_\infty$  by(13.80)

$$e \cdot \tilde{E}_\infty = \frac{1}{4\pi} \left\{ a(\tilde{E}, \bar{\mathbf{v}}) - \sum g \cdot \mathbf{v}_T dA \right\}$$

for any unit vector  $e$ . This allows us to use the solution of the truncated problem  $\tilde{E} \in X$  to approximate the far field pattern. A partial justification of this approach is that the far field pattern  $\tilde{E}_\infty$  does not depend on the choice of  $\mathbf{v}$ , provided (13.78) is satisfied. To see this, suppose  $\mathbf{v}^{(1)}$  and  $\mathbf{v}^{(2)}$  satisfy (13.78) and give rise to far field patterns  $\tilde{E}_\infty^{(1)}$  and  $\tilde{E}_\infty^{(2)}$ . Then

$$e \cdot \tilde{E}_\infty^{(1)} - e \cdot \tilde{E}_\infty^{(2)} = \frac{1}{4\pi} \left\{ a(\tilde{E}, \overline{\mathbf{v}^{(1)} - \mathbf{v}^{(2)}}) - \sum g \cdot (\mathbf{v}^{(1)} - \mathbf{v}^{(2)}) dA \right\}.$$

But  $\mathbf{v}^{(1)} - \mathbf{v}^{(2)} \in X$ , so that using (13.76) we have  $e \cdot \tilde{E}_\infty^{(1)} - e \cdot \tilde{E}_\infty^{(2)} = 0$ .

We can now define the approximate far field pattern for  $E_b$  based on (13.80). One possible approach, used in [220], is to simply substitute  $E_b$  for  $\tilde{E}$  in (13.80). However, it is then necessary to integrate over  $\Omega$  which is a fixed volume independent of mesh size to evaluate  $a(E_b, \mathbf{v}_T)$ . This gets very time consuming for small  $b$  [220]. In addition, using (13.80) directly, the discrete far field pattern depends on the choice of  $\mathbf{v}$  which is undesirable.

Again using the ideas of super-convergent flux recovery, we instead use the method of Chapter 12 and define  $\mathbf{v}_b \in V_b$  as follows (recall  $V_b$  is the subspace of  $H(\text{curl}; \Omega)$  used to construct  $X_b$  in Chapter 7):(13.81)

$$(\mathbf{v}_b)_T(y) \text{ interpolates } ((\hat{x} \times e) \times \hat{x})_T \exp(-i \kappa \hat{x} \cdot y) \text{ on } \Gamma.$$

Then we define the approximate far field pattern  $E_{h,\infty}$  by(13.82)

$$e \cdot E_{h,\infty}(\hat{x}) = \frac{1}{4\pi} \left\{ a(E_h, \bar{u}_h) - \sum g \cdot u_{h,T} dA \right\}.$$

This definition is also independent of the choice of  $v_b$  provided  $v_b$  satisfies (13.81). The proof is exactly the same as for the independence of  $\tilde{E}_\infty$  on  $v$  but using (13.77) in place of (13.76).

In practice, we choose  $v_b \in V_b$  to interpolate  $((\hat{x} \times e) \times \hat{x})_T \exp(-i\kappa \hat{x} \cdot y)$  on  $\Gamma$ , and to interpolate zero elsewhere in  $\Omega$ . Thus,  $v_b$  is only non-zero on elements sharing a face or edge with  $\Gamma$ . To evaluate  $a(E_h, \bar{v}_{h,b})$  requires integration only over the elements in this “skin”, and the integrals can be computed precisely since  $v_b$  is a finite element function. On the other hand, in our analysis of the error, we shall choose  $v_b$  as a suitable interpolant of a function defined on  $\Omega$ .

To start our analysis of this procedure, note that

$$e \cdot E_\infty - e \cdot E_{h,\infty} = e \cdot E_\infty - e \cdot \tilde{E}_\infty + e \cdot E_{h,\infty}.$$

Our assumption is that  $\Sigma$  is chosen far enough from  $\Gamma$  to ensure that the error due to the absorbing boundary condition  $e \cdot E_\infty - e \cdot \tilde{E}_\infty$  is sufficiently small. It remains to estimate the term  $e \cdot \tilde{E}_\infty - e \cdot E_{h,\infty}$ . To do that we define two auxiliary functions. First let  $\xi \in H(\text{curl}; \Omega)$  be such that(13.83a)

$$\xi_T \in L_t^2(\Sigma),$$

$$\xi_T = (u - u_h)_T \text{ on } \Gamma, \quad (13.83b)$$

$$a(\varphi, \bar{\xi}) = 0 \text{ for all } \varphi \in X. \quad (13.83c)$$

Then let  $\zeta \in H(\text{curl}; \Omega)$  be such that(13.84a)

$$\zeta_T \in L_t^2(\Sigma),$$

$$\zeta_T = u_T \text{ on } \Gamma,$$

$$a(\varphi, \bar{\zeta}) = 0 \text{ for all } \varphi \in X. \quad (13.84b)$$

In both cases existence of the solution is shown by the use of the Fredholm alternative and unique continuation as in the analysis of (13.76) given in Chapter 4 . Of course,  $\xi$  is the weak solution of

$$\begin{aligned} \nabla \times \nabla \times \xi - \kappa^2 \xi &= 0 && \text{in } \Omega, \\ u \times \xi &= u \times (u - u_h) && \text{on } \Gamma, \\ \nabla \times \xi \times u - i \kappa \xi_T &= 0 && \text{on } \Sigma. \end{aligned}$$

The function  $\zeta$  satisfies the same problem but with  $v_b$  omitted. We then have the following basic error estimate.

**Lemma 13.20** Suppose  $\tilde{E}$  satisfies (13.76) and  $E_b$  satisfies (13.77). If  $\tilde{E}_\infty$  is given by (13.80) and  $E_{h,\infty}$  by (13.82), then

$$e \cdot \tilde{E}_\infty - e \cdot E_{h,\infty} = \frac{1}{4\pi} \left\{ a(\tilde{E} - E_h, \bar{u}_h - \bar{z}) - \sum g \cdot \xi_T dA \right\},$$

where  $\xi$  satisfies (13.83) and  $\zeta$  satisfies (13.84).

**Proof** Let  $v = v_b + \xi$ , then

$$e \cdot \tilde{E}_\infty - e \cdot E_{h,\infty} = \frac{1}{4\pi} \left\{ a(\tilde{E} - E_h, \bar{v}_h) + a(\tilde{E}, \bar{\xi}) - \int \sum g \cdot \xi_T dA \right\}.$$

But, since  $\tilde{E} \in X$ , eqn (13.83) implies

$$e \cdot \tilde{E}_\infty - e \cdot E_{h,\infty} = \frac{1}{4\pi} \left\{ a(\tilde{E} - E_h, \bar{v}_h) - \int \sum g \cdot \xi_T dA \right\}.$$

Since  $E - E_b \in X$ , (13.84) implies that we may subtract  $\zeta$  and we are done.  $\square$

Using the estimate in the previous lemma, we can estimate the error in the computed far field pattern as follows.

**Corollary 13.21** *Under the conditions of the previous lemma, for any unit vector  $e$ ,*

$$|e \cdot \tilde{E}_\infty - e \cdot E_{h,\infty}| \leq C \left\{ \| \tilde{E} - E_h \|_X \| T_h z - z \|_X + \| u \times (r_h u - u) \|_{Y(\Gamma)} \right\},$$

where  $r_b$  is the interpolation operator into the edge finite element space  $V_b$ .

**Remark 13.22** *The regularity of  $\tilde{E}$  and  $\zeta$  is in general the same. Thus, if  $\| E - E_b \|_X = O(h^\delta)$ , then  $\| r_b \zeta - \zeta \|_X = O(h^\delta)$  so that the first term on the right hand side of this estimate converges at twice the rate of convergence of the finite element solution. This is the “superconvergence” referred to in superconvergent flux recovery. Since*

$$\| u \times (T_h u - u) \|_{Y(\Gamma)} \leq C \| u \times (T_h u - u) \|_{H(\text{Div}; \Gamma)}$$

and since  $v$  is constant on each triangle, we have

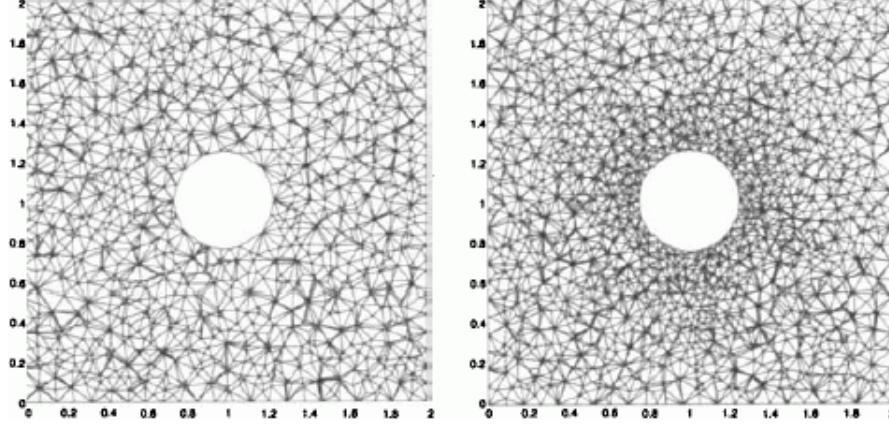
$$\| u \times (T_h u - u) \|_{H(\text{Div}; \Gamma)} \leq Ch^k,$$

where  $k$  is the degree of the edge finite element space used to compute  $E_b$ . Hence, we obtain

$$|e \cdot E_\infty - e \cdot E_{h,\infty}| \leq C(h^{2\delta} + h^k),$$

which suggests only  $O(h^\delta)$  as a maximum rate of convergence. However, the error  $\| v \times (r_b v - v) \|_{H(\text{Div}; \Gamma)}$  can be controlled a priori since  $v$  is a known function on  $\Gamma$ . Hence, the error from this term can be controlled by suitably refining the surface grid on  $\Gamma$ . Moreover, for smooth domains, it may well be possible to use negative norm estimates of the error  $v \times (r_b v - v)$  as was done for the Helmholtz equation in two dimensions in [226]. It is likely that this would provide a better estimate only for quadratic or higher order edge elements and has yet to be done.

Fig. 13.14. A slice through the tetrahedral mesh used in this study indicating the geometric resolution of the two meshes: *left*, mesh 1; *right*, mesh 2. The maximum element diameter is roughly comparable in the two cases, but mesh 2 approximates the boundary of the sphere better.<sup>7</sup>



**Proof of Corollary 13.21** We select  $v_b = r_b \zeta$  and use the continuity of  $a(\cdot, \cdot)$  and the Cauchy–Schwarz inequality to obtain(13.85)

$$\begin{aligned} |e \cdot \tilde{E}_\infty - e \cdot E_{h,\infty}| &\leq C \|\tilde{E} - E_h\|_X \|r_h z - z\|_X \\ &\quad + \|g\|_{L_t^2(\Sigma)} \|\xi_T\|_{L_t^2(\Sigma)}. \end{aligned}$$

Now we need to relate the norm  $\|\xi_T\|_{L_t^2(\Sigma)}$  to boundary data on  $\Gamma$ . Considering (13.83) we see that  $v \times (v - v_b) \in Y(\Gamma)$ . Furthermore, using a cutoff function, we know that there is a function  $\xi_0 \in H(\text{curl}; \Omega)$  such that  $v \times \xi_0 = v \times (v - v_b)$  on  $\Gamma$  and  $\xi_{0,T} = 0$  on  $\Sigma$ . Thus if  $u = \xi - \xi_0 \in X$ , we see that  $u$  satisfies

$$a(\varphi, \bar{u}) = a(\varphi, \bar{\xi}_0) \text{ for all } \varphi \in X.$$

Using the *a priori* estimate obtained by applying the Fredholm alternative to (13.83), we obtain

$$\|\xi\|_X \leq \|u\|_X + \|\xi_0\|_X \leq C \|\xi_0\|_{H(\text{curl}; \Omega)} \leq C \|v \times (v_h - v)\|_{Y(\Gamma)}.$$

Using this estimate in (13.85) completes the proof.  $\square$

Now we present the result of a numerical experiment to test this approach to computing the far field pattern taken from [224]. We use linear first kind edge elements ( $k = 1$ ) on a tetrahedral mesh to approximate backscattering by a sphere. The domain  $D$  is a sphere of radius 0.25 and  $\Sigma$  is taken to be the surface of the cube  $[-1, 1]^3$ . Of course, the sphere is not a Lipschitz polyhedron, but we

<sup>7</sup> Reprinted from *Journal of Computational Physics*, 170, Phase-accuracy comparisons and improved far-field estimates for 3-D edge elements on tetrahedral meshes, 614–641, P. Monk and A.K. Parrot, Copyright 2001, with permission from Elsevier Science.

can expect that a surface triangulation will adequately approximate the sphere to the accuracy of linear elements (this has not been proved).

The incoming wave is a plane wave in the direction  $d = (1, 0, 0)^\top$  with  $p = (0, 1, 0)^\top$ . For an incoming plane wave  $E^i$ , the *radar cross section* in the direction  $\hat{x}$  is denoted by  $\text{RCS}(\hat{x})$  and given by

$$\text{RCS}(\hat{x}) = \frac{4\pi|E_\infty(\hat{x})|^2}{|E^i|^2}.$$

For a sphere of radius  $a$ , we then define the *normalized echo area* by

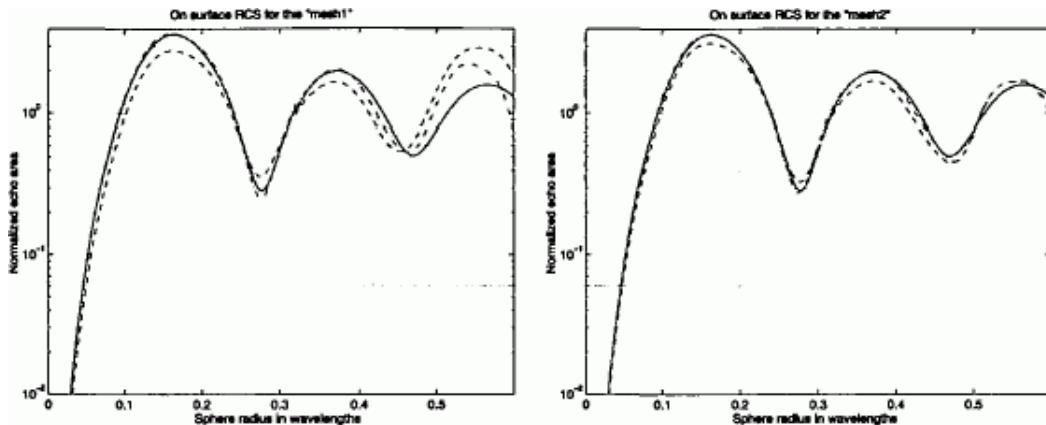
$$\text{NEA} = \frac{1}{\pi a^2} \text{RCS}(-d).$$

Here  $\text{RCS}(-d)$ , the *backscattered* RCS, is measured in the direction from which the plane wave is incident (i.e.  $-d$ ). Using the MIE series from Section 9.5.2, it is possible to compute the exact RCS and hence the NEA for a range of wavenumbers  $\kappa$ .

Using two meshes of tetrahedra with different surface approximate properties (mesh 1 has  $\approx 70000$  elements and mesh 2 has  $\approx 110000$  elements, see Fig. 13.14), we can find the time-harmonic linear ( $k = 1$ ) edge element solution to (13.77). In fact, in the example presented here, the solution is computed using a broadband excitation in the time domain. The time-harmonic solution is then extracted using the Fourier transform (see [224]). Using this field, we can compute an approximation to the NEA either by using our proposed method in (13.82) or using (13.79) directly by simply substituting  $E_b$  for  $E$ .

Figure 13.15 shows the results for mesh 1 and mesh 2. In both cases our proposed method (13.82) improves the accuracy of the NEA. Ultimately for high  $\kappa$  the computed echo area becomes inaccurate most likely due to a breakdown in accuracy of the finite element method as the wavelength of the radiation decreases (see Section 13.3). The improved resolution of the sphere in mesh 2 explains the improved results for that mesh!

Fig. 13.15. Comparison of the computed NEA as a function of sphere radius in wavelengths ( $a/\lambda = \pi a/2\pi$ ). In both panels, the solid line indicates the exact solution computed via the MIE series. The dashed line is computed via (13.79) and the improved dashed dot line is computed via (13.82). On the left is the result for mesh 1, on the right for mesh 2.<sup>8</sup>



<sup>8</sup> Reprinted from *Journal of Computational Physics*, 170, Phase-accuracy comparisons and improved far-field estimates for 3-D edge elements on tetrahedral meshes, 614–641, P. Monk and A.K. Parrot, Copyright 2001, with permission from Elsevier Science.

# 14 INVERSE PROBLEMS

## 14.1 Introduction

So far, in this book we have studied the direct scattering problem of computing the interaction of incident electromagnetic fields with a known target. In this final chapter, we shall discuss the inverse scattering problem of determining the shape of a scatterer from a knowledge of the incident field and the corresponding far field pattern. On the one hand, we shall not use finite element methods except to obtain data for testing the inverse algorithm. On the other hand, much of our previous theoretical work can be used to investigate the inverse algorithm. In addition, my own study of direct scattering has been largely motivated by work on inverse problems.

We shall limit ourselves to the inverse scattering problem for time-harmonic waves in the resonance region. By “resonance region” we mean that the wavelength of the electromagnetic waves ( $2\pi/\lambda$ ) is the same order of magnitude as the size of features we wish to detect (e.g. the diameter of the scatterer). Of course, this is exactly the frequency range we have been discussing in the preceding chapters.

Contrasting the direct and inverse scattering problems, we can summarize the situation as follows:

*Direct Problem* We assume that we know the relevant details of the scatterer such as its shape, electromagnetic parameters  $\epsilon$ ,  $\sigma$  and  $\mu$ , and its position. We then seek to approximate the scattered field due to a given incident field. One example of this type of problem, as we have seen, is to predict the radar cross section of a given aircraft. This involves finding the far field pattern of the scattered field. The direct problem is also sometimes referred to as the “forward problem”.

*Inverse Problem* By contrast, the model inverse scattering problem presented in this chapter assumes that measurements of the far field pattern due to a variety of incoming waves are available (typically at a number of different frequencies — although we shall only consider the single frequency case here). From these scattering data, it is desired to determine information about the scatterer.

As an example of an inverse problem, we could seek to determine the shape of a scatterer or perhaps the presence of undesirable flaws inside the scatterer. The best known inverse problem is most likely the radar problem [52]. In this problem we seek to determine the velocity and position of an object from measurements

of the far field pattern at the same position as the source of the incident field (termed backscattering data). Other applications include non-destructive testing of objects by microwave interrogation [199], microwave medical imaging [99, 100] and mine detection [27]. In these latter applications it is possible, in principle, to gather more than just backscattered data. It may also be desirable to obtain more than just positional information. For example, in the non-destructive testing application it might be desirable to image flaws to determine if they threaten the integrity of the object under investigation.

The inverse scattering problem we shall investigate is a model problem in that we have simplified certain practical aspects. In particular, we have chosen an inverse problem that fits into the basic theoretical framework established in the foregoing chapters. Considerable elaboration of the basic method is possible.

The problem we wish to discuss assumes that a bounded scatterer in a homogeneous, isotropic and infinite background medium is illuminated by plane waves with direction  $d \in \mathbb{R}^3$  and polarization  $p \in \mathbb{C}^3$  given by(14.1)

$$E^i(x) = i\kappa(d \times p) \times d \exp(i\kappa x \cdot d),$$

where  $|d| = 1$ . This rather elaborate version of the plane wave (compared to (1.20)) allows us to use an arbitrary polarization  $p$  and still have  $E^i$  as a solution of the background homogeneous Maxwell system. This incident field reduces to the standard one if  $p$  is perpendicular to  $d$ . The incident field  $E^i$  is scattered by an unknown bounded object, and we assume that the far field pattern of the scattered electric field  $E_\infty(\hat{x}, d, p)$  is known for all observation directions  $\hat{x} \in \partial B_1$ , all directions of incident field  $d \in \partial B_1$  and all polarizations  $p \in \mathbb{C}^3$ . In reality, only measurements of  $E_\infty$  would be available, and so  $E_\infty$  cannot be assumed exactly known. We shall deal with this problem later.

We have assumed the knowledge of a tremendous amount of data. In fact, we shall only use a discrete set of  $\hat{x}$  and  $d$ . In addition,  $E_\infty$  is linear in  $p$  and orthogonal to  $\hat{x}$ , so for each  $d$  two measurements corresponding to linearly independent polarizations orthogonal to  $d$  suffice to determine  $E_\infty$  for all  $p$ . Note that we assume that both the amplitude and phase of the components of  $E_\infty$  are known. From this data, we wish to determine the shape of the unknown scatterer.

There are a variety of methods that can be used to attack this inverse problem. A good survey can be found in Colton [87]. A very flexible option is to use an optimization approach. Let  $S$  denote a scatterer of the type to be reconstructed. For example,  $S$  could denote the surface of a perfect conducting scatterer. Then we denote by  $E_\infty(\hat{x}, d, p, S)$  the far field pattern of the scattered field due to  $S$  when the incident field (14.1) is used. Then we seek an optimal  $S^*$  that gives the best fit to the data by solving, for example,(14.2)

$$\inf_S \int_{\partial B_1} \int_{\partial B_1} \| E_\infty(\cdot, d, p, S) - E_\infty(\cdot, d, p) \|_{L^2(\partial B_1)}^2 dA(d) dA(p).$$

The reader will have noted that my description of  $S$  is rather imprecise, and that this approach begs the question whether  $S^*$  actually exists. The first of these difficulties is a drawback of the optimization approach. We need specific *a priori* information about the unknown scatterer to make the formulation precise. For example, we might assume that we are reconstructing a perfectly conducting scatterer that is starlike with respect to the origin. Then  $S$  denotes the unknown boundary parametrized by a suitable function of angular polar coordinates. If in fact there are two scatterers present, or the boundary of the scatter is imperfectly conducting, this approach might fail.

It is also necessary to establish conditions under which (14.2) has a minimizer  $S^*$ . Often this requires modifying the basic least-squares term in (14.2) to regularize the problem or limit the search for a minimizer to some compact set (see, e.g., [43]).

Numerically, it is necessary to establish that a suitable gradient of the functional in (14.2) with respect to the parameters defining a discrete approximation to  $S$  can be computed in an efficient way (see, e.g., [126]). Then large-scale optimization methods are needed to actually compute an approximation to  $S^*$  (see, e.g., [166] in the acoustic context). Efficient and special purpose optimization algorithms have been proposed in [129, 130] in the context of electromagnetics.

The advantage of this method is that it is rather general. The functional can be modified to incorporate incomplete data (e.g. measurements only on a portion of  $\partial B_i$ ), and even allow for measurements of the magnitude of  $E_\infty$  alone. The general applicability of the method, and the fact that it is possible to reconstruct details of the scatterer (e.g. the function  $\epsilon_r$  if the scatterer is a dielectric and  $\epsilon_r$  is included as an unknown in the optimization scheme) are advantages of this approach. For more details, see, e.g., [130].

Another fruitful approach is to linearize the map from the scattering data  $S$  to the far field pattern  $E_\infty(\cdot; d, p, S)$  about a reference configuration, and then predict  $S^*$  by solving the linearized problem. For example, a linearization valid at low frequency and low contrast ( $\epsilon_r$  close to unity) is given by the Born approximation (see, e.g., [45, 78, 124]). When the Born approximation is applicable, extremely fast algorithms can be constructed (for example [45, 117]). However, if the approximation is applied outside its domain of applicability, poor reconstructions can result. Of course, it is possible to iterate the linearization technique about successively improving estimates of the scatterer to obtain a Newton method for the inverse scattering problem (see, e.g., [254]). This method has a similar complexity to the optimization technique outline in the previous paragraphs.

The linear sampling method (LSM), which is the subject of this chapter, attempts to provide the speed and simplicity of linearization methods without making any special assumptions or approximations. The LSM only reconstructs the boundary of the scatterer, so it does not immediately produce information about the nature of the scatterer such as the conductivity of the material. The LSM requires, at the least, data for  $\hat{x}$  in some open subdomain of the unit

sphere, for  $d$  in some possibly disjoint subdomain and for all  $p$ . But the method has important advantages. First, it only requires the solution of linear ill-posed problems and, second, it does not require *a priori* information about the scatterer such as that needed, for example, to implement (14.2) (it is required to know the approximate position and size of the scatterer).

The LSM has its roots in the study of far field patterns for acoustic waves by Colton and Kirsch [91]. This study resulted in the dual space method of Colton and Monk [97, 98], again for acoustic problems. The dual space method was a variant of the optimization method discussed above. A crucial insight was then provided by Colton and Kirsch [92], who noted that the boundary could be determined without optimization. This idea was extensively developed in the acoustic context (see, e.g., [96, 88]) and more recently in the electromagnetic context in [101, 194, 90, 155, 69].

In parallel to the development of the LSM, Kirsch (see, e.g., [187]) has introduced a family of sampling methods based on various factorizations of the far field operator (we shall define this operator in the next section — see (14.4)). These methods have a more complete mathematical justification than the LSM, but are more limited in applicability (at least the theoretical analysis cannot be established in as great a generality). Sampling methods have also been developed by other authors (see, e.g., [242, 171, 253, 254]).

## 14.2 The linear sampling method

In this section we shall describe in a formal way the LSM and give a heuristic justification for its success. The theorems to support this argument will follow after we have discussed how to implement the method and shown some examples.

Regardless of the physics of the scatterer, provided certain resonances are avoided, the LSM is the same. We introduce an artificial source point  $\zeta \in \mathbb{R}^3$  and artificial polarization  $q \in \mathbb{R}^3$  with  $|q| = 1$ . Then we seek a function  $F = L_t^2(\partial B_1) \rightarrow L_t^2(\partial B_1)$  such that(14.3)

$$\int_{\partial B_1} E_\infty(\hat{x}, d, g(d)) dA(d) = \frac{i\kappa}{4\pi} (\hat{x} \times q) \times \hat{x} \exp(-i\kappa \hat{x} \cdot z)$$

for all  $\hat{x} \in \partial B_1$ . Note that  $E_\infty(\hat{x}, d, p)$  is a linear function of the polarization  $p$  and hence the left hand side of (14.3) defines a linear operator. We shall discuss later how to actually solve (14.3), but for now, motivated by (14.3), merely define the *far field operator*  $E_g^i$  by(14.4)

$$(FG)(\hat{x}) = \int_{\partial B_1} E_\infty(\hat{x}, d, g(d)) dA(d)$$

for  $\hat{x} \in \partial B_1$ . Note that, by superposition,  $F_g$  is the far field pattern for the electric field due to the incoming wave  $E_g^s$  of the form(14.5)

$$E_g^i(x) = i\kappa \int_{\partial B_1} \exp(i\kappa x \cdot d) g(d) dA(d).$$

This is an example of an electric *Herglotz wave function* with kernel  $i\zeta g$ . We shall study such functions more in Section 14.3.2. Let us denote the corresponding scattered field by  $E_g^S$  so that  $Fg = E_{g,\infty}^S$ .

Turning now to the right-hand side of (14.3), we see that this is the far field pattern of an electric dipole at  $\zeta$  with polarization  $q$  so that if we define(14.6)

$$E_{e,\infty}(\hat{x}, z, q) = \frac{i\kappa}{4\pi} (\hat{x} \times q) \times \hat{x} \exp(-i\kappa \hat{x} \cdot z)$$

then  $E_{e,\infty}$  is the far field pattern of  $E_e$  given by

$$E_e(x) = \frac{-1}{i\kappa} \nabla \times \nabla \times (q\Phi(x, z)),$$

where  $\Phi$  is the fundamental solution to the Helmholtz equation given by (9.1). We can thus rewrite (14.3) as(14.7)

$$E_{g,\infty}^S(\hat{x}) = E_{e,\infty}(\hat{x}, z, q) \quad \text{forall } \hat{x} \in \partial B_1.$$

Now let us assume that  $\zeta \in D$  where  $D$  denotes the unknown support of the scatterer (i.e. we want to reconstruct  $\Gamma = \partial D$ ). Suppose also that (14.3) is solvable so that (14.7) holds. Since the two far field patterns agree, Corollary 9.29 of Rellich's lemma and the unique continuation result in Theorem 4.13 show that

$$E_g^S(x) = E_e(x, z, q) \quad \text{for } x \in \mathbb{R}^3 \setminus D.$$

Now let  $x \in \Gamma$  and let  $\zeta \in D$  approach  $x$ . Due to the singularity in  $\Phi$  at  $x = \zeta$  the norm of the right-hand side blows up in the  $H_{loc}(\text{curl}, \mathbb{R}^3 \setminus D)$ . On the left-hand side, we have a weighted integral of scattered fields, and due to our assumption that  $\partial D$  is a Lipschitz polyhedron, the fields are in  $H_{loc}(\text{curl}, \mathbb{R}^3 \setminus D)$ . The only way for the left-hand side to become unbounded is for  $\|g(\cdot, z, q)\|_{L_t^2(\partial B_1)}$  to become unbounded as  $\zeta \rightarrow x$ . We see that the boundary of  $D$  is indicated by the growth of  $\|g(\cdot, z, q)\|_{L_t^2(\partial B_1)} < \infty$ . Later, after we have discussed how to solve (14.3) in more detail, we shall show that if  $\zeta \notin D$ , the procedure for computing  $g$  will also result in a function with large norm. Thus, we can say that  $D$  is indicated by the region of  $\zeta \in \mathbb{R}^3$  where  $F: L_t^2(\partial B_1) \rightarrow L_t^2(\partial B_1)$ .

Note that nowhere have we used any *a priori* information about the boundary conditions on  $\Gamma$  or the scattering mechanism in  $D$  (e.g.  $D$  could be a penetrable scatterer where  $\epsilon_r \neq 1$  or  $\mu_r \neq 1$ ). This makes clear that the implementation of the method is independent (within some limitations we shall mention later in this chapter) of *a priori* information on the scatterer.

The general scheme for finding  $\Gamma$  is now clear. We solve (14.3) for many  $\zeta$  in the region of  $\mathbb{R}^3$  where we expect  $D$  to lie. These sample points  $\zeta$  must have sufficient density that some will lie in  $D$  (hence the need for *a priori* data on the size and approximate position of the scatterer). Regions where  $g$  has small norm establish an outline of  $D$  (this can then be refined by using further sample points in the neighborhood of this approximate reconstruction, see [89]). Note

that (14.3) is a linear first kind integral equation. The combination of sampling and linearity gives rise to the name “linear sampling method” or LSM (clearly a major drawback of the scheme compared to other algorithms such as ART (cf. [232]) or MUSIC (cf. [125]) is a poor acronym).

**HEALTH WARNING** *The preceding paragraphs are the only place in this book where I have knowingly made a false argument! The heuristic argument supporting the LSM fails in general because for most scatterers (14.3) does not have a solution! There are some scatterers (e.g. a perfectly conducting sphere) where  $g$  does exist, but in general we cannot be sure of existence.*

In Section 14.3, we shall remedy this failing by giving a theoretical justification of the LSM. Next, however, we show how to implement the method assuming certain properties for  $F$  that will be proved later.

### 14.2.1 Implementing the LSM

In view of our upcoming discussion we make the following definition.

**Definition 14.1** Let  $\mathcal{A} : U \subset \chi \rightarrow v \subset Y$  be an operator from a subset  $U$  of a Hilbert space  $\chi$  into a subset  $v$  of a Hilbert space  $Y$ . The problem of finding  $\varphi \in U$  such that

$$\mathcal{A}\varphi = f$$

where  $f \in v$  is *wellposed* if  $\mathcal{A} : U \rightarrow v$  is bijective and the inverse operator  $\mathcal{A}^{-1} : v \rightarrow U$  is continuous. Otherwise the equation is *illposed*.

We wish to compute solutions of (14.3) (in fact, in view of the *Health Warning*, only approximate solutions). But the far field operator  $F$  is compact, since  $E_\infty$  is a smooth (even analytic) function of its arguments. The next lemma shows that this compactness implies that  $F^1$ , even if it exists, cannot be bounded and hence (14.3) is illposed.

**Lemma 14.2** *The far field operator  $F : L_t^2(\partial B_1) \rightarrow L_t^2(\partial B_1)$  given by (14.4) is compact and the far field equation (14.3) is illposed.*

**Proof** The fact that  $F$  is compact follows from the smoothness of the kernel  $E_\infty$  and Theorem 3.6. Suppose  $F^1$  exists, and is bounded. Then  $F^1 F = I$  and so  $I$  is compact (since  $F^1$  is bounded and  $F$  is compact, see Theorem 2.31). Hence, by Lemma 2.32,  $L_t^2(\partial B_1)$  is finite dimensional. This is a contradiction, so (14.3) is illposed.  $\square$

Since  $F^1$ , if it exists, cannot be bounded, we must seek approximate but stable solutions by modifying (14.3) (shortly we shall show that  $F$  is injective and so has an inverse on its range). This process is termed *regularization* and has been studied in great detail for many years (see, e.g., [137, 186]). Next, we shall outline enough of this theory to understand the approach we have adopted.

The method we shall use to regularize (14.3), the Tikhonov regularization with Morozov's discrepancy principle, is but one of many possible regularization

techniques. It has the virtue of giving reasonable reconstructions in practice. To understand the method, we first need to generalize the matrix singular value decomposition to operators in the standard way. Let  $F^*$  denote the  $L_t^2(\partial B_1)$  adjoint of  $F$  so that

$$(Fg, h)_{L_t^2(\partial B_1)} = \left( g, F^* h \right)_{L_t^2(\partial B_1)} \quad \text{forall } h, g \in L_t^2(\partial B_1).$$

Then  $F^*$  is bounded and so  $FF$  is bounded and self-adjoint. In addition,  $FF$  is compact, since  $F$  is compact. Hence by the Hilbert-Schmidt theory (see Theorem 2.36), we know that  $FF$  has an orthonormal sequence of eigenfunctions  $\varphi_n \in L_t^2(\partial B_1)$  and eigenvalues  $\mu_n^2$  such that  $\mu_1 \geq \mu_2 \geq \dots \geq \mu_n \rightarrow 0$  as  $n \rightarrow \infty$ , and

$$F^* F \varphi_n = \mu_n^2 \varphi_n, \quad n = 1, 2, \dots$$

We can define a second sequence of vectors  $\xi_n = (1/\mu_n)F\varphi_n$  so that (14.8)

$$F\varphi_n = \mu_n \xi_n, \quad F^* \xi_n = \mu_n \varphi_n, \quad \text{for } n = 1, 2, \dots$$

The set  $\{(\mu_n, \varphi_n, \xi_n)\}_{n=1}^\infty$  is called a *singular system* for  $F$ . Using this singular system, we have the following theorem (for a more general version and a discussion of the result, see Theorems 4.7 and 4.8 of [94]).

**Theorem 14.3** *Assuming  $F$  is injective and compact, any  $g \in L_t^2(\partial B_1)$  can be written as*

$$g = \sum_{n=1}^{\infty} (g, \varphi_n)_{L_t^2(\partial B_1)} \varphi_n \quad \text{and} \quad Fg = \sum_{n=1}^{\infty} \mu_n (g, \varphi_n)_{L_t^2(\partial B_1)} \xi_n.$$

*The generalized far field equation of finding  $g \in L_t^2(\partial B_1)$  such that  $Fg = f$  where  $f \in L_t^2(\partial B_1)$  is solvable if and only if  $f \in N(F^*)^\perp$  and satisfies*

$$\sum_{n=1}^{\infty} \frac{1}{\mu_n^2} \left| (f, \xi_n)_{L_t^2(\partial B_1)} \right|^2 < \infty.$$

*In this case, the solution is given by*

$$g = \sum_{n=1}^{\infty} \frac{1}{\mu_n} (f, \xi_n)_{L_t^2(\partial B_1)} \varphi_n.$$

**Remark 14.4** *In Lemma 14.14 we shall show that for at least one class of scattering problems,  $F$  is injective. In our application,  $f = E_{\epsilon_\infty}$ . The final equation of the theorem shows that the solution is very sensitive to perturbations of  $f$  in the higher Fourier modes, since  $\mu_n \rightarrow 0$  as  $n \rightarrow 0$ .*

Now we can describe the first part of the regularization strategy for (14.3). As mentioned before, we choose to apply the *Tikhonov regularization* [137, 186, 279].

**Lemma 14.5** *The operator  $(\alpha I + F^* F) : L_t^2(\partial B_1) \rightarrow L_t^2(\partial B_1)$ ,  $\alpha > 0$ , has a bounded inverse.*

**Proof** Let  $\tilde{c}(g, h) = \alpha(g, h)_{L_t^2(\partial B_1)} + (Fg, Fh)_{L_t^2(\partial B_1)}$ . Then  $\tilde{c}(\cdot, \cdot)$  satisfies the conditions of the Lax–Milgram Lemma 2.21 and the assertion of the lemma follows.  $\square$

Now we define the regularized solution of (14.3) by

$$g_\alpha = (\alpha I + F^* F)^{-1} F^* E_{e,\infty}.$$

Using the singular system (14.8)(14.9)

$$g_\alpha = \sum_{n=1}^{\infty} \frac{\mu_n}{\alpha + \mu_n^2} (E_{e,\infty}, \xi_n)_{L_t^2(\partial B_1)} \varphi_n.$$

If  $\alpha$  is chosen appropriately we can have  $\mu_n / (\alpha + \mu_n^2) \approx \mu_n^{-1}$  for the first few modes (i.e. small  $n$ ), but, since  $\mu_n \rightarrow 0$  as  $n \rightarrow \infty$ , we have

$$\frac{\mu_n}{\alpha + \mu_n^2} \approx \frac{\mu_n}{\alpha} \text{ for } n \text{ large enough}$$

so that the growth of the higher-order modes in expansion (14.9) are controlled. Clearly the choice of  $\alpha$  is of critical importance to the method. If  $\alpha$  is too large  $g_\alpha$  will be inaccurate, but if  $\alpha$  is too small stability will be compromised. We should note that the Tikhonov regularization is a valid regularization in the sense that if a solution  $g$  to (14.3) exists then in the limit as  $\alpha \rightarrow 0$ , we have, from (14.9),  $g_\alpha \rightarrow g$  in  $L_t^2(\partial B_1)$ .

Since we shall be solving the regularized version of (14.3) for many  $\tilde{z}$  with a wide range of norms expected for  $g$ , we need an automatic way to pick  $\alpha$  in a reasonable way. We do this using the Morozov discrepancy principle [137, 279, 157]. Let us define the norm of the residual,

$$R(\alpha) = \| Fg_\alpha - E_{e,\infty} \|_{L_t^2(\partial B_1)}.$$

If  $\alpha$  is small we expect  $R(\alpha)$  to be small, since many terms in the expansion (14.9) will be close to those for  $F^* E_{e,\infty}$  (but the norm of  $g_\alpha$  may be large due to instability). More precisely,

$$Fg_\alpha = \sum_{n=1}^{\infty} \frac{\mu_n^2}{\alpha + \mu_n^2} (E_{e,\infty}, \xi_n) \xi_n$$

and

$$E_{e,\infty} = \sum_{n=1}^{\infty} (E_{e,\infty}, \xi_n)_{L_t^2(\partial B_1)} \xi_n.$$

Thus,

$$R(\alpha)^2 = \sum_{n=1}^{\infty} \frac{\alpha^2}{(\alpha + \mu_n^2)^2} \left| (E_{e,\infty}, \xi_n)_{L_t^2(\partial B_1)} \right|^2. \quad (14.10)$$

This is a continuous and monotone increasing function of  $\alpha$ . In fact,  $R(\alpha) \rightarrow \|E_{e,\infty}\|_{L_t^2(\partial B_1)}$  as  $\alpha \rightarrow \infty$  and  $R(\alpha) \rightarrow 0$  as  $\alpha \rightarrow 0$ . On the other hand, (14.11)

$$\|g_\alpha\|_{L_t^2(\partial B_1)}^2 = \sum_{n=1}^{\infty} \frac{\mu_n^2}{(\alpha + \mu_n^2)^2} \left| (E_{e,\infty}, \xi_n)_{L_t^2(\partial B_1)} \right|^2$$

is monotone decreasing as  $\alpha \rightarrow \infty$ . The idea of the *Morozov discrepancy principle* is to balance  $R(\alpha)$  and  $\|g_\alpha\|_{L_t^2(\partial B_1)}$ , taking into account possible errors in  $F$  and  $E_{e,\infty}$  (the former come from measurement errors in  $E_e$ , while the latter are needed for our upcoming theory). Suppose we have available only measurements  $F_\delta$  of  $F$  (via its kernel) and an approximation  $E_{e,\infty}^\epsilon$  to  $E_{e,\infty}$  such that

$$\|F - F_\delta\|_{L_t^2(\partial B_1) \rightarrow L_t^2(\partial B_1)} \leq \delta \quad \text{and} \quad \|E_{e,\infty}^\epsilon - E_{e,\infty}\|_{L_t^2(\partial B_1)} \leq \epsilon.$$

The magnitude of the residual generated by actually computing with  $F_\delta$  and  $E_{e,\infty}^\epsilon$  can be estimated formally as follows. Suppose  $g_{\epsilon,\delta}$  solves exactly  $F g_{\epsilon,\delta} = E_{e,\infty}^\epsilon$ . Then

$$F g_{\epsilon,\delta} = (F - F_\delta) g_{\epsilon,\delta} = (F - F_\delta) g_{\epsilon,\delta} + \left( E_{e,\infty}^\epsilon - E_{e,\infty} \right) + E_{e,\infty}.$$

So  $\|F g_{\epsilon,\delta} - E_{e,\infty}\|_{L_t^2(\partial B_1)} \leq \delta \|g_{\epsilon,\delta}\|_{L_t^2(\partial B_1)} + \epsilon$ . It thus makes sense to choose  $\alpha$  so that  $R(\alpha) = \delta \|g_{\epsilon,\delta}\|_{L_t^2(\partial B_1)} + \epsilon$ . Of course,  $g_{\epsilon,\delta}$  is not available. In addition, in our applications,  $\epsilon$  is small compared to  $\delta \|g_{\epsilon,\delta}\|_{L_t^2(\partial B_1)}$ , so we actually choose  $\alpha$  to satisfy the following equality

$$R(\alpha) = \delta \|g_\alpha\|_{L_t^2(\partial B_1)}.$$

Squaring both sides and using the expansions for  $R(\alpha)$  and  $g_\alpha$  in (14.10) and (14.11), we see that this implies finding the zero of

$$\mu(\alpha) = \sum_{n=1}^{\infty} \frac{\alpha^2 - \delta^2 \mu_n^2}{(\mu_n^2 + \alpha)^2} \left| (E_{e,\infty}, \xi_n)_{L_t^2(\partial B_1)} \right|^2.$$

The function  $\mu(\alpha)$  is obviously monotone increasing and so there is at most one zero (in fact,  $\mu(\alpha)$  is negative for  $\alpha$  small enough and positive if  $\alpha > \delta \mu_n$ , and there is one zero). Despite this nice property, solving  $\mu(\alpha) = 0$  for all sample points  $\zeta$  significantly slows the inverse algorithm. In addition,  $\mu(\alpha)$  is rather flat, so the magnitude of the zero can vary over many orders of magnitude in a single inverse problem. We usually work with  $\ln(\alpha)$  rather than  $\alpha$  itself. Of course, we do not need a very accurate solution of this problem. For noisy data,  $F$  is not known and we must work with the singular system for  $F_\delta$  rather than  $F$ .

Now we discuss one method for solving (14.3) from [90]. The left-hand side of (14.3) is convenient for theory, but obscures the dependence of the far field

operator on  $g$ . We first derive an equivalent form. Let  $(e_1(\hat{x}), e_2(\hat{x}), \hat{x})$  denote an orthonormal basis for  $\mathbb{R}^3$ . Since the far field pattern  $E_\infty$  is radial, we know that  $\hat{x} \cdot E_\infty(\hat{x}, d, p) = 0$  (see Corollary 9.5). Hence, (14.3) is equivalent to the two scalar equations, (14.12)

$$\int_{\partial B_1} e_j(\hat{x}) \cdot E_\infty(\hat{x}, d, g(d)) dA(d) = e_j(\hat{x}) \cdot E_e(\hat{x}, z, q)$$

for all  $\hat{x} \in \partial B_1$  and  $j = 1, 2$ . Now we assume that the far field pattern satisfies the reciprocity relation:  $s \cdot E_\infty(\hat{x}, d, r) = r \cdot E_\infty(-d, -\hat{x}, s)$  for all  $\hat{x}, d \in \partial B_1$  and  $r, s \in \mathbb{C}^3$ . We shall prove this shortly in Theorem 14.15 for a particular class of scatterers. Using the reciprocity relation, we may rewrite (14.12) as (14.13)

$$\int_{\partial B_1} E_\infty(-d, -\hat{x}, e_j(\hat{x})) \cdot g(d) dA(d) = e_j(\hat{x}) \cdot E_{e,\infty}(\hat{x}, z, d)$$

for all  $\hat{x} \in \partial B_1$  and  $j = 1, 2$ . This clarifies the fact that the far field operator is linear in  $g$ .

For the numerical study we shall present shortly, we now replace the integral in (14.13) by a sum using a quadrature approximation. We first approximate the surface  $\partial B_1$  using a triangulation. Each triangle  $T$  of this mesh corresponds, by radial projection, to a segment  $S$  of the unit sphere. Suppose  $T$  has vertices  $a_j^T, j = 1, 2, 3$ . Then for any smooth function  $f$  on  $\partial B_1$ , we approximate

$$\int_T f dA \approx \frac{1}{3} \text{area}(T) \sum_{j=1}^3 f(a_j).$$

If the triangulation has  $N_H$  vertices given by  $d_m, m = 1, \dots, N_H$ , we obtain

$$\int_{\partial B_1} f dA \approx \sum_{m=1}^{N_H} \omega_m f(d_m),$$

where the quadrature weights are  $\omega_m, m = 1, \dots, N_H$ . In our numerical experiments,  $N_H = 42$  and we choose the evaluation points for  $E_\infty$  to agree with the incident directions so that successively  $\hat{x} = d_m, 1 \leq m \leq N_H$ . In view of the fact that  $d_j \cdot g(d) = 0$ , we may write  $g = g_1 e_1(d) + g_2 e_2(d)$  where  $g_1, g_2 \in \mathbb{C}$ . Thus, the fully discrete problem corresponding to (14.3) is to find  $g_1, g_2, 1 \leq j \leq N_H$ , such that

$$\sum_{j=1}^{N_H} \sum_{n=1}^2 \omega_j E_\infty(-d_j, -d_m, e_n(d_m)) \cdot e_n(d_j) g_{n,j} = e_l(d_m) \cdot E_{e,\infty}(d_m, z, q)$$

for  $l = 1, 2$  and  $m = 1, 2, \dots, N_H$ . This gives rise to  $2N_H$  linear equations in  $2N_H$  unknowns. By enumerating the equations and unknowns in this problem, it may be written as

(14.14)

$$A_\infty \vec{g} = \vec{f},$$

where  $\vec{f}$  is a vector given by the right hand side above and  $A_\infty$  is a  $2N_H \times 2N_H$  matrix.

In our forthcoming numerical study, we shall show results where  $E_\infty$  is computed numerically. The use of computed data for the inverse problem can give rise to unwitting “inverse crimes”. This phrase, coined in [94], refers to the possibility that the computed data for an inverse algorithm can be especially well suited to the inversion scheme under investigation. To avoid such crimes and also to simulate measurement error, we actually replace  $A_\infty$  by  $A_\infty^\epsilon$  given by

$$\left(A_\infty^\epsilon\right)_{l,m} = (A_\infty)_{l,m} (1 + \epsilon (R_1)_{l,m} + i \epsilon (R_2)_{l,m}) \quad 1 \leq l, m \leq 2N_H,$$

where  $\epsilon > 0$  is an error parameter and  $R_1$  and  $R_2$  are  $(2N_H) \times (2N_H)$  matrices of random numbers uniformly distributed in  $[-1,1]$ . Of course, for a physical measurement device, the error is unlikely to be uniform, but this suffices to investigate the numerical scheme. We chose  $\epsilon = 0.01$  in the results shown later.

Now we apply the Tikhonov regularization to (14.14) with  $A_\infty$  replaced by  $A_\infty^\epsilon$  so that the regularized solution, still denoted by  $\vec{g}$ , satisfies

$$\left(aI + \left(A_\infty^\epsilon\right)^* A_\infty^\epsilon\right) \vec{g} = \left(A_\infty^\epsilon\right)^* \vec{f},$$

where  $\alpha > 0$  is the regularization parameter. Using the singular value decomposition of  $A_\infty^\epsilon$ , we can easily solve this equation for  $\vec{g}$ . We then use the Morozov discrepancy principle described earlier in this section to find an appropriate  $\alpha$  for each  $\zeta$  and  $q$  by taking the Morozov parameter corresponding to the noise level  $\delta = \|A_\infty - A_\infty^\epsilon\|$  (the spectral matrix norm). Note that this choice of  $\delta$  ignores errors due to the numerical scheme for computing the data  $A_\infty$ .

Once we have computed  $\vec{g}$  for a given  $\zeta$  and  $q$ , we can compute  $\|\vec{g}\|_H$  given by

$$\|\vec{g}\|_H^2 = \sum_{j=1}^{N_H} \omega_j |g_j|^2.$$

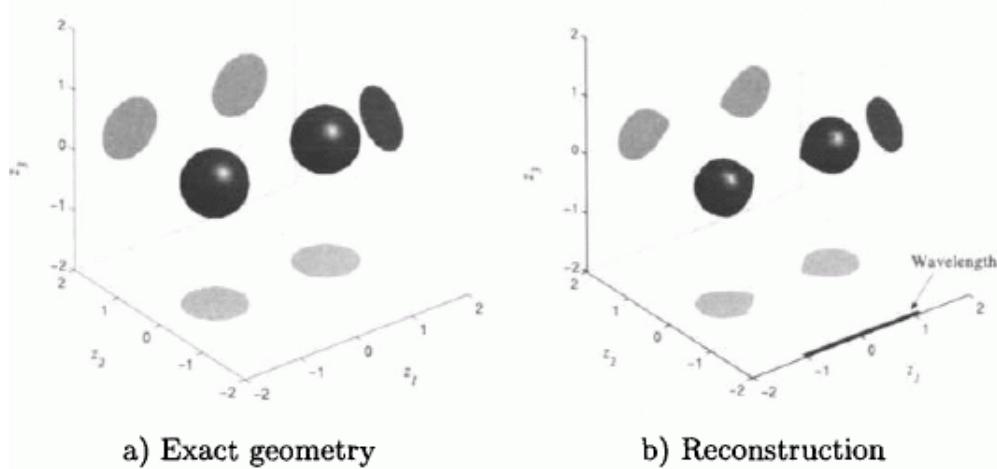
We expect that the boundary  $\partial D$  will be indicated by large values of  $\|\vec{g}\|_H$ , and that  $\|\vec{g}\|_H$  will also be large outside  $D$ . Of course, this statement begs the question of what is “large”.

In practice, this is a difficult question to answer. One approach is to first reconstruct a sphere or other known target with similar size and boundary conditions as the unknown scatterer. A parameter  $C$  is then chosen so that the boundary of the object is indicated by the surface of points  $\zeta$  such that  $G(\zeta) = C$  where(14.15)

$$G(\zeta) = \frac{1}{3} \left( \|\vec{g}(q = e_1)\|_H^{-1} + \|\vec{g}(q = e_2)\|_H^{-1} + \|\vec{g}(q = e_3)\|_H^{-1} \right).$$

Note that here we have combined data for each polarization  $q$  in a single function  $G(\zeta)$ , since  $e_j, j = 1, 2, 3$ , are the three standard unit vectors.

Fig. 14.1. Reconstruction of two perfectly conducting spheres. (a) Surface of the true or exact scatterer. (b) Reconstruction using the LSM with criterion (14.16) used to choose the isosurface shown. Note that the wavelength of the incident field (shown in (b)) is larger than the diameter of either sphere.<sup>9</sup>



The “calibration” approach just outlined was studied in [89] and used in a number of subsequent publications [88]. If we do not know the boundary condition on the unknown scatterer, this approach fails. In the numerical study reported here from [90] we just choose a value for  $C$  *a posteriori*.

### 14.2.2 Numerical results with the LSM

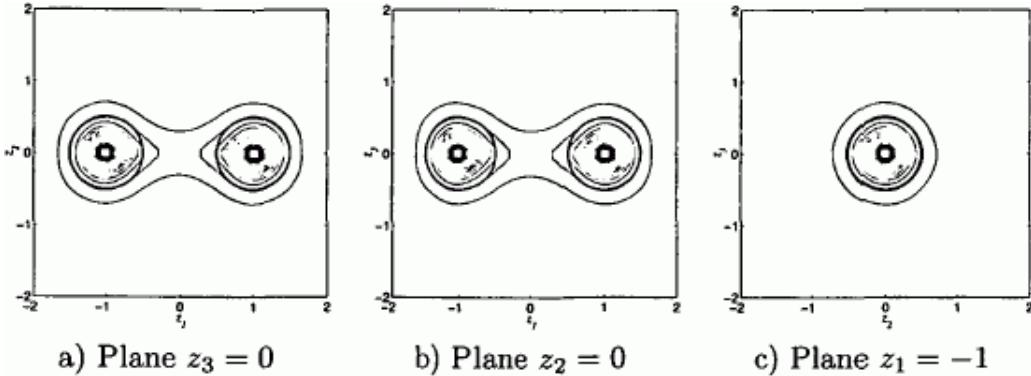
Most of the numerical results we show are from [90]. In that study we choose the sampling points  $\vec{x}$  to lie on a uniform grid in the cube  $[-2, 2]^3$  with a mesh size of 0.1. Thus, we sampled at  $(41)^3$  points. We could have used the adaptive approach of [89], but preferred to spend computer time on the reconstruction rather than human programming time on the more sophisticated algorithm. For this reason we do not report computer times.

Data for the inverse problem were obtained using a finite element like code for approximating the scattering problem (actually the perfectly conducting scattering problem) using a truncated domain and the Silver–Muller radiation condition as described in Section 13.5. The method actually used was the “ultra weak variational formulation” of Maxwell’s equations discretized using plane-waves of [74] with four plane-wave directions per tetrahedron. This code was conveniently available at the time of the study from which these results are reproduced.

We start with the problem of reconstructing two perfectly conducting spheres. The wavenumber is  $\kappa = 3$ , and so the two spheres are approximately one-half

<sup>9</sup> Reprinted from *SIAM J. Sci. Comput.*, 24, The linear sampling method for solving the electromagnetic inverse scattering problem, 719–731, D. Colton, H. Haddar and P. Monk, Copyright 2002, with permission from SIAM.

Fig. 14.2. To provide more details of the reconstruction shown in Fig. 14.14 we show here contour maps of  $g(\vec{z})$  on three planes through the scatterer. We also superimpose the exact scatter for comparison. Clearly, the outlines of the scatterers are visible, but there is a slight pinching of the reconstruction towards one another.<sup>10</sup>



wavelength in diameter and their centers are separated by approximately one wavelength. This choice of  $\varkappa$  was dictated by limitations on the forward code, but results in a difficult inverse problem since the scatterers are less than a wavelength in diameter. Generally, the fidelity of a reconstruction improves as  $\varkappa$  increases and the wavelength of the incident field becomes smaller than the scatterers. However, the reconstruction shown in Fig. 14.1 is acceptable. These plots show the isosurface of  $\|\vec{g}\|_H$  given by(14.16)

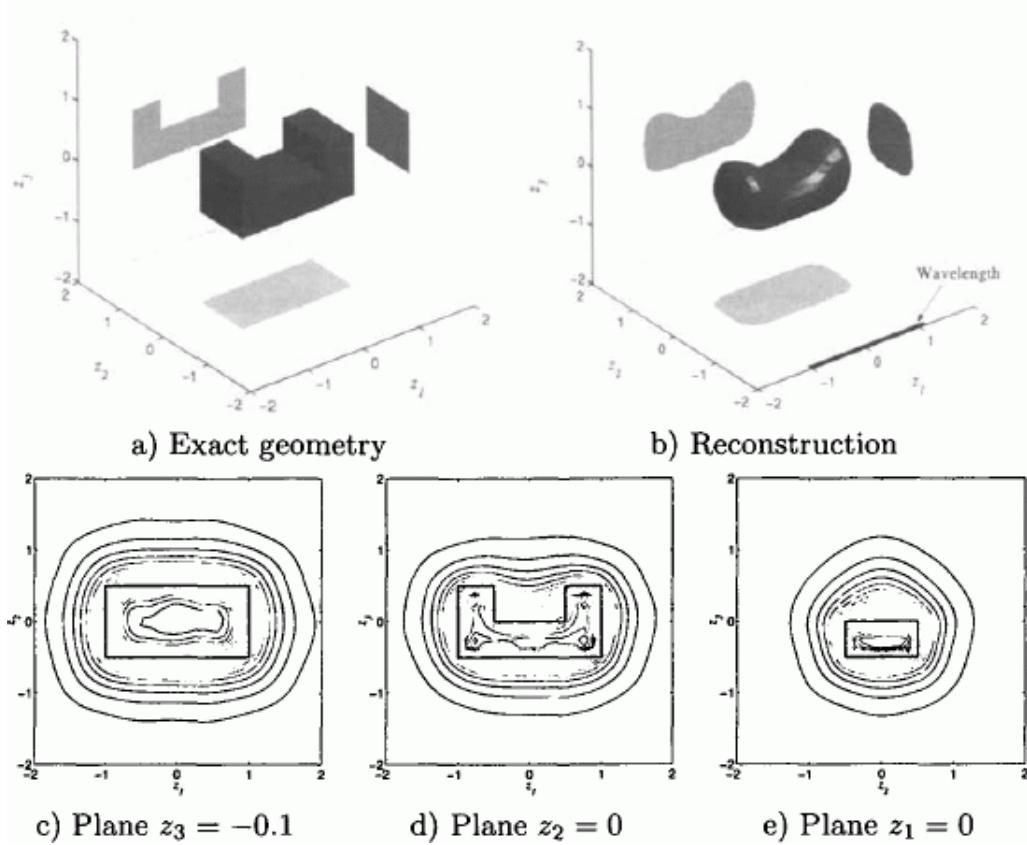
$$\mathcal{G}(z) = 0.2 \max_z \mathcal{G}(z),$$

where  $G(\vec{z})$  is given by (14.15) and includes information from all polarizations  $q$ . Figure 14.2 shows 20 contours of  $G(\vec{z})$  on different cross-sections through the scatterer. These reveal that the reconstruction is slightly pinched towards each other. This is typical of reconstructions of objects that are close together. If the balls were still closer together the pinching could become more obvious, even closer and the balls would be reconstructed as an elongated “dumbbell”. In these computations, the error parameter  $\varepsilon$  was set to 0.01.

It is reasonable to ask if it is helpful to use all polarization data (by this we mean different  $q$ ). Next we show the reconstruction of a U-shaped scatterer in Fig. 14.3 . Parameters are as in the previous problem. Clearly, it is possible for the LSM to reconstruct non-convex scatterers. Of course the indentation of the U is smoothed, but the indentation, which is much less than a wavelength deep and so is difficult to reconstruct, is evident. In Fig. 14.3 we used  $G(\vec{z})$  given by (14.15), including data for all polarizations  $q$ . The contour level drawn is

<sup>10</sup> Reprinted from *SIAM J. Sci. Comput.*, 24 , The linear sampling method for solving the electromagnetic inverse scattering problem, 719–731, D. Colton, H. Haddar and P. Monk, Copyright 2002, with permission from SIAM.

Fig. 14.3. (a) The exact U-shaped scatterer and (b) its reconstruction using criterion (14.17). In the remaining panels we show contour maps of  $g(\vec{z})$  on surfaces through the scatterer. In this case, all polarization data are used ( $g$  is given by (14.15)). Clearly, it is possible to reconstruct non-convex objects.<sup>11</sup>



(14.17)

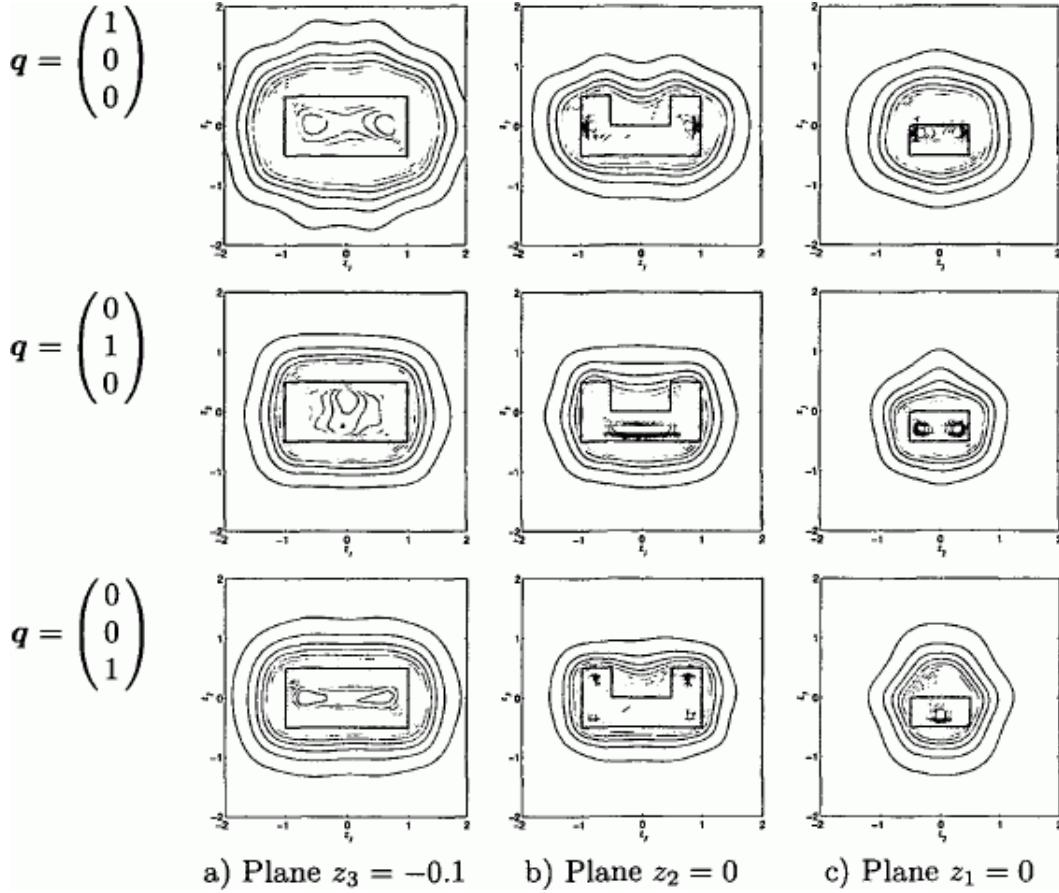
$$\mathcal{G}(z) = 0.3 \max_z \mathcal{G}(z),$$

Figure 14.4 shows contour plots of  $\left\| \vec{g} (q = e_j) \right\|_{H}^{-1}$  for  $j = 1, 2, 3$  on various sections through the scatterer. These individual components of  $G$  each emphasize different aspects of the scatterer. For example,  $\left\| \vec{g} (q = e_3) \right\|_{H}^{-1}$  provides resolution of the vertical parts of  $U$  and so on. It is clearly important to combine all these polarization data in order to get an accurate reconstruction.

We have commented a number of times that the LSM does not require *a priori* knowledge concerning the boundary data on the unknown scatterer. Of course, the fidelity of the construction will depend on the properties of the actual scatterers present (and so on boundary data), but the method need not be

<sup>11</sup> Reprinted from *SIAM J. Sci. Comput.*, 24, The linear sampling method for solving the electromagnetic inverse scattering problem, 719–731, D. Colton, H. Haddar and P. Monk, Copyright 2002, with permission from SIAM.

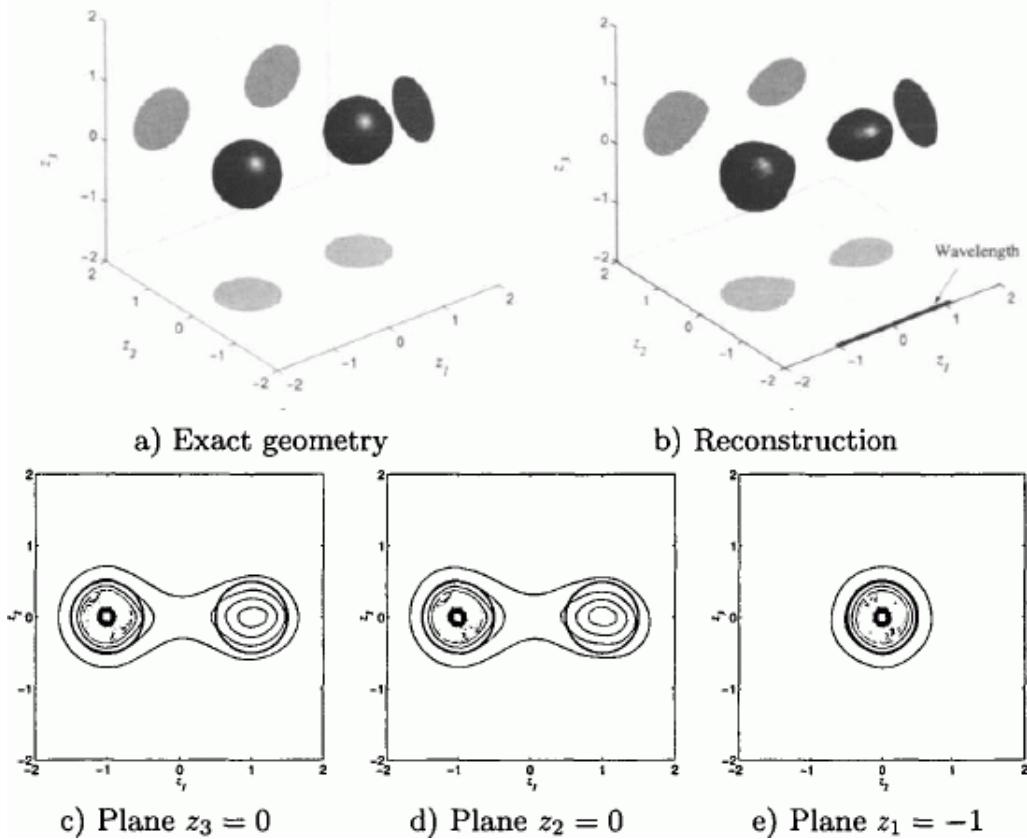
Fig. 14.4. To investigate the influence of the polarization  $q$  on the reconstruction, we show the individual components of  $g$ . The top row of figures shows  $\|g(q = e_1)\|_{L^2(\partial\Omega)}^{-1}$ , etc. Combining the results for all three choices of  $q$  (see Figure 14.3) results in a better reconstruction than for any single choice of  $q$ <sup>12</sup>



(indeed cannot be!) adjusted for this. In the next and last example, we show the reconstruction of two balls. One is perfectly conducting, but the other is penetrable conducting with  $\epsilon_r = 2 + 2i$ . In this case,  $\nu = 3$ . The reconstruction is shown in Fig. 14.5. The penetrable sphere is reconstructed almost as well as the perfectly conducting ball. Criterion (14.16) (the same as for two perfectly conducting balls) is used to draw the three-dimensional pictures. This case is not covered by the theory presented in the next section. For further details concerning this example, see [155].

<sup>12</sup> Reprinted from *SIAM J. Sci. Comput.*, 24, The linear sampling method for solving the electromagnetic inverse scattering problem, 719–731, D. Colton, H. Haddar and P. Monk, Copyright 2002, with permission from SIAM.

Fig. 14.5. Reconstruction of a perfectly conducting sphere next to a penetrable sphere (with  $\epsilon_r = 2 + 2i$ ). (a) Surface of the true or exact scatterer. (b) Reconstruction using the LSM with criterion (14.16) used to choose the isosurface shown. In the remaining panels we show sections through the reconstruction. Clearly, the penetrable sphere is less easily visible than the perfectly conducting sphere.<sup>13</sup>



### 14.3 Mathematical aspects of inverse scattering

We are now going to show that the heuristic arguments given in the previous section can be given a solid foundation. This requires that we make concrete assumptions about the scatterer. We shall consider only the simplest case when  $D$  is a bounded perfectly conducting obstacle in a uniform background medium (for extensions to other situations, see [90, 155, 70]).

Thus, we assume that the total field  $E \in H_{\text{loc}}(\text{curl}; \mathbb{R}^3 \setminus D)$  satisfies (14.18)

$$\nabla \times \nabla \times E - \kappa^2 E = 0 \quad \text{in } \mathbb{R}^3 \setminus D,$$

<sup>13</sup> Reprinted from *Inverse Problems*, 18, The linear sampling method for solving the electromagnetic inverse scattering problem, 891–906, H. Haddar and P. Monk, Copyright 2002, with permission from IOP Publishing Ltd.

$$\mathbf{E} = \mathbf{E}^s + \mathbf{E}^i \text{ in } \mathbb{R}^3 \setminus D, \quad (14.19)$$

$$\mathbf{v} \times \mathbf{E} = 0 \text{ on } \Gamma = \partial D, \quad (14.20)$$

$$\lim_{\rho \rightarrow 0} \rho \left( (\nabla \times \mathbf{E}^s) \times \hat{\mathbf{x}} - ik \mathbf{E}^s \right) = 0, \quad (14.21)$$

where  $\mathbf{E}$  is given by (14.1) and  $D$  is a Lipschitz polyhedral domain with simply connected complement.

Before we start our analysis we need to re-examine the space  $Y(\Gamma)$  where  $\Gamma = \partial D$  defined in (3.50). In particular, we need to revert to the original definition of Chen *et al.* [77], since we need to consider traces on  $\Gamma$  of functions on  $D$  and on  $\mathbb{R}^3 \setminus D^-$ . We define (14.22)

$$Y(\Gamma) = \left\{ f \in \left( H^{-1/2}(\Gamma) \right)^3 \middle| \text{there is a function } \mathbf{u} \in H_0(\mathbf{curl}; B_R) \text{ with } \mathbf{v} \times \mathbf{u} = f \text{ on } \Gamma \right\},$$

where  $R$  is large enough, so that  $\Gamma \subset B_R$ .

Note that the choice  $f \in (H^{1/2}(\Gamma))^3$  follows from Theorem 3.29. The space  $Y(\Gamma)$  is equipped with the norm

$$\|f\|_{Y(\Gamma)} = \inf_{\substack{\mathbf{u} \in H_0(\mathbf{curl}; B_R) \\ \mathbf{v} \times \mathbf{u} = f}} \|\mathbf{u}\|_{H(\mathbf{curl}; B_R)}.$$

It is possible to show that  $Y(\Gamma)$  is a Banach space. Using the extension result for  $H(\mathbf{curl}; D)$  functions in Theorem 3.34, we can prove that  $\|\cdot\|_{Y(\Gamma)}$  is equivalent to the norm

$$\|f\|_{Y(\Gamma),1} = \inf_{\substack{\varphi \in H(\mathbf{curl}; D) \\ \mathbf{v} \times \varphi = f \text{ on } \Gamma}} \frac{((f, \varphi))_1}{\|\varphi\|_{H(\mathbf{curl}; D)}},$$

where

$$((f, \varphi))_1 = \int_D (\nabla \times \mathbf{u} \cdot \varphi - \mathbf{u} \cdot \nabla \times \varphi) dV.$$

and  $\mathbf{u} \in H(\mathbf{curl}; D)$  is any function such that  $\mathbf{v} \times \mathbf{u} = f$  on  $\Gamma$ . Using the extension theorem again, this time applied to  $H(\mathbf{curl}; B_R \setminus D^-)$ , shows that  $\|f\|_{Y(\Gamma)}$  is also equivalent to

$$\|f\|_{Y(\Gamma),2} = \inf_{\substack{\varphi \in H(\mathbf{curl}; B_R \setminus \bar{D}) \\ \mathbf{v} \times \varphi = f \text{ on } \Gamma, \mathbf{v} \times \varphi = 0 \text{ on } \partial B_R}} \frac{((f, \varphi))_2}{\|\varphi\|_{H(\mathbf{curl}; B_R \setminus \bar{D})}},$$

where

$$((f, \varphi))_2 = \int_{B_R \setminus \bar{D}} (\nabla \times \mathbf{u} \cdot \varphi - \mathbf{u} \cdot \nabla \times \varphi) dV.$$

and  $\mathbf{u} \in H(\mathbf{curl}; B_R \setminus D^-)$  is any function such that  $\mathbf{v} \times \mathbf{u} = f$  on  $\Gamma$  and  $\mathbf{v} \times \mathbf{u} = 0$  on  $\partial B_R$ . Thus, traces of functions defined on either side of  $\Gamma$  suffice to characterize  $Y(\Gamma)$ . In addition,  $Y(\Gamma)$  is in fact a Hilbert space. These facts are

not surprising, since  $Y(\Gamma)$  can be given an intrinsic characterization. It is in fact the appropriate generalization (non-trivial indeed!) to Lipschitz boundaries of the space  $H^{1/2}(\text{Div}; \Gamma)$ . Detailed results can be found in [63]. We do not give the results here, because the indirect characterization in (14.22) suffices for us.

The plan of this section is now as follows. First, we show that the data used for the inverse problem uniquely determine the scatterer. Then, in Section 14.3.2 we examine in detail properties of the Herglotz wave functions defined in (14.5). Using these functions, we then verify various properties of the far field operator  $F$  defined in (14.4) that justify the use of regularization to solve (14.3) and our particular numerical implementation. Finally, we prove the desired results justifying the LSM in Section 14.3.4. Our presentation follows [94, 95, 70].

### 14.3.1 Uniqueness for the inverse problem

Our next result shows that an exact knowledge of the far field pattern for all measurement points  $\mathcal{O}$ , all directions  $d$  and all polarizations  $p$  uniquely determines  $D$ . Thus, at least in theory, there is sufficient information in the data to determine the boundary of  $D$ . This is a comforting result, if we are about to try to solve the inverse scattering problem (but, of course, the result says nothing about the stability of the reconstruction with respect to measurement error). This version of the theorem is taken from [194].

**Theorem 14.6** *Let  $D_1$  and  $D_2$  be two perfectly conducting Lipschitz polyhedral scatterers such that at a fixed wavenumber  $\kappa$  the far field patterns for both scatterers coincide for all directions  $d$  and polarizations  $p$ . Then  $D_1 = D_2$ .*

This theorem is proved most easily using a remarkable reciprocity result first proved by Potthast [254], termed a “mixed reciprocity” result since it gives an equality between solutions of (14.18)–(14.21), and a solution of the same problem except that the incident wave is due to an incident field from a dipole source. Let  $E_{dp}$  and  $E_{dp}^s$  be the total field and scattered field solutions of (14.18)–(14.21) when (14.23)

$$E_{dp}^i(x) = -\frac{1}{ik} \nabla_x \times \nabla_x \times (q \Phi(x, z)), \quad x \neq z,$$

where  $\Phi$  is, as usual, the fundamental solution of the Helmholtz equation (see (9.1)) and  $q \in C^3$ ,  $q \neq 0$ , is the polarization of the dipole source located at  $z \in R^3$ .

We may write  $E_{dp} = E_{dp}(x, z, q)$  (or  $E_{dp}^s = E_{dp}^s(x, z, q)$ ). Similarly, we denote by  $E(x, d, p)$  or  $E^s(x, d, p)$  the solutions of (14.18)–(14.21) with the incident plane wave given by (14.1).

**Theorem 14.7** *Suppose  $D$  is a perfect conductor, then*

$$p \cdot E_{dp,\infty}(\hat{x}, z, q) = \frac{1}{4\pi} q \cdot E^S(z, -\hat{x}, p)$$

for all  $\mathcal{O} \in \partial B_1$ ,  $z \in R^3 \setminus D^-$  and  $p, q \in C^3$ .

**Remark 14.8** In proving this theorem, it is useful to use the magnetic fields  $H = (1/ik) \nabla \times E$  and  $H^i = (1/ik) \nabla \times E^i$ , both for point source and plane wave solutions.

**Proof of Theorem 14.7** The proof uses the representation theorem for the far field pattern and (3.51). By the Stratton–Chu formula in Theorem 9.4, we have(14.24)

$$\begin{aligned} E^S(z, \hat{x}, p) &= \nabla_z \times \int_{\Gamma} v \times E^S(y, -\hat{x}, p) \Phi(z, y) dA(y) \\ &\quad - \frac{1}{ik} \nabla_z \times \nabla_z \times \int_{\Gamma} v \times H^S(y, -\hat{x}, p) \Phi(z, y) dA(y) \\ &= \nabla_z \times \int_{\Gamma} v \times E^S(y, -\hat{x}, p) \Phi(z, y) dA(y) \\ &\quad - \frac{1}{ik} \nabla_z \times \nabla_z \times \int_{\Gamma} v \times H(y, -\hat{x}, p) \Phi(z, y) dA(y) \\ &\quad \left( \nabla_z \times \int_{\Gamma} v \times E^i(y) \Phi(z, y) dA(y) \right. \\ &\quad \left. - \frac{1}{ik} \nabla_z \times \nabla_z \times \int_{\Gamma} v \times H^i(y) \Phi(z, y) dA(y) \right) \end{aligned}$$

The term in parentheses on the last lines of the above equation evaluates to zero, since  $\zeta \notin D$  so that  $E^i$  satisfies Maxwell's equations in  $D$  and  $\Phi(\zeta, y)$  satisfies the Helmholtz equation in  $D$  (see the proof of the first Stratton–Chu formula in Theorem 9.1). Furthermore,  $v \times E = 0$  on  $\Gamma$ , so (14.24) may be written as

$$\begin{aligned} q \cdot E^S(z, -\hat{x}, p) &= -\frac{1}{ik} q \cdot \nabla_z \times \nabla_z \times \int_{\Gamma} v \times H(y, -\hat{x}, p) \Phi(z, y) dA(y) \\ &= \int_{\Gamma} v \times H(y, -\hat{x}, p) \cdot E_{dp}^i(y) dA(y), \end{aligned}$$

where we have used the fact that for constant vectors  $r, s \in \mathbb{C}^3$ , we have

$$r \cdot \nabla_z \times \nabla_z \times (s \Phi(z, y)) = s \cdot \nabla_z \times \nabla_z \times (r \Phi(z, y))$$

and  $E_{dp}^i$  is given by (14.23). Again using the boundary condition on  $\Gamma$ , we have(14.25)

$$\begin{aligned} q \cdot E^S(z, -\hat{x}, p) &= \int_{\Gamma} v \times H(y, -\hat{x}, p) \cdot E_{dp}(y, z, q) dA(y) \\ &\quad - \int_{\Gamma} v \times H(y, -\hat{x}, p) \cdot E_{dp}^s(y, z, q) dA(y) \\ &= - \int_{\Gamma} v \times H(y, -\hat{x}, p) \cdot E_{dp}^s(y, z, q) dA(y) \\ &= - \int_{\Gamma} \{v \times H(y, -\hat{x}, p) \cdot E_{dp}^s(y, z, q) \\ &\quad + v \times H_{dp}^s(y, z, q) \cdot E(y, -\hat{x}, p)\} dA(y) \\ &= \int_{\Gamma} \{v \times E_{dp}^s(y, z, q) \cdot H^i(y) \\ &\quad + v \times H_{dp}^s(y, z, q) \cdot E^i(y)\} dA(y) \\ &= + \int_{\Gamma} \{v \times E_{dp}^s(y, z, q) \cdot H^s(y, -\hat{x}, p) \\ &\quad + v \times H_{dp}^s(y, z, q) \cdot E^s(y, -\hat{x}, p)\} dA(y). \end{aligned}$$

The last integral here vanishes, because both  $E_{dp}^s, H_{dp}^s$  and  $(E^s, H^s)$  are solutions of Maxwell's equations in  $\mathbb{R}^3 \setminus D^-$  satisfying the radiation condition (see the proof of the Stratton–Chu formula in Theorem 9.1 for example). Hence, using the definition of  $E^i$  and  $H^i$  and comparing the first integral on the right hand side to the representation for  $E_{dp,\infty}$  given in Corollary 9.5 shows that

$$\begin{aligned} \int_{\Gamma} \left\{ v \times E_{dp}^s(y, z, q) \cdot H^i(y) + v \times H_{dp}^s(y, z, q) \cdot E^i(y) \right\} dA(y) \\ = 4\pi p \cdot E_{dp,\infty}(\hat{x}, z, q). \end{aligned}$$

Use of this in (14.25) proves the result.  $\square$

We are now in a position to prove the uniqueness theorem. Besides the previously proved result on mixed reciprocity, this proof also makes use of the well-posedness of the forward problem in the exterior of  $D$  guaranteed by Theorem 10.8. In particular, let  $X$  be the space defined in (4.3). We know that  $\|E\|_X$  is bounded in terms of the data for the scattering problem. We use the spaces  $X_1$  where the domain is  $\Omega_1$  exterior to  $D_1$ , and  $X_2$  where  $\Omega_2$  is exterior to  $D_2$  chosen large enough that  $\Omega_1 \cap \Omega_2$  contains a neighbourhood of  $D_1$  and  $D_2$ .

**Proof of Theorem 14.6** By Rellich's lemma in the form of Corollary 9.29 and unique continuation (see Theorem 4.13), the equality of the far field patterns implies that the scattered field due to  $D_1$ , denoted by  $E_1^s$ , equals the scattered field  $E_2^s$  due to  $D_2$  in  $G$  which denotes the unbounded component of  $\mathbb{R}^3 \setminus (D_1^- \cup D_2^-)$ . This holds for all directions  $d$  and all polarizations. Hence, via the mixed reciprocity result of Lemma 14.7, we see that  $E_{1,dp,\infty}$  (the far field pattern due to a dipole point source at  $\zeta \in G$  scattered from  $D_1$ ) equals the far field pattern  $E_{2,dp,\infty}$  due to the same source and domain  $D_2$ . Again, by Rellich's lemma and unique continuation, we conclude that  $E_{1,dp}^s = E_{2,dp}^s$  in  $G$ , for all source points  $\zeta \in G$ .

Now suppose  $D_1 \neq D_2$ . Then (perhaps after renumbering the domains!) there is a point  $x^* \in \partial D_1$ ,  $x^* \notin D_2^-$ . Since  $D_1$  is a Lipschitz polyhedron, we can also choose  $x^*$  to lie in the interior of a face of  $\partial D_1$ , which then has a well-defined normal  $v(x^*)$ . Let

$$z_n = x^* + \frac{1}{n} v(x^*)$$

for  $n \geq N_0$ , such that  $\zeta_n \in G$ . Denoting by  $E_{2,dp}^s(x, z)$  the scattered field due to a point source at  $\zeta$  and  $D_2$ , we have

$$\| E_{1,d,p}^s(\cdot, z_n) - \bar{E}_{2,d,p}(\cdot, x^*) \|_{H(\text{curl}, \Omega_1 \cap \Omega_2)} \leq \| E_{2,d,p}^s(\cdot, z_n) - E_{2,d,p}^s(\cdot, x^*) \|_{X_2}.$$

The right-hand side tends to zero as  $n \rightarrow \infty$  due to the well-posedness of the forward problem for scattering from  $D_2$  in the  $X_2$  norm (see Theorem 10.8 and recall  $x^* \notin D_2^-$ , so the boundary data on  $D_2$  converges in  $Y(\partial D_2)$ ).

On the other hand, because of the boundary condition

$$\mathbf{v} \times E_{1,d,p}^s(\cdot, z_n) = -\mathbf{v} \times E_{d,p}^i(\cdot, z_n) \text{ on } \partial D,$$

where  $E_{d,p}^i$  is the dipole source at  $z_n$  and the fact that  $E_{d,p}^i(\cdot, x^*) \notin H(\text{curl}, \Omega_1 \cap \Omega_2)$  we have

$$\lim_{n \rightarrow \infty} \| E_{1,d,p}^s(\cdot, z_n) \|_{H(\text{curl}, \Omega_1 \cap \Omega_2)} \rightarrow \infty.$$

This is a contraction. Thus, it must be impossible to find such a point  $x^*$  and we conclude that  $D_1 = D_2$  as required.  $\square$

### 14.3.2 Herglotz wave functions

A principal tool of the analysis of the LSM is the Herglotz wave function. Let us recall the definition.

**Definition 14.9** A field  $E_g$  is an *electric Herglotz wave function* if there is a function  $g \in L_t^2(\partial B_1)$  such that

$$E_g(x) = \int_{\partial B_1} g(d) \exp(i \kappa x \cdot d) dA(d) \quad \text{forall } x \in \mathbb{R}^3.$$

Surprisingly such functions appear frequently in scattering applications. Note that  $E_g$  is defined (and analytic) for any  $x$ . In fact,  $E_g$  is characterized by the following lemma due to Hartmann and Wilcox [158].

**Lemma 14.10** A function  $E$  is a solution of Maxwell's equation (14.18) in all of  $\mathbb{R}^3$  subject to the growth condition  $\|E\|_{H(\text{curl}; \mathbb{R})} = O(R^{1/2})$  if and only if  $E$  is an electric Herglotz wave function.

**Proof** Suppose  $E_g$  is an electric Herglotz wave function. Then

$$\nabla \cdot E_g = i \kappa \int_{\partial B_1} \exp(i \kappa x \cdot d) d \cdot g(d) dA(d) = 0 \text{ in } \mathbb{R}^3,$$

where we have used the fact that  $g$  is a tangential vector field. Hence,

$$\nabla \times \nabla \times E_g - \kappa^2 E_g = -\Delta E_g - \kappa^2 E_g = 0 \text{ in } \mathbb{R}^3.$$

The fact that  $E_g$  satisfies the growth condition is obvious. For the proof of the reverse implication of the lemma, see Theorem 6.30 of [94].  $\square$

**Lemma 14.11** Suppose  $E_g$  is an electric Herglotz wave function. Then  $E_g = 0$  for all  $x$  if and only if  $g = 0$ .

**Proof** Following Colton and Kress [94], we note that each component of  $E_g$  satisfies the Helmholtz equation and

$$(E_g)_l = \int_{\partial B_1} g_l(d) \exp(i \kappa x \cdot d) dA(d), \quad l = 1, 2, 3.$$

Using the Funk–Hecke formula (9.54), we obtain

$$\int_{\partial B_1} g_l(d) Y_n^m(d) dA(d) = 0, \quad m = -n, \dots, n, \quad n = 0, 1, 2, \dots,$$

and the completeness of the spherical harmonics in  $L^2(\partial B_1)$  guaranteed by Lemma 9.11 shows that  $g_l = 0$ .  $\square$

We now state and prove a critical result for our upcoming theory. Indeed, this result is usually the most difficult result to prove when verifying the LSM in a particular situation. Our proof is taken from [95] with slight modifications to apply to Lipschitz domains.

**Theorem 14.12** *Let  $D$  be a bounded Lipschitz domain with connected complement. Then any function  $E \in H(\text{curl}; D)$  satisfying  $\nabla \times E - \kappa^2 E = 0$  in  $D$  (in the sense of distributions) can be approximated by an electric Herglotz wave function  $E_g$  to arbitrary accuracy in the  $H(\text{curl}; D)$  norm.*

**Remark 14.13** *The assumption that  $D$  has connected complement is essential. Indeed it is possible to show, via simple examples, that the theorem does not hold on domains for which the complement is disconnected. On the other hand, we allow the scatterer to be disconnected. Note also that the theorem holds even if  $\kappa$  is a Maxwell eigenvalue.*

**Proof of Theorem 14.12** Let

$$M(D) = \{E \in H(\text{curl}; D) \mid \nabla \times \nabla \times E - \kappa^2 E = 0 \text{ in } D\}.$$

We consider the map  $H_D: L_t^2(\partial B_1) \rightarrow M(D)$  defined, for  $g \in L_t^2(\partial B_1)$  by

$$(H_D g)(x) = \int_{\partial B_1} \exp(i \kappa x \cdot d) g(d) dA(d).$$

We want to show that  $H_D(L_t^2(\partial B_1))$  is dense in  $M(D)$ . This is done by characterizing a suitable adjoint operator  $H_D^*$  and showing  $H_D^*$  is injective. Once we have shown that  $H_D^*$  is injective, the result follows by Lemma 2.15.

In particular,  $H_D^*: M(D) \rightarrow L_t^2(\partial B_1)$  is defined by requiring that, if  $h \in M(D)$ , then

$$\langle H_D(g), h \rangle_{H(\text{curl}; D)} = \left( g, H_D^*(h) \right)_{L_t^2(\partial B_1)}$$

for all  $g \in L_t^2(\partial B_1)$ . We start by characterizing  $H_D^*$ . Using the definition of the  $H(\text{curl}; D)$  inner product and the integration by parts identity (3.51), we obtain

$$\begin{aligned} \langle H_D(g), h \rangle_{H(\text{curl}, D)} &= \int_D \int_{\partial B_1} (\exp(i \kappa x \cdot d) g \cdot \bar{h} \\ &\quad + \nabla_x \times (\exp(i \kappa x \cdot d) g(d)) \cdot \nabla \times \bar{h}(x)) dA(d) dV(x) \\ &= \int_{\partial B_1} g(d) \cdot \overline{\left\{ \int_D \exp(-i \kappa x \cdot d) (h(x) + \nabla \times \nabla \times h(x)) dV(x) \right\}} dA(d) \\ &\quad + \overline{\int_D \exp(-i \kappa x \cdot d) (v(x) \times \bar{g}(d)) \cdot \nabla \times h dA(x)} \Big\} dA(d). \end{aligned}$$

Here we have used the fact that  $\nabla \times \nabla \times \mathbf{b} = \kappa^2 \mathbf{b} \in (L^2(D))^3$  to justify the integration by parts. Since  $\nabla \times \nabla \times \mathbf{b} - \kappa^2 \mathbf{b} = 0$  in  $D$  in the usual weak sense, we have

$$\begin{aligned} \left( H_D^*(\mathbf{h}) \right)(\mathbf{g}) &= \left\{ d \times \left[ \int_D \exp(-i \kappa \mathbf{x} \cdot \mathbf{d}) (1 + \kappa^2) \mathbf{h} dV(\mathbf{x}) \right. \right. \\ &\quad \left. \left. - \int_{\Gamma} \exp(-i \kappa \mathbf{x} \cdot \mathbf{d}) \mathbf{v} \times \nabla \times \mathbf{h} dA(\mathbf{x}) \right] \right\} \times \mathbf{d}. \end{aligned}$$

Now from the proof of Corollary 9.5, we see that this is the far field pattern of (14.26)

$$\begin{aligned} V(\mathbf{y}) &= \frac{(1 + \kappa^2)}{\kappa^2} \nabla_{\mathbf{y}} \times \nabla_{\mathbf{y}} \times \int_D \Phi(\mathbf{y}, \mathbf{x}) \mathbf{h}(\mathbf{x}) dV(\mathbf{x}) \\ &= -\frac{1}{\kappa^2} \nabla_{\mathbf{y}} \times \nabla_{\mathbf{y}} \times \int_{\partial D} \Phi(\mathbf{y}, \mathbf{x}) \mathbf{v} \times (\nabla \times \mathbf{h}) dA(\mathbf{x}). \end{aligned}$$

Using the fact that  $\mathbf{h}$  satisfies Maxwell's equations in  $D$  and that  $\nabla_{\Gamma} \cdot (\nabla \times \nabla \times \mathbf{h}) = -\kappa^2 \mathbf{v} \cdot \mathbf{h}$  for  $\mathbf{h} \in M(D)$  (note that  $\mathbf{v} \cdot \mathbf{h}$  is well defined in  $H^{1/2}(\Gamma)$ , since  $\nabla \cdot \mathbf{h} = 0$  in  $D$ ), we obtain

$$\begin{aligned} V(\mathbf{y}) &= (1 + \kappa^2) \int_D \Phi(\mathbf{y}, \mathbf{x}) \mathbf{h} dV(\mathbf{x}) - \int_{\Gamma} \Phi(\mathbf{y}, \mathbf{x}) \mathbf{v}(\mathbf{x}) \times (\nabla \times \mathbf{h}) dA(\mathbf{x}) \\ &= -\frac{1}{\kappa^2} \nabla \int_{\Gamma} \Phi(\mathbf{y}, \mathbf{x}) \mathbf{v} \cdot \mathbf{h} dA(\mathbf{x}). \end{aligned}$$

This formula defines  $V$  for any  $\mathbf{y} \in \mathbb{R}^3 \setminus \bar{D}$ . Then, using the fact that  $\Phi$  is the fundamental solution of the Helmholtz equation (proceeding along the lines of the proof of Theorem 9.1), we see that the same formula defines  $V$  for  $\mathbf{y} \in D$ . Then, if  $V_+ = V|_{\mathbb{R}^3 \setminus \bar{D}}$  and  $V_- = V|_D$ , we have

$$\begin{aligned} \nabla \times \nabla \times V_+ - \kappa^2 V_+ &= 0 \quad \text{in } \mathbb{R}^3 \setminus \bar{D}, \\ \nabla \times \nabla \times V_- - \kappa^2 V_- &= (1 + \kappa^2) \mathbf{h} \quad \text{in } D. \end{aligned}$$

As  $\mathbf{y} \rightarrow \Gamma = \partial D$ , the integrand on  $\Gamma$  in the definition of  $V$  becomes singular. This results in the classical jump properties relating traces of  $V_+$  to  $V_-$  on  $\Gamma$ . Since this is the only place we shall use such properties we have not discussed them in detail. However, they are well known (see, e.g., [94]) for smooth boundaries.

and the same relations also hold for Lipschitz boundaries (see [215, 62, 64]). In particular, it can be shown that

$$\left. \begin{aligned} \mathbf{v} \times V_+ &= \mathbf{v} \times V_-, \\ \mathbf{v} \times \nabla \times V_+ &= \mathbf{v} \times \nabla \times V_- - \mathbf{v} \times \nabla \times h, \end{aligned} \right\} \text{on } \Gamma.$$

Now suppose  $H^*_D b = 0$ . Since  $H^*_D b$  is the far field pattern of  $V_+$ , this implies, by Corollary 9.29 of Rellich's lemma and the unique continuation result in Theorem 4.13, that  $V_+ = 0$ . But using the equations satisfied by  $b$  and  $V_-$  and the integration by parts result (3.51), we have

$$\begin{aligned} \|h\|_{H(\text{curl}; D)}^2 &= \int_D (\nabla \times h \cdot \nabla \times \bar{h} + h \cdot \bar{h}) dV \\ &= \int_D (\nabla \times \nabla \times h \cdot \bar{h} + h \cdot \bar{h}) dV + \int_\Gamma \nabla \times h \cdot \mathbf{v} \times \bar{h} dA \\ &= (1 + \kappa^2) \int_D h \cdot \bar{h} dV - \int_\Gamma \bar{h} \cdot \mathbf{v} \times \nabla \times h dA \\ &= \int_D (\nabla \times \nabla \times V_- - \kappa^2 V_-) \cdot \bar{h} dV - \int_\Gamma \bar{h} \cdot \mathbf{v} \times \nabla \times h dA \\ &= \int_D (\nabla \times V_- \cdot \nabla \times \bar{h} - \kappa^2 V_- \cdot \bar{h}) dV + \int_\Gamma \bar{h} \cdot \mathbf{v} \times \nabla \times V_+ dA \\ &= \int_D V_- \cdot (\nabla \times \nabla \times \bar{h} - \kappa^2 \bar{h}) dV + \int_\Gamma \mathbf{v} \times V_+ \cdot \nabla \times \bar{h} dA = 0, \end{aligned}$$

where we have used the fact that  $\mathbf{v} \times V_+ = 0$  and  $\mathbf{v} \times \mathbf{v} \times V_+ = 0$  on  $\Gamma$ . Thus,  $b = 0$  and so  $H^*_D$  is injective. By Lemma 2.15 we have thus proved  $\overline{H_D(L_t^2(\partial B_1))} = M(D)$ . This completes the proof.  $\square$

### 14.3.3 The far field operators F and B

Now we study the far field operator  $F$  and a related operator  $B$  which we shall define shortly. Of course,  $F$  is defined by (14.4). Our first task is to verify that the far field operator is injective with dense range. This implies that the Tikhonov regularization is an appropriate procedure for solving (14.3). Again the fact that  $D$  has connected complement is a critical assumption.

**Theorem 14.14** Suppose  $\kappa$  is not an interior Maxwell eigenvalue for  $D$ . Then the far field operator  $F: L_t^2(\partial B_1) \rightarrow L_t^2(\partial B_1)$  is injective with dense range.

**Proof** Suppose  $F_g = 0$ , then

$$\int_{\partial B_1} E_\infty(\hat{x}, d, g(d)) dA(d) = 0 \text{ for all } \hat{x} \in \partial B,$$

where  $E_\infty$  is the far field pattern of the solution  $E \in H_{\text{loc}}(\text{curl}; \mathbb{R}^3 \setminus \overline{D})$  satisfying (14.18)–(14.21). The left-hand side is the far field pattern of the scattered field due to the incoming wave given by the Herglotz wave function

$$E_g^i = \int_{\partial B_1} E^i(x, d, g(d)) dA(d) = i \kappa \int_{\partial B_1} g(d) \exp(i \kappa x \cdot d) dA(d).$$

This follows by superposition of the solutions of the scattering problem. Since the far field vanishes, we conclude that, due to Corollary 9.29 of Rellich's lemma and the unique continuation result in Theorem 4.13, the scattered field due to this incident field vanishes identically outside  $D$ . Hence  $E_g^i$  satisfies

$$\begin{aligned} \nabla \times \nabla \times E_g^i - \kappa^2 E_g^i &= 0 \quad \text{in } D, \\ v \times E_g^i &= 0 \quad \text{on } \Gamma. \end{aligned}$$

The second equation follows from the perfect conducting boundary condition on  $\Gamma$ . Since  $\kappa$  is not an interior Maxwell eigenvalue,  $E_g^i \neq 0$ . Hence, by Lemma 14.11,  $g = 0$  and injectivity is proved.

Now we prove that  $F$  has dense range. Let  $F^*: L_t^2(\partial B_1) \rightarrow L_t^2(\partial B_1)$  denote the adjoint to  $F$ . By definition,

$$\begin{aligned} (Fg, h)_{L_t^2(\partial B_1)} &= \int_{\partial B_1} \int_{\partial B_1} \exp(i \kappa x \cdot d) g(d) \cdot \bar{h}(x) dA(d) dA(x) \\ &= \int_{\partial B_1} g(d) \cdot \overline{\left( \int_{\partial B_1} \exp(-i \kappa x \cdot d) h(x) dA(x) \right)} dA(d). \end{aligned}$$

Thus,  $(F^*h)(d) = (F\tilde{h})(-d)$ , where  $\tilde{h}(\hat{x}) = \overline{h(-\hat{x})}$ . Since  $F$  is injective, it follows that  $F^*$  is injective, so that Lemma 2.15 shows that  $F(L_t^2(\partial B_1))$  is dense in  $(L_t^2(\partial B_1))^\perp$ .  $\square$

In fact, the proof has shown a slightly stronger result than that stated in the theorem. We see that  $F$  is injective with dense range unless  $\kappa$  is a Maxwell eigenvalue of  $D$  and the corresponding eigenfunction is a Herglotz wave function.

We next show that the reciprocity result used in the formulation of the numerical scheme (see the derivation of (14.13)) is valid.

**Theorem 14.15** *The electric far field pattern due to the scattering of plane wave of the form (14.1) by a perfect conductor satisfies the reciprocity relation*

$$q \cdot E_\infty(\hat{x}, d, p) = p \cdot E_\infty(-d, -\hat{x}, q)$$

for all  $\mathcal{O}$ ,  $d \in \partial B_1$  and all  $q, p \in \mathbb{R}^3$ .

**Remark 14.16** Colton and Kress [94] point out that this result holds for imperfect conductors as well.

**Proof of Theorem 14.15** From Corollary 9.5, we see that

$$\begin{aligned} 4\pi p \cdot E_\infty(\hat{x}, d, p) &= \int_{\Gamma} \left( v \times E^s(y, d, p) \cdot H^i(y, -\hat{x}, q) \right. \\ &\quad \left. + v \times H^s(y, d, p) \cdot E^i(y, -\hat{x}, q) \right) dA(y) \end{aligned}$$

where  $H^i = (1/i\kappa)^\vee \times E^i$  and  $H^s = (1/i\kappa)^\vee \times E^s$ .

Similarly,

$$4\pi p \cdot E_\infty(-d, -\hat{x}, q) = \int_{\Gamma} \left( v \times E^s(y, x, q) \cdot H^i(y, d, p) + v \times H^s(y, \hat{x}, q) \cdot E^i(y, d, p) \right) dA(y).$$

Subtracting these expressions and adding and subtracting suitable combinations of incident fields, we obtain

$$\begin{aligned} & 4\pi(q \cdot E_\infty(\hat{x}, d, p) - p \cdot E_\infty(-d, -\hat{x}, q)) \\ &= \int_{\Gamma} \left( v \times E(y, d, p) \cdot H^i(y, \hat{x}, q) + v \times H(y, d, p) \cdot E^i(y, -\hat{x}, q) \right) dA(y) \\ & - \int_{\Gamma} \left( v \times E^i(y, d, p) \cdot H^i(y, -x, q) + v \times H^i(y, d, p) \cdot E^i(y, -\hat{x}, q) \right) dA(y) \\ & - \int_{\Gamma} \left( v \times E^s(y, -\hat{x}, q) \cdot H(y, d, p) + v \times H^s(y, -\hat{x}, q) \cdot E^i(y, d, p) \right) dA(y) \\ & + \int_{\Gamma} \left( v \times E^s(y, -\hat{x}, q) \cdot H^s(y, d, p) + v \times E^s(y, -\hat{x}, q) \cdot H^s(y, d, p) \right) dA(y) \end{aligned}$$

The second and fourth integral vanish. To see that the second integral vanishes, we can use (3.51) and the fact that both  $E^i$  and  $H^i$  are smooth solutions of Maxwell's equations in  $D$ . A similar argument in  $\mathbb{R}^3 \setminus \bar{D}$  shows that the fourth integral vanishes. In this case, it is necessary to use an argument like that in the proof of Theorem 9.2 (via the radiation condition) to show that there is no contribution from infinity.

Combining the remaining integrals shows that

$$\begin{aligned} & 4\pi(q \cdot E_\infty(\hat{x}, d, p) - p \cdot E_\infty(-d, -\hat{x}, q)) \\ &= \int_{\Gamma} \{ v \times E(y, d, p) \cdot H(y, -\hat{x}, q) + v \times H(y, d, p) \cdot E(y, -\hat{x}, q) \} dA(y). \end{aligned}$$

The fact that  $v \times E = 0$  on  $\Gamma$  and hence  $E_\tau = 0$  on  $\Gamma$  shows that the right hand side of this equation vanishes and completes the proof.  $\square$

Now we need some results concerning the far field operator  $B$  which we define next. For  $\lambda \in Y(\Gamma)$  let  $E_\lambda \in H_{loc}(\text{curl}; \mathbb{R}^3 \setminus \bar{D})$  satisfy (14.27)

$$\begin{aligned} \nabla \times \nabla \times E_\lambda - \kappa^2 E_\lambda &= 0 \text{ in } \mathbb{R}^3 \setminus \bar{D}, \\ v \times E_\lambda &= \lambda \text{ on } \Gamma, \end{aligned} \tag{14.28}$$

$$\lim_{\rho \rightarrow \infty} \rho (\nabla \times E_\lambda \times \hat{x} - i \kappa E_\lambda) = 0. \tag{14.29}$$

Then we define  $\mathcal{B}: Y(\Gamma) \rightarrow (L^2(\partial B_1))$  by  $B\lambda = E_{\lambda, \infty}$ . Thus,  $B$  maps boundary data to the corresponding far field pattern.

**Lemma 14.17** *The operator  $B$  is a bounded, compact and linear map.*

**Proof** The operator  $B$  is the composition of two operators. The first is the operator mapping  $\lambda$  to the solution  $E$  in the neighborhood of  $\Gamma$  (bounded as a map from  $Y(\Gamma)$  into  $X$  by Theorem 10.8). The second operator maps this data to the far field using the integral representation in (13.80) (choosing  $v = 0$  in a neighborhood of  $\Sigma$ ). This is clearly bounded and compact and hence the composition of these two functions yields the conclusion of this lemma.  $\square$

**Lemma 14.18** *The operator  $B$  is injective with dense range.*

**Proof** The injectivity of  $B$  is implied by Corollary 9.29 of Rellich's lemma and the unique continuation result in Theorem 4.13.

To prove density, we will show that the dual operator  $B^T: L_t^2(\partial B_1) \rightarrow Y(\Gamma)$  is injective. The dual is defined by(14.30)

$$\langle\langle B(\lambda), g \rangle\rangle_{\partial B_1} = \left\langle \left\langle \lambda, B^T(g) \right\rangle \right\rangle_{\Gamma}$$

for all  $g \in L_t^2(\partial B_1)$  and  $\lambda \in Y(\partial D)$  where  $\langle\langle \cdot, \cdot \rangle\rangle_{\partial B_1}$  represents the bilinear form on  $L_t^2(\partial B_1) \times L_t^2(\partial B_1)$  given by

$$\langle\langle \lambda, \mu \rangle\rangle_{\partial B_1} = \int_{\partial B_1} \lambda \cdot \mu dA$$

and  $\langle\langle \cdot, \cdot \rangle\rangle_{\Gamma}$  is the corresponding duality pairing on  $\Gamma$ .

From the representation for  $E_\infty$  in Corollary 9.5, we have(14.31)

$$\begin{aligned} B(\lambda) &= \frac{i\kappa}{4\pi} \hat{x} \times \int_{\Gamma} \left[ v \times E_{\lambda} + \frac{1}{i\kappa} (v \times \nabla \times E_{\lambda}) \times \hat{x} \right] \\ &\quad \cdot \exp(-i\kappa \hat{x} \cdot y) dA(y), \end{aligned}$$

where  $E_\lambda$  solves the scattering problem (14.27)–(14.29). Because  $E_\lambda$  satisfies Maxwell's equations in the exterior of  $D$ , we know that  $v \times v \times E_\lambda$  is well defined in  $Y(\Gamma)$ .

Interchanging order of integration in (14.31), we have(14.32)

$$\begin{aligned} \langle\langle B(\lambda), g \rangle\rangle_{\partial B_1} &= \frac{i\kappa}{4\pi} \int_{\Gamma} \int_{\partial B_1} \exp(-i\kappa \hat{x} \cdot y) g(\hat{x}) \\ &\quad \cdot \left\{ \hat{x} \times (v \times E_{\lambda}) + \frac{1}{i\kappa} (\hat{x} \times (v \times \nabla \times E_{\lambda}) \times \hat{x}) \right\} dA(\hat{x}) dA(y). \end{aligned}$$

Now let us define a Herglotz wave function  $E_g$  by

$$E_g(y) = \int_{\partial B_1} g(\hat{x}) \exp(-i\kappa \hat{x} \cdot y) dA(\hat{x})$$

and note that, since

$$\nabla \times E_g = i\kappa \int_{\partial B_1} (g(\hat{x}) \times \hat{x}) \exp(-i\kappa \hat{x} \cdot y) dA(\hat{x}),$$

we have

$$\int_{\partial B_1} \exp(-i \kappa \hat{x} \cdot y) g(x) \cdot \hat{x} \times (\nu \times E_\lambda) dA(\hat{x}) = \frac{1}{i \kappa} (\nabla \times E_g) \cdot \nu \times E_\lambda.$$

Also, since  $g(x^\wedge) \cdot x^\wedge = 0$ ,

$$\int_{\partial B_1} \exp(-i \kappa \hat{x} \cdot y) g(\hat{x}) \cdot (\hat{x} \times (\nu \times \nabla \times E_\lambda)) \times \hat{x} dA(\hat{x}) = E_g \cdot \nu \times \nabla \times E_\lambda.$$

Putting these equations in (14.32), we obtain the formula

$$\langle \langle B(\lambda), g \rangle \rangle_{\partial B_1} = \frac{1}{4\pi} \int_{\Gamma} [(\nu \times \nabla \times E_\lambda) \cdot E_g - (\nu \times \nabla \times E_g) \cdot E_\lambda] dA.$$

Now define  $\tilde{E} \in H_{loc}(\text{curl}; \mathbb{R}^3 \setminus D^-)$  to satisfy

$$\begin{aligned} \nabla \times \nabla \times \tilde{E} - \kappa^2 \tilde{E} &= 0 \quad \text{in } \mathbb{R}^3 \setminus \bar{D}, \\ \nu \times \tilde{E} &= \nu \times E_g \quad \text{on } \Gamma, \\ \lim_{\rho \rightarrow \infty} \rho (\nabla \times \tilde{E} \times \hat{x} - i \kappa \bar{E}) &= 0. \end{aligned}$$

Then we may write

$$\langle \langle B(\lambda), g \rangle \rangle_{\partial B_1} = \frac{1}{4\pi} \int_{\Gamma} (\nu \times \nabla \times E_\lambda) \cdot \tilde{E} - (\nu \times \nabla \times E_g) \cdot E_\lambda dA.$$

Using Green's formula (3.51) and arguments similar to those used to prove the Stratton–Chu formula in Theorem 9.2, we can show that since  $E_\lambda$  and  $\tilde{E}$  are both radiating solutions of Maxwell's equations, we have

$$\int_{\Gamma} (\nu \times \nabla \times E_\lambda) \cdot \tilde{E} - (\nu \times \nabla \times \tilde{E}) \cdot E_\lambda dA = 0.$$

Hence,

$$\begin{aligned} \langle \langle B(\lambda), g \rangle \rangle_{\partial B_1} &= \frac{1}{4\pi} \int_{\Gamma} (\nu \times E_\lambda) \cdot (\nabla \times E_g - \nabla \times \tilde{E}) dA \\ &= \frac{1}{4\pi} \int_{\Gamma} \lambda \cdot (\nabla \times E_g - \nabla \times \tilde{E}) dA, \end{aligned}$$

and we have shown that

$$4\pi B^\top(g) = (\nabla \times E_g - \nabla \times \bar{E})_T \in Y(\Gamma)'.$$

Now suppose  $B^\top(g) = 0$ . Then

$$\left. \begin{aligned} \nu \times E_g &= \nu \times \tilde{E} \\ \nu \times (\nabla \times E_g) &= \nu \times (\nabla \times \tilde{E}) \end{aligned} \right\} \text{on } \Gamma.$$

Since  $E_g$  is a solution of Maxwell's equation in  $D$  and  $\tilde{E}$  is a solution in  $\mathbb{R}^3 \setminus D^-$ , these relations imply that  $\tilde{E}$  can be extended by  $E_g$  into  $D$  and the extended function satisfies the homogeneous Maxwell's equation in all of  $\mathbb{R}^3$  together with the radiation condition. Hence, by Corollary 9.29 of Rellich's lemma and the unique continuation result in Theorem 4.13, we have  $\tilde{E} = 0$  and so  $E_g = 0$ . This implies, by Lemma 14.11, that  $g = 0$ . Hence, we have proved that  $B^\top$  is injective. The density result of this theorem now follows from Lemma 2.15.  $\square$

It will be useful in the next section to note that we may rewrite the far field equation (14.3) as

$$B(\gamma_t E_g) = \frac{1}{i\kappa} E_{e,\infty},$$

where  $\gamma E_g = v \times E_g$  on  $\Gamma$ .

#### 14.3.4 Mathematical justification of the LSM

Here we prove our main theorem justifying the LSM.

**Theorem 14.19** *Assume that  $\kappa$  is not a Maxwell eigenvalue for  $D$ . Then if  $F$  is the far field operator (14.4) corresponding to the perfectly conducting scattering problem (14.18)–(14.21), we have:*

(1) *If  $\zeta \in D$ , then for every  $\epsilon > 0$  there exists a solution  $g_\epsilon(\cdot, z, p) \in L_t^2(\partial B_1)$  satisfying the inequality*

$$\| Fg_\epsilon(\cdot, z, p) - E_{e,\infty}(\cdot, z, p) \|_{L_t^2(\partial B_1)} < \epsilon.$$

Moreover, this solution satisfies

$$\lim_{z \rightarrow \Gamma} \| E_{g_\epsilon}(\cdot, z, p) \|_{H(\text{curl}; D)} = \infty, \quad \text{and} \quad \lim_{z \rightarrow \Gamma} \| g_\epsilon(\cdot, z, p) \|_{L_t^2(\partial B_1)} = \infty.$$

where  $E_g(\cdot, z)$  is the electric Herglotz wave function with kernel  $g$ .

(2) *If  $\zeta \in \mathbb{R}^3 \setminus D^-$ , then for every  $\epsilon > 0$  and  $\delta > 0$  there exists a solution  $g_{\delta,\epsilon}(\cdot, z, p) \in L_t^2(\partial B_1)$  such that*

$$\| Fg_{\delta,\epsilon}(\cdot, z, p) - E_{e,\infty}(\cdot, z, p) \|_{L_t^2(\partial B_1)} < \epsilon + \delta,$$

and, in addition,

$$\lim_{\delta \rightarrow 0} \| E_{g_{\delta,\epsilon}}(\cdot, z, p) \|_{H(\text{curl}; D)} = \infty, \quad \text{and} \quad \lim_{\delta \rightarrow 0} \| g_{\delta,\epsilon}(\cdot, z, p) \|_{L_t^2(\partial B_1)} = \infty,$$

where  $E_{g_{\delta,\epsilon}}(\cdot, z, p)$  is the electric field of the electromagnetic Herglotz pair with kernel  $g_{\delta,\epsilon}$ .

**Remark 14.20** *It is important to verify that both the norms of  $g$  and  $E_g$  blow up as  $\zeta$  approaches the boundary.*

**Proof of Theorem 14.19** First, let  $\zeta \in D$ . In this case  $E_{\infty}(\cdot, z, p)$  is in the range of  $B$  since it is the far field pattern of the electric dipole  $E_e(x, z, p)$  which is the solution of the exterior mixed boundary problem (14.18)–(14.21)

with incoming wave  $E^i = E_e|_{\Gamma}$ . Let  $E \in H(\text{curl}; D)$  be the weak solution of the interior boundary value problem

$$\begin{aligned} \nabla \times \nabla \times E - \kappa^2 E &= 0 \quad \text{in } D, \\ v \times E &= v \times E_e \quad \text{on } \Gamma. \end{aligned}$$

From Theorem 14.12 and the definition of  $Y(\Gamma)$  for every  $\epsilon > 0$  there is a  $g_\epsilon(\cdot, z, p) \in L_t^2(\partial B_1)$  such that the corresponding electric Herglotz function  $E_{g_\epsilon}(\cdot, z, p)$  satisfies

$$\| Y_t(E - E_{g_\epsilon}(\cdot, z, p)) \|_{Y(\Gamma)} < \epsilon.$$

The continuity of the operator  $B$  (see Theorem 14.17) and the fact that  $\gamma_i E = \gamma_i E_e$  implies that(14.33)

$$\| B(Y_t E_{g_\epsilon}(\cdot, z, p)) - \frac{1}{i\kappa} E_{e,\infty}(\cdot, z, p) \|_{L_t^2(\partial B_1)} < C\epsilon$$

for some positive constant  $C$ . Furthermore, if  $\zeta \rightarrow \Gamma$  then

$$\| E_e(\cdot, z, p) \|_{H(\text{curl}; (\mathbb{R}^3 \setminus \bar{D}) \cap B_R)} \rightarrow \infty.$$

The well-posedness of the exterior perfectly conducting boundary value problem implies

$$\lim_{z \rightarrow \Gamma} \| Y_t E_e \|_{Y(\Gamma)} \rightarrow \infty, \quad \text{and so} \quad \lim_{z \rightarrow \Gamma} \| Y_t E_{g_\epsilon}(\cdot, z, p) \|_{Y(\Gamma)} \rightarrow \infty.$$

Hence the kernel and the corresponding electric Herglotz function blow up in norm as  $\zeta \rightarrow \Gamma$ .

Now let  $\zeta \in \mathbb{R}^3 \setminus D^-$ . For these points  $E_{e,\infty}(\cdot, \zeta, p)$  is not in the range of  $B$ . To see this, suppose it is in the range of  $B$ . Then, due to Rellich's lemma and unique continuation, the field due to electric dipole  $E_e(x, \zeta, p)$  has to be a solution to Maxwell's equation in  $\mathbb{R}^3 \setminus D^-$  which is not possible since it has a singularity at  $\zeta$ .

However from Theorem 14.18, using Tikhonov regularization, we can construct a regularized solution to the far field equation (14.3). In particular, if  $f_z^\alpha \in Y(\Gamma)$  is the regularized solution of

$$B(f_z) = -\frac{1}{i\kappa} E_{e,\infty}(\cdot, z, p)$$

corresponding to the regularization parameter  $\alpha$  chosen by the Morozov discrepancy principle we may choose  $\alpha$  small enough so that(14.34)

$$\| B(f_z^\alpha) - \frac{1}{i\kappa} E_{e,\infty}(\cdot, z, p) \|_{L_t^2(\partial B_1)} < \delta,$$

for an arbitrary small  $\delta > 0$ . In addition, because  $E_{e,\infty}$  is not in the range of  $B$ ,

(14.35)

$$\lim_{a \rightarrow 0} \| f_z^a \|_{Y(\Gamma)} = \infty.$$

Using Theorem 14.18 and the continuity of the operator  $B$ , we can find an electric Herglotz function  $E_{g_{a,\epsilon}}(\cdot, z, p)$  with  $g_{a,\epsilon}(\cdot, z, p) \in L_t^2(\partial B_1)$  such that (14.36)

$$\| B(Y_t E_{g_{a,\epsilon}(\cdot, z, p)}) - B(f_z^a) \|_{L_t^2(\partial B_1)} < \epsilon.$$

Now combining (14.34) and (14.36), we obtain

$$\| B(Y_t E_{g_{a,\epsilon}(\cdot, z, p)}) - \frac{1}{i\kappa} E_{e,\infty}(\cdot, z, p) \|_{L_t^2(\partial B_1)} < \epsilon + \delta.$$

Furthermore, since  $\gamma_t E_{g_{a,\epsilon}(\cdot, z, p)}$  approximates  $f_z^a$  in  $Y(\Gamma)$ , (14.35) implies that

$$\lim_{a \rightarrow 0} \| \gamma_t E_{g_{a,\epsilon}(\cdot, z, p)} \|_{Y(\Gamma)} = \infty, \text{ and so } \lim_{a \rightarrow 0} \| g_{a,\epsilon}(\cdot, z, p) \|_{L_t^2(\partial B_1)} = \infty.$$

# APPENDIX A COORDINATE SYSTEMS

## A.1 Cartesian coordinates

(1) Unit vectors:

$$e_1 = (1, 0, 0)^\top, e_2 = (0, 1, 0)^\top, e_3 = (0, 0, 1)^\top.$$

$$(2) \boldsymbol{x} = (x_1, x_2, x_3)^\top = x_1 e_1 + x_2 e_2 + x_3 e_3.$$

(3) Gradient:

$$\nabla p = \frac{\partial p}{\partial x_1} e_1 + \frac{\partial p}{\partial x_2} e_2 + \frac{\partial p}{\partial x_3} e_3.$$

(4) Divergence:

$$\nabla \cdot \boldsymbol{v} = \frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} + \frac{\partial v_3}{\partial x_3}.$$

(5) Curl:

$$\nabla \times \boldsymbol{v} = \left( \frac{\partial v_3}{\partial x_2} - \frac{\partial v_2}{\partial x_3} \right) e_1 - \left( \frac{\partial v_3}{\partial x_1} - \frac{\partial v_1}{\partial x_3} \right) e_2 + \left( \frac{\partial v_2}{\partial x_1} - \frac{\partial v_1}{\partial x_2} \right) e_3.$$

(6) Laplacian:

$$\Delta p = \frac{\partial^2 p}{\partial x_1^2} + \frac{\partial^2 p}{\partial x_2^2} + \frac{\partial^2 p}{\partial x_3^2}.$$

## A.2 Spherical coordinates

(1) Unit vectors:

$$e_\rho = \sin \theta \cos \varphi e_1 + \sin \theta \sin \varphi e_2 + \cos \theta e_3,$$

$$e_\theta = \cos \theta \cos \varphi e_1 + \cos \theta \sin \varphi e_2 - \sin \theta e_3,$$

$$e_\varphi = -\sin \varphi e_1 + \cos \varphi e_2.$$

$$(2) \boldsymbol{x} = \rho \sin \theta \cos \varphi e_1 + \rho \sin \theta \sin \varphi e_2 + \rho \cos \theta e_3.$$

(3) Gradient:

$$\nabla p = \frac{\partial p}{\partial \rho} e_\rho + \frac{1}{\rho} \frac{\partial p}{\partial \theta} e_\theta + \frac{1}{\rho \sin \theta} \frac{\partial p}{\partial \varphi} e_\varphi.$$

(4) Divergence:

$$\nabla \cdot \boldsymbol{v} = \frac{1}{\rho^2} \frac{\partial}{\partial \rho} \left( \rho^2 v_\rho \right) + \frac{1}{\rho \sin \theta} \frac{\partial}{\partial \theta} (\sin(\theta) v_\theta) + \frac{1}{\rho \sin \theta} \frac{\partial v_\varphi}{\partial \varphi}.$$

(5)  $\text{Curl}:(A.1)$ 

$$\begin{aligned}\nabla \times v = & -\frac{1}{\rho \sin \theta} \left( \frac{\partial}{\partial \theta} (\sin \theta v_\varphi) - \frac{\partial v_\theta}{\partial \varphi} \right) e_\rho \\ & + \frac{1}{\rho} \left( \frac{1}{\sin \theta} \frac{\partial v_\rho}{\partial \varphi} - \frac{\partial}{\partial \rho} (\rho v_\theta) \right) e_\theta \\ & + \frac{1}{\rho} \left( \frac{\partial}{\partial \rho} (\rho v_\theta) - \frac{\partial v_\rho}{\partial \theta} \right) e_\varphi.\end{aligned}$$

(6) Laplacian:

$$\Delta p = \frac{1}{\rho^2} \frac{\partial}{\partial \rho} \left( \rho^2 \frac{\partial p}{\partial \rho} \right) + \frac{1}{\rho^2 \sin \theta} \frac{\partial}{\partial \theta} \left( \sin(\theta) \frac{\partial p}{\partial \theta} \right) + \frac{1}{\rho^2 \sin^2 \theta} \frac{\partial^2 p}{\partial \varphi^2}.$$

# APPENDIX B VECTOR AND DIFFERENTIAL IDENTITIES

## B.1 Vector identities

$$(1) \quad \mathbf{a} \times \mathbf{b} = -\mathbf{b} \times \mathbf{a}.$$

$$(2) \quad \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} = (\mathbf{c} \times \mathbf{a}) \cdot \mathbf{b}.$$

## B.2 Differential identities

These differential identities are valid for smooth functions/vector functions:(B.1)

$$\begin{aligned} \nabla \times (\nabla p) &= 0, \\ \nabla \cdot (\nabla \times \mathbf{v}) &= 0, \end{aligned} \tag{B.2}$$

$$\nabla \cdot (\varphi \mathbf{v}) = \nabla \varphi \cdot \mathbf{v} + \varphi \nabla \cdot \mathbf{v}, \tag{B.3}$$

$$\nabla \times (\varphi \mathbf{v}) = \varphi \nabla \times \mathbf{v} + (\nabla \varphi) \times \mathbf{v}, \tag{B.4}$$

$$\nabla \times (\mathbf{u} \times \mathbf{v}) = \mathbf{u}(\nabla \cdot \mathbf{v}) - (\mathbf{u} \cdot \nabla)\mathbf{v} + (\mathbf{v} \cdot \nabla)\mathbf{u} - \mathbf{v}(\nabla \cdot \mathbf{u}), \tag{B.5}$$

$$\nabla \times (\nabla \times \mathbf{u}) = \nabla(\nabla \cdot \mathbf{u}) - \Delta \mathbf{u}, \tag{B.6}$$

$$\nabla \cdot (\mathbf{u} \times \mathbf{v}) = \mathbf{v} \cdot \nabla \times \mathbf{u} - \mathbf{u} \cdot \nabla \times \mathbf{v}, \tag{B.7}$$

$$\nabla \times \nabla \times \{x \mathbf{u}(x)\} = -x \Delta \mathbf{u}(x) + \nabla \left\{ \mathbf{u}(x) + \rho \frac{\partial \mathbf{u}}{\partial \rho}(x) \right\}. \tag{B.8}$$

In the (B.6) and (B.8),  $\Delta \mathbf{u} = (\Delta u_1, \Delta u_2, \Delta u_3)$  in Cartesian coordinates only.

## B.3 Differential identities on a surface

Let  $S$  be a smooth surface with unit normal  $\mathbf{v}$  and let  $v$  and  $p$  be smooth functions defined a neighborhood of  $S$ . The following identities hold:

$$\begin{aligned} \nabla_S p &= (\mathbf{v} \times \nabla p|_S) \times \mathbf{v}, \\ -\nabla_S \times p &= -\mathbf{v} \times \nabla_S p, \\ \nabla_S \times \mathbf{v} &= -\nabla_S \cdot (\mathbf{v} \times \mathbf{v}), \\ \nabla_S \cdot \mathbf{v} &= \nabla_S \times (\mathbf{v} \times \mathbf{v}), \\ \nabla_S \cdot (\mathbf{v} \times \mathbf{v}) &= -\mathbf{v} \cdot (\nabla \times \mathbf{v})|_S. \end{aligned}$$

The differential equalities in this and the previous subsection can be extended to less smooth functions as discussed in the text.

# REFERENCES

- [1] Abboud, T. and Nédélec, J.C. (1992). Electromagnetic waves in an inhomogeneous medium. *J. Math. Anal. Appl.*, **164**, 40–58.
- [2] Adams, R.A. (1975). *Sobolev Spaces*, Volume 65 of *Pure and Applied Mathematics*. Academic Press, New York.
- [3] Ahagon, A., Fujiwara, K., and Nakata, T. (1996). Comparison of various kinds of edge elements for electromagnetic field analysis. *IEEE Trans. Mag.*, **32**, 898–901.
- [4] Ainsworth, M. and Coyle, J. (2001). Computation of Maxwell eigenvalues on curvilinear domains using  $hp$ -version nédélec elements. To appear in Proceedings of ENUMATH 2001.
- [5] Ainsworth, M. and Coyle, J. (2001). Hierarchic  $hp$ -edge element families for Maxwell's equations on hybrid quadrilateral/triangular meshes. *Comput. Meth. Appl. Mech. Eng.*, **190**, 6709–733.
- [6] Ainsworth, M. and Coyle, J. (2002). Conditioning of hierarchic  $p$ -version nédélec elements on meshes of curvilinear quadrilaterals and hexahedra. Accepted for publication in SIAM J. Numer. Anal.
- [7] Ainsworth, M. and Coyle, J. (2002). Hierarchic finite element bases on unstructured tetrahedral meshes. In press.
- [8] Alonso, A. and Valli, A. (1996). Some remarks on the characterization of the space of tangential traces of  $H(\text{rot}; \Omega)$  and the construction of an extension operator. *Manuscripta Mathematica*, **89**, 159–178.
- [9] Alonso, A. and Valli, A. (1999). An optimal domain decomposition preconditioner for low-frequency time-harmonic Maxwell equations. *Math. Comput.*, **68**, 607–31.
- [10] Ammari, H. and Bao, G. (2002). Maxwell's equations in a perturbed periodic structure. *Adv. Comput. Math.*, **16**, 99–112.
- [11] Ammari, H. and Nédélec, J.C. (1999). Generalized impedance boundary conditions for the Maxwell equations as singular perturbations problems. *Communi. Partial Diff. Eqns*, **24**, 821–49.
- [12] Amrouche, C., Bernardi, C., Dauge, M., and Girault, V. (1998). Vector potentials in three-dimensional nonsmooth domains. *Math. Meth. Appl. Sci.*, **21**, 823–64.
- [13] Andersen, L.S. and Volakis, J.L. (1998). Hierarchical tangential vector finite elements for tetrahedra. *IEEE Microwave and Guided Wave Letters*, **8**, 127–9.
- [14] Andersen, L.S. and Volakis, J.L. (1999). Condition numbers of various FEM matrices. *Journal of Electromagnetic Waves and Applications*, **13**, 1663–79.

- [15] Angell, T.S. and Kirsch, A. (1992). The conductive boundary condition for Maxwell's equations. *SIAM J. Appl. Math.*, **52**, 1597–610.
- [16] Anselone, P.M. (1971). *Collectively Compact Operator Approximation Theory*. Prentice-Hall, Englewood Cliffs.
- [17] Arnold, D., Boffi, D., Falk, R.S., and Gastaldi, L. (2001). Finite element approximation on quadrilateral meshes. *Commun. Numer. Meth. Eng.*, **17**, 805–12.
- [18] Arnold, D.N., Falk, R.S., and Winther, R. (2000). Multigrid in  $H(\text{div})$  and  $H(\text{curl})$ . *Numer. Math.*, **85**, 197–217.
- [19] Assous, F. and Ciarlet Jr., P. (1997). Une caractérisation de l'orthogonal de  $\nabla(H^2(\Omega) \cap H_0^1(\Omega))$  dans  $L^2(\Omega)$ . *C. R. Acad. Sci. Paris, Série 1*, **325**, 605–10.
- [20] Assous, F., Ciarlet Jr., P., Raviart, P.A., and Sonnendrücker, E. (1999). Characterization of the singular part of the solution of Maxwell's equations in a polyhedral domain. *Math. Meth. Appl. Sci.*, **22**, 485–99.
- [21] Babuška, I. and Miller, A. (1984). The post-processing approach in the finite element method – Part 1: Calculation of displacements, stresses and other higher derivatives of the displacements. *Int. J. Numer. Meth. Eng.*, **20**, 1085–109.
- [22] Babuška, I. and Miller, A. (1984). The post-processing approach in the finite element method – Part 2: The calculation of stress intensity factors. *Int. J. Numer. Meth. Eng.*, **20**, 1110–29.
- [23] Babuška, I. and Miller, A. (1984). The post-processing approach in the finite element method – Part 3: A posteriori error estimates and adaptive mesh selection. *Int. J. Numer. Meth. Eng.*, **20**, 2311–24.
- [24] Babuška, I. and Aziz, A.K. (1972). Survey lectures on the mathematical foundations of the finite element method. In *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations* (ed. A. Aziz), pp. 5–359. Academic Press, New York.
- [25] Bao, G. and Dobson, D.C. (2000). On the scattering by a biperiodic structure. *Proc. Am. Math. Soc.*, **128**, 2715–23.
- [26] Barton, M.L. and Cendes, Z.J. (1987). New vector finite elements for three-dimensional magnetic computation. *J. Appl. Phys.*, **61**, 3919–21.
- [27] Baum, C. (1999). *Detection and Identification of Visually Obscured Targets*. Taylor and Francis, Philadelphia.
- [28] Bayliss, A., Goldstein, C.I., and Turkel, E. (1985). On accuracy conditions for the numerical computation of waves. *J. Comput. Phys.*, **59**, 396–404.
- [29] Bécache, E., Joly, P., and Tsogka, C. (2001). An analysis of new mixed finite elements for the approximation of wave propagation problems. *SIAM J. Numer. Anal.*, **37**, 1053–84.
- [30] Bécache, E., Joly, P., and Tsogka, C. (2002). A new family of mixed finite elements for the linear elastodynamic problem. *SIAM J. Numer. Anal.*, **39**, 2109–32.
- [31] Beck, R., Deuflhard, P., Hiptmair, R., Hoppe, R., and Wohlmuth, B.

- (1998). Adaptive multilevel methods for edge element discretizations of Maxwell's equations. *Surveys of Mathematics in Industry*, **8**, 271–312.
- [32] Beck, R., Hiptmair, R., Hoppe, R., and Wohlmuth, B. (2000). Residual based a-posteriori error estimators for eddy current computation. *RAIRO - Math. Model. Numer. Anal.*, **34**, 159–82.
- [33] Beck, R., Hiptmair, R., and Wohlmuth, B. (2000). A hierarchical error estimator for eddy current computation. In *ENUMATH99: Proceedings of the 3rd European Conference on Numerical Mathematics and Advanced Applications* (ed. P. Neittaanmäki and T. Tiihonen), pp. 110–20. World Scientific, Singapore.
- [34] Ben Belgacem, F. and Bernardi, C. (1999). Spectral element discretization of the Maxwell equations. *Math. Comput.*, **68**, 1497–520.
- [35] Ben Belgacem, F., Bernardi, C., Costabel, M., and Dauge, M. (1997). Un résultat de densité pour les équations de Maxwell. *C. R. Acad. Sci. Paris, Série I*, **324**, 731–36.
- [36] Bérenger, J.P. (1994). A perfectly matched layer for the absorption of electromagnetics waves. *J. Comput. Phys.*, **114**, 185–200.
- [37] Bérenger, J.P. (1996). Perfectly matched layer for the FDTD solution of wave-structure interaction problems. *IEEE Trans. Antennas Propagat.*, **44**, 110–7.
- [38] Bermúdez, A. and Pedreira, D.G. (1992). Mathematical analysis of a finite element method without spurious solutions for computation of dielectric waveguides. *Numer. Math.*, **61**, 39–57.
- [39] Bermúdez, A., Pedreira, D.G., and Joly, P. (2002). A hybrid approach for the computation of guided modes in integrated optics. *Adv. Comput. Math.*, **16**, 229–61.
- [40] Bernardi, C. (1989). Optimal finite element interpolation on curved domains. *SIAM J. Numer. Anal.*, **26**, 1212–40.
- [41] Bernardi, C., Dauge, M., and Maday, Y. (2000). Compatibilité de traces aux arêtes et coins d'un polyèdre. *C. R. Acad. Sci. Paris, Série I*, **331**, 679–84.
- [42] Bespalov, A.N. (1988). Finite element method for the eigenmode problem of a RF cavity. *Sov. J. Numer. Anal. Math. Modell.*, **3**, 163–78.
- [43] Biegler, L.T., Coleman, T.F., Conn, A.R., and Santosa, F.N. (ed.) (1997). *Large-scale Optimization and Applications. Part I: Optimization in Inverse Problems and Design*. Springer, New York.
- [44] Birman, M. and Solomyak, M. (1987).  $L^2$  theory of the Maxwell operator in arbitrary domains. *Russ. Math. Surv.*, **42**, 75–96.
- [45] Bleistein, N., Cohen, J.K., and Stockwell Jr, J.W. (2001). *Mathematics of Multidimensional Seismic Imaging, Migration, and Inversion*. Springer, New York.
- [46] Boffi, D. (2000). Fortin operators and discrete compactness for edge elements. *Numer. Math.*, **87**, 229–46.

- [47] Boffi, D. (2001). A note on the de Rham complex and a discrete compactness property. *Appl. Math. Lett.*, **14**, 33–8.
- [48] Boffi, D., Brezzi, F., and Gastaldi, L. (2000). On the problem of spurious eigenvalues in the approximation of linear elliptic problems in mixed form. *Math. Comput.*, **69**, 121–40.
- [49] Boffi, D., Fernandez, P., and Perugia, I. (1999). Computational models of electromagnetic resonators: Analysis of edge element approximation. *SIAM J. Numer. Anal.*, **36**, 1264–90.
- [50] Boffi, D. and Gastaldi, L. (2002). Edge finite elements for the approximation of Maxwell resolvent operator. *RAIRO - Math. Model. Numer. Anal.*, **36**, 293–305.
- [51] Bonnet Ben-Dhia, A.S., Hazard, C., and Lohrengel, S. (1999). A singular field method for the solution of Maxwell's equations in polyhedral domains. *SIAM J. Appl. Math.*, **59**, 2028–44.
- [52] Borden, B. (1999). *Radar Imaging of Airborne Targets*. Institute of Physics, Bristol.
- [53] Bossavit, A. (1988). Mixed finite elements and the complex of Whitney forms. In *The Mathematics of Finite Elements and Applications VI* (ed. J. Whiteman), pp. 137–44. Academic Press, London.
- [54] Bossavit, A. (1998). *Computational Electromagnetism*. Academic Press, San Diego.
- [55] Bossavit, A. and Vérité, J.C. (1982). A mixed FEM-BEM method to solve 3-D eddy current problems. *IEEE Trans. Mag.*, **18**, 431–5.
- [56] Boyse, W.E., Lynch, D.R., Paulsen, K.D., and Minerbo, G.N. (1992). Node based finite element modelling of Maxwell's equations. *IEEE Trans. Antennas Propagat.*, **40**, 642–51.
- [57] Braess, D. (2001). *Finite Elements* (2nd edn). Cambridge University Press, Cambridge.
- [58] Bramble, J.H., Pasciak, J.E., and Xu, J. (1991). The analysis of multigrid algorithms with non-nested spaces or non-inherited quadratic forms. *Math. Comput.*, **56**, 1–34.
- [59] Brandts, J.H. (1994). Superconvergence and a posteriori error estimation for triangular mixed finite element methods. *Numer. Math.*, **68**, 311–24.
- [60] Brenner, S.C. and Scott, L.R. (1994). *The Mathematical Theory of Finite Element Methods*. Springer, New York.
- [61] Brezzi, F. and Fortin, M. (1991). *Mixed and Hybrid Finite Element Methods*. Springer, New York.
- [62] Buffa, A., Costabel, M., and Schwab, C. (2001). Boundary element methods for Maxwell's equations on non-smooth domains. To appear in *Numer. Math.*
- [63] Buffa, A., Costabel, M., and Sheen, D. (2003). On the traces of  $H(\text{curl}, \Omega)$  in Lipschitz domains. Preprint.
- [64] Buffa, A., Hiptmair, R., von Petersdorff, T., and Schwab, C. (2002). Boundary element methods for Maxwell equations in Lipschitz domains. to appear

- in *Numer. Math.*
- [65] Burnett, D.S. (1994). A three-dimensional acoustic infinite element based on a prolate spheroidal multipole expansion. *J. Acoust. Soc. Am.*, **96**, 2798–816.
  - [66] Burnett, D.S. and Holford, R.L. (1998). Multipole-based 3-D infinite elements: an ellipsoidal acoustic element and a spherical electromagnetic element. In *Computational Methods for Unbounded Domains* (ed. T. Geers), pp. 52–62. IUTAM.
  - [67] Cai, W. and Yu, T. (2000). Fast calculation of dyadic Green's functions for electromagnetic scattering in a multilayered medium. *J. Comput. Phys.*, **165**, 1–21.
  - [68] Cai, X.C. and Widlund, O.B. (1993). Multiplicative Schwarz algorithms for some nonsymmetric and indefinite problems. *SIAM J. Numer. Anal.*, **30**, 936–52.
  - [69] Cakoni, F. and Colton, D. (2002). Combined far field operators in electromagentic inverse scattering theory. To appear in *Math. Meth. Appl. Sci.*
  - [70] Cakoni, F., Colton, D., and Monk, P. (2003). The direct and inverse scattering problems for partially coated obstacles. Submitted for publication.
  - [71] Caorsi, S., Fernandes, P., and Raffetto, M (2000). On the convergence of Galerkin finite element approximations of electromagnetic eigenproblems. *SIAM J. Numer. Anal.*, **38**, 580–607.
  - [72] Cecot, W., Demkowicz, L., and Rachowicz, W. (2000). A three-dimensional infinite element for Maxwell's equations. TICAM Report 00–20, TICAM, University of Texas at Austin, USA.
  - [73] Cessenat, M. (1996). *Mathematical Methods in Electromagnetism*. World Scientific, Singapore.
  - [74] Cessenat, O. (1996). *Application d'une nouvelle formulation variationnelle aux équations d'ondes harmoniques. Problèmes de Helmholtz 2D et de Maxwell 3D*. Ph. D. thesis, Université Paris IX Dauphine.
  - [75] Chatterjee, A., Jin, J.M., and Volakis, J.L. (1993). Edge-based finite elements and vector ABC's applied to 3D scattering. *IEEE Trans. Antennas Propagat.*, **41**, 221–26.
  - [76] Chen, Z. and Dai, S. (2002). On the efficiency of adaptive finite element methods for elliptic problems with discontinuous coefficients. *SIAM J. Sci. Comput.* (to appear).
  - [77] Chen, Z., Du, Q., and Zou, J. (2000). Finite element methods with matching and nonmatching meshes for Maxwell equations with discontinuous coefficients. *SIAM J. Numer. Anal.*, **37**, 1542–70.
  - [78] Chew, W. (1990). *Waves and Fields in Inhomogeneous Media*. Van Nostrand Reinhold, New York.
  - [79] Chew, W. C. and Weedon, W. H. (1994). A 3D perfectly matched medium from modified Maxwell's equations with stretched coordinates. *Microwave Opt. Technol. Lett.*, **7**(13), 599–604.

- [80] Ciarlet, P.G. (1978). *The Finite Element Method for Elliptic Problems*, Volume 4 of *Studies In Mathematics and Its Applications*. North-Holland, New York.
- [81] Clément, P. (1975). Approximation by finite element functions using local regularization. *RAIRO Anal. Numér.*, **9**, 77–84.
- [82] Cohen, G.C. (2002). *Higher-order numerical methods for transient wave equations*. Springer, Berlin.
- [83] Cohen, G. and Monk, P. (1998). Gauss point mass lumping schemes for Maxwell's equations. *Numer. Meth. Partial Diff. Eqns.*, **14**, 63–88.
- [84] Cohen, G. and Monk, P. (1999). Mur-Nédélec finite element schemes for Maxwell's equations. *Comput. Meth. Appl. Mech. Eng.*, **169**, 197–217.
- [85] Collino, F. and Monk, P. (1998). Optimizing the perfectly matched layer. *Comput. Meth. Appl. Mech. Eng.*, **164**, 157–71.
- [86] Collino, F. and Monk, P. (1998). The perfectly matched layer in curvilinear coordinates. *SIAM J. Sci. Comput.*, **19**, 2061–90.
- [87] Colton, D.L. (2002). Inverse acoustic and electromagnetic scattering theory. To appear in *Proceedings of the MSRI Conference on Inverse Problems*, G. Uhlmann, ed., Cambridge University Press, Cambridge.
- [88] Colton, D., Coyle, J., and Monk, P. (2000). Recent developments in inverse acoustic scattering theory. *SIAM Rev.*, **42**, 369–414.
- [89] Colton, D., Giebermann, K., and Monk, P. (2000). A regularized sampling method for solving three dimensional inverse scattering problems. *SIAM J. Sci. Comput.*, **21**, 2316–30.
- [90] Colton, D., Haddar, H., and Monk, P. (2002). The linear sampling method for solving the electromagnetic inverse scattering problem. *SIAM Journal on Computing*, **14**, 719–31.
- [91] Colton, D. and Kirsch, A. (1984). Dense sets and far field patterns in acoustic wave propagation. *SIAM J. Math. Anal.*, **15**, 996–1006.
- [92] Colton, D. and Kirsch, A. (1996). A simple method for solving inverse scattering problems in the resonance region. *Inu Prob.*, **12**, 383–93.
- [93] Colton, D. and Kress, R. (1983). *Integral Equation Methods in Scattering Theory*. Wiley, New York.
- [94] Colton, D. and Kress, R. (1998). *Inverse Acoustic and Electromagnetic Scattering Theory* (2nd edn). Number 93 in Applied Mathematical Sciences. Springer, New York.
- [95] Colton, D. and Kress, R. (2001). On the denseness of Herglotz wave functions and electromagnetic Herglotz pairs in Sobolev spaces. *Math. Meth. Appl. Sci.*, **24**, 1289–303.
- [96] Colton, D., Kress, R., and Monk, P. (1997). Inverse scattering from an orthotropic medium. *J. Comput. Appl. Math.*, **81**, 269–98.
- [97] Colton, D. and Monk, P. (1985). A novel method for solving the inverse scattering problem for time-harmonic acoustic waves in the resonance region. *SIAM J. Appl. Math.*, **45**, 1039–53.
- [98] Colton, D. and Monk, P. (1986). A novel method for solving the inverse

- scattering problem for time-harmonic acoustic waves in the resonance region II. *SIAM J. Appl. Math.*, **46**, 506–23.
- [99] Colton, D. and Monk, P. (1994). The detection and monitoring of leukemia using electromagnetic waves: mathematical theory. *Innu Prob.*, **10**, 1235–51.
- [100] Colton, D. and Monk, P. (1995). The detection and monitoring of leukemia using electromagnetic waves: Numerical analysis. *Innu Prob.*, **11**, 329–42.
- [101] Colton, D. and Potthast, R. (1999). The inverse electromagnetic scattering problem for an anisotropic medium. *Quart. J. Mech. Appl. Math.*, **52**, 349–72.
- [102] Costabel, M. (1990). A remark on the regularity of solutions of Maxwell's equations on Lipschitz domains. *Math. Meth. Appl. Sci.*, **12**, 365–8.
- [103] Costabel, M. and Dauge, M. (1998). Espaces fonctionnels Maxwell: Les gentils, les méchants et les singularités. Report available at <http://www.maths.univ-rennes1.fr/~dauge/>.
- [104] Costabel, M. and Dauge, M. (1998). Un résultat de densité pour les équations de Maxwell régularisées dans un domaine Lipschitzien. Technical report, IRMAR, Université de Rennes 1, France. Report available at <http://www.maths.univ-rennes1.fr/~costabel/>.
- [105] Costabel, M. and Dauge, M. (1999). Maxwell and Lamé eigenvalues on polyhedra. *Math. Meth. Appl. Sci.*, **22**, 243–58.
- [106] Costabel, M. and Dauge, M. (2000). Singularities of electromagnetic fields in polyhedral domains. *Arch. Rat. Mech. Anal.*, **151**, 221–76.
- [107] Costabel, M., Dauge, M., and Nicaise, S. (1999). Singularities of Maxwell interface problems. *RAIRO - Math. Model. Numer. Anal.*, **33**, 627–49.
- [108] Cox, S.J. and Dobson, D.C. (1999). Maximizing band gaps in twodimensional photonic crystals. *SIAM J. Appl. Math.*, **59**, 2108–20.
- [109] Crouzeix, M. and Raviart, P.-A. (1973). Conforming and non-conforming finite element methods for solving the stationary Stokes problem. *RAIRO Anal. Numér.*, **7**, R-3, 33–76.
- [110] Crowley, C.W., Silvester, P.P., and Hurwitz, H. (1988). Covariant projection elements for 3D vector field problems. *IEEE Trans. Mag.*, **24**, 397–400.
- [111] Cutzach, P.-M. and Hazard, C. (1998). Existence, uniqueness and analyticity properties for electromagnetic scattering in a two-layered medium. *Math. Meth. Appl. Sci.*, **21**, 433–61.
- [112] Dauge, M. (1988). *Elliptic Boundary Value Problems on Corner Domains*, Volume 1341 of *Lecture Notes in Mathematics*. Springer, Berlin.
- [113] Dauge, M., Costabel, M., and Martin, D. (1999). Numerical investigation of a boundary penalization method for Maxwell equations. Report available at <http://www.maths.univ-rennes1.fr/~dauge/>.
- [114] Dauge, M., Costabel, M., and Martin, D. (2001). Weighted regularization of Maxwell equations in polyhedral domains. Report available at <http://www.maths.univ-rennes1.fr/~dauge/>.

- [115] Dautray, R. and Lions, J.-L. (1990). *Spectral Theory and Applications*, Volume 3 of *Mathematical Analysis and Numerical Methods for Science and Technology*. Springer, Berlin.
- [116] de La Bourdonnaye, A. (1995). Some formulations coupling finite element and integral equation methods for Helmholtz equation and electromagnetism. *Numer. Math.*, **69**, 257–68.
- [117] Deming, R.W. and Devaney, A.J. (1997). Diffraction tomography for multimonostatic ground penetrating radar imaging. *Inu. Prob.*, **13**, 29–45.
- [118] Demkowicz, L., Gopalakrishnan, J., and Pasciak, J.E. (2002). Analysis of a multigrid algorithm for time harmonic Maxwell equations. In press.
- [119] Demkowicz, L. and Monk, P. (2001). Discrete compactness and the approximation of Maxwell's equations in  $\mathbb{R}^3$ . *Math. Comput.*, **70**, 507–23.
- [120] Demkowicz, L., Monk, P., and Vardapetyan, L. (2000). de Rham diagram for  $hp$  finite element spaces. *Comput. Math. Appl.*, **39**, 29–38.
- [121] Demkowicz, L. and Pal, M. (1998). An infinite element for Maxwell's equations. *Comput. Meth. Appl. Mech. Eng.*, **164**, 77–94.
- [122] Demkowicz, L. and Vardapetyan, L. (1998). Modelling electromagnetic absorbtion/scattering problems using  $hp$ -adaptive finite elements. *Comput. Methods Appl. Mech. Engrg.*, **152**, 103–24.
- [123] Descloux, J., Nassif, N., and Rappaz, J. (1978). On spectral approximation. I. The problem of convergence. *RAIRO Anal. Númer.*, **12**, 97–112.
- [124] Devaney, A.J. (1984). Acoustic tomography. In *Inverse Problems in Acoustic and Elastic Waves* (ed. F. Santosa, et al.), pp. 250–73. SIAM, Philadelphia.
- [125] Devaney, A.J. (2003). Super-resolution processing of multi-static data using time-reversal and MUSIC. to appear in J. Opt. Soc. Am.
- [126] Dierkes, T., Dorn, O., Natterer, F., Palamodov, V., and Sielschot, H. (2002). Fréchet derivatives for some bilinear inverse problems. *SIAM J. Appl. Math.*, **62**, 2092–113.
- [127] Dobson, D.C., Gopalakrishnan, J., and Pasciak, J.E. (2000). An efficient method for band structure calculations in 3D photonic crystals. *J. Comput. Phys.*, **161**, 668–79.
- [128] Dobson, D.C. and Pasciak, J.E. (2001). Analysis of an algorithm for computing electromagnetic Bloch modes using Nedelec spaces. *Comput. Meth. Appl. Math.*, **1**, 138–53.
- [129] Dorn, O., Bertete-Aguirre, H., Berryman, J.G., and Papanicolaou, G.C. (1999). A nonlinear inversion method for 3D-electromagnetic imaging using adjoint fields. *Inu. Prob.*, **15**, 1523–58.
- [130] Dorn, O., Miller, E., and Rappaport, C. (2000). A shape reconstruction method for electromagnetic tomography using adjoint fields and level sets. *Inu. Prob.*, **16**, 1119–56.
- [131] Douglas Jr., J., Santos, J.E., and Sheen, D. (2000). A nonconforming mixed finite element method for maxwell's equations. *Math. Meth. Appl. Sci.*, **10**, 593–613.

- [132] Dubois, F. (1990). Discrete vector potential representation of a divergence free vector field in three dimensional domains: Numerical analysis of a model problem. *SIAM J. Numer. Anal.*, **27**, 1103–42.
- [133] Dubois, F. (2000). Du tourbillon au champ de vitesse. Lecture notes.
- [134] Edelsbrunner, H. (2001). *Geometry and Topology for Mesh Generation*. Cambridge University Press, Cambridge.
- [135] Elmkies, A. and Joly, P. (1996). Éléments finis et condensation de masse pour les équations de Maxwell: le cas 2D. Technical Report 3035, INRIA, France.
- [136] Elmkies, A. and Joly, P. (1997). Éléments finis d'arête et condensation de masse pour les équations de Maxwell: le cas 3D. *C. R. Acad. Sd. Paris, Série 1*, **325**, 1217–22.
- [137] Engl, H.W., Hanke, M., and A.Neubauer (1996). *Regularization of Inverse Problems*. Kluwer, Dordrecht.
- [138] Eriksson, K. and Johnson, C. (1987). Error estimates and automatic time step control for nonlinear parabolic problems, I. *SIAM J. Numer. Anal.*, **24**, 12–23.
- [139] Falk, R.S. and Osborn, J.E. (1980). Error estimates for mixed methods. *RAIRO Anal. Numér.*, **14**, 269–77.
- [140] George, P.-L. and Borouchaki, H. (1998). *Delaunay Triangulation and meshing*. Editions Hermès, Paris.
- [141] Giles, M.B. and Süli, E. (2002). Adjoint methods for PDEs: *a posteriori* error analysis and postprocessing. *Acta Numerica*, **11**, 145–236.
- [142] Girault, V. (1988). Incompressible finite element methods for Navier–Stokes equations with nonstandard boundary conditions in  $\mathbb{R}^3$ . *Math. Comput.*, **51**, 53–8.
- [143] Girault, V. and Raviart, P.A. (1986). *Finite Element Methods for Navier–Stokes Equations*. Springer, New York.
- [144] Givoli, D. (1991). Non reflecting boundary conditions. *J. Comput. Phys.*, **94**, 1–29.
- [145] Goldstein, C.I. (1981). The finite element method with non-uniform mesh sizes applied to the exterior Helmholtz problem. *Numer. Math.*, **38**, 61–82.
- [146] Golub, G.H. and Van Loan, C.F. (1996). *Matrix computations* (3rd edn). Johns Hopkins University, Baltimore.
- [147] Gopalakrishnan, J. and Pasciak, J.E. (2003). Overlapping Schwarz preconditioners for indefinite time harmonic Maxwell equations. *Math. Comput.*, **72**, 1–15.
- [148] Grarinaru, V. and Hiptmair, R. (1999). Whitney elements on pyramids. *ETNA*, **8**, 154–68.
- [149] Graglia, R.D., Wilton, D.R., and Peterson, A.F. (1997). Higher order interpolatory vector bases for computational electromagnetics. *IEEE Trans. Antennas Propagat.*, **45**, 329–342.
- [150] Graglia, R.D., Wilton, D.R., Peterson, A.F., and Gheorma, I.L. (1998). Higher order interpolatory vector bases on prism elements. *IEEE Trans.*

- Antennas Propagat.*, **46**, 442–50..
- [151] Grisvard, P. (1985). *Elliptic Problems in Nonsmooth Domains*. Pitman, London.
- [152] Grote, M. J. (2000). Non-reflecting boundary conditions for electromagnetic scattering. *Int. J. Numer. Modell.*, **13**, 397–416.
- [153] Grote, M.J. and Keller, J.B. (1995). On nonreflecting boundary conditions. *J. Comput. Phys.*, **122**, 231–43.
- [154] Grote, M.J. and Keller, J.B. (1998). Nonreflecting boundary conditions for Maxwell's equations. *J. Comput. Phys.*, **139**, 327–42.
- [155] Haddar, H. and Monk, P. (2002). The linear sampling method for solving the electromagnetic inverse medium problem. *Inv. Prob.*, **18**, 891–906.
- [156] Hanouzet, B. and Sesquès, M. (1990). Influnce des termes de courbure dans les conditions aux limites artificielles pour les équations de Maxwell. *C.R. Acad. Sci. Paris*, **311 Série I**, 561–4.
- [157] Hansen, P.C. (1998). *Rank-Deficient and Discrete Ill-posed Problems*. SIAM, Philadelphia.
- [158] Hartman, P. and Wilcox, C. (1961). On solutions of the Helmholtz equation in exterior domains. *Math. Zeit.*, **75**, 228–55.
- [159] Hazard, C. and Lenoir, M. (1996). On the solution of time-harmonic scattering problems for Maxwell's equations. *SIAM J. Math. Anal.*, **27**, 1597–630.
- [160] Hesthaven, J.S. and Warburton, T. (2002). Nodal high-order methods on unstructured grids. I. Time-domain solution of Maxwell's equations. *J. Comput. Phys.*, **181**, 186–221.
- [161] Hiptmair, R. (1998). Multigrid method for Maxwell's equations. *SIAM J. Numer. Anal.*, **36**, 204–25.
- [162] Hiptmair, R. (2001). Higher order Whitney forms. *Progress in Electromagnetics Research*, **32**, 271–99.
- [163] Hiptmair, R. (2002). Coupling of finite elements and boundary elements in electromagnetic scattering. In press.
- [164] Hiptmair, R. (2002). Finite elements in computational electromagnetism. *Acta Numerica*, **11**, 237–339.
- [165] Hiptmair, R. (2002). Symmetric coupling for eddy current problems. *SIAM J. Numer. Anal.*, **40**, 41–65.
- [166] Hohage, T. (2001). On the numerical solution of a three-dimensional inverse medium scattering problem. *Inv. Prob.*, **17**, 1743–63.
- [167] Hsiao, G., Monk, P., and Nigam, N. (2002). Error analysis of a finite element-integral equation scheme for approximating the time-harmonic Maxwell system. *SIAM J. Numer. Anal.*, **40**, 198–219.
- [168] Ihlenburg, F. (1998). *Finite Element Analysis of Acoustic Scattering*, Volume 132 of *Applied Mathematical Sciences*. Springer, Berlin.
- [169] Ihlenburg, F. and Babuška, I. (1995). Finite element solution of the Helmholtz equation with high wavenumber Part I: The h-version of the FEM. *Computers Math. Applic.*, **30**, 9–37.

- [170] Ihlenburg, F. and Babuška, I. (1997). Finite element solution of the Helmholtz equation with high wave number Part II: the  $hp$  version of the FEM. *SIAM J. Numer. Anal.*, **34**, 315–58.
- [171] Ikehata, M. (1998). Reconstruction of an obstacle from the scattering amplitude at a fixed frequency. *Inverse Prob.*, **14**, 949–54.
- [172] Ishimaru, A. (1991). *Electromagnetic wave propagation, radiation, and scattering*. Prentice-Hall, Englewood Cliffs.
- [173] Jami, A. and Lenoir, M. (1978). A variational formulation for exterior problems in linear hydrodynamics. *Comput. Meth. Appl. Mech. Eng.*, **16**, 341–59.
- [174] Jeffreys, H. and Jeffreys, B.S. (1972). *Methods of Mathematical Physics* (3rd edn). Cambridge University Press, Cambridge.
- [175] Jerison, D.S. and Kenig, C.E. (1981). The Neumann problem on Lipschitz domains. *Bull. Amer. Math. Soc.*, **4**, 203–7.
- [176] Jerison, D.S. and Kenig, C.E. (1982). Boundary value problems on lipschitz domains. In *Studies in Partial Differential Equations* (ed. W. Littmann), pp. 1–68. Math. Assoc. America, Washington DC. MAA Studies in Mathematics No. 23.
- [177] Jin, Jian-Ming (1993). *The Finite Element Method in Electromagnetics*. Wiley, New York.
- [178] Joly, P., Poirier, C., Roberts, J.E., and Trouve, P. (1996). A new nonconforming finite element method for the computation of electromagnetic guided waves I: mathematical analysis. *SIAM J. Numer. Anal.*, **33**, 1494–525.
- [179] Joly, P., Poirier, C., Roberts, J.-E., and Trouve, P. (1991). A new nonconforming finite element method for computation of electromagnetic guided waves. In *Computing methods in applied science and engineering* (ed. R. Glowinski), pp. 433–44. Nova Science Publishers.
- [180] Kanellopoulos, V.N. and Webb, J.P. (1991). A numerical study of vector absorbing boundary conditions for the finite-element solution of Maxwell's equations. *IEEE Microwave and Guided Wave Lett.*, **1**, 325–7.
- [181] Keller, J.B. (1962). Geometrical theory of diffraction. *J. Opt. Soc. Am.*, **52**, 116–30.
- [182] Keller, J.B. and Givoli, D. (1989). Exact non-reflecting boundary conditions. *J. Comput. Phys.*, **82**, 172–92.
- [183] Kikuchi, F. (1986). An isomorphic property of two Hilbert spaces appearing in electromagnetism: analysis by the mixed formulation. *Japan. J. Appl. Math.*, **3**, 53–8.
- [184] Kikuchi, F. (1987). Mixed and penalty formulations for finite element analysis of an eigenvalue problem in electromagnetism. *Comput. Meth. Appl. Mech. Eng.*, **64**, 509–21.
- [185] Kikuchi, F. (1989). On a discrete compactness property for the Nédélec finite elements. *J. Fac. Sci. Univ. Tokyo, Sect. 1A Math.*, **36**, 479–90.
- [186] Kirsch, A. (1996). *An Introduction to the Mathematical Theory of Inverse*

- Problems*. Springer, New York.
- [187] Kirsch, A. (1998). Characterization of the shape of a scattering obstacle using the spectral data of the far field operator. *Inv. Prob.*, **14**, 1489–512.
- [188] Kirsch, A. and Monk, P. (1990). Convergence analysis of a coupled finite element and spectral method in acoustic scattering. *IMA J. Numer. Anal.*, **9**, 425–47.
- [189] Kirsch, A. and Monk, P. (1994). An analysis of the coupling of finite element and Nyström methods in acoustic scattering. *IMA J. Numer. Anal.*, **14**, 523–44.
- [190] Kirsch, A. and Monk, P. (1995). A finite element/spectral method for approximating the time harmonic Maxwell system in  $\mathbb{R}^3$ . *SIAM J. Appl. Math.*, **55**, 1324–44.
- [191] Kirsch, A. and Monk, P. (1998). Corrigendum to “A finite element/spectral method for approximating the time-harmonic Maxwell system in  $\mathbb{R}^3$ ” (SIAM J. Appl. Math., **55** (1995) pp. 1324–44). *SIAM J. Appl. Math.*, **58**, 2024–8.
- [192] Kirsch, A. and Monk, P. (2002). A finite element method for approximating electromagnetic scattering from a conducting object. *Numer. Math.*, **92**, 501–34.
- [193] Kress, R. (1999). *Linear Integral Equations* (2nd edn). Springer, Berlin.
- [194] Kress, R. (2002). Specific theoretical tools. In *Scattering* (ed. R. Pike and P. Sabatier), pp. 37–210. Academic Press, San Diego.
- [195] Křížek, M. and Neittaanmäki, P. (1984). Finite element approximation for div-rot system with mixed boundary conditions in non-smooth plane domains. *Apl. Mat.*, **29**, 272–85.
- [196] Křížek, M. and Neittaanmäki, P. (1984). On the validity of Friedrichs' inequalities. *Math. Scand.*, **54**, 17–26.
- [197] Křížek, M. and Neittaanmäki, P. (1985). Solvability of a first order system in three-dimensional non-smooth domains. *Apl. Mat.*, **30**, 307–15.
- [198] Křížek, M. and Neittaanmäki, P. (1989). On time-harmonic Maxwell equations with nonhomogeneous conductivities: solvability and FE-approximation. *Apl. Mat.*, **34**, 480–99.
- [199] Langenberg, K.J. (1987). Applied inverse problems for acoustic, electromagnetic and elastic wave scattering. In *Basic Methods of Tomography and Inverse Problems* (ed. P. Sabatier), pp. 124–467. Adam Hilger, Bristol.
- [200] Lassas, M. and Sommersalo, E. (2001). Analysis of the PML equations in general convex geometry. *Proc. Roy. Soc. Edinburgh Sect A*, **131**, 1183–207.
- [201] Lassas, M. and Sommersalo, E. (1998). On the existence and convergence of the solution of pml equations. *Computing*, **60**, 229–42.
- [202] Lax, P. (2002). *Functional Analysis*. Wiley, New York.
- [203] Lebedev, N.N. (1972). *Special Functions and Their Applications*. Dover, New York.
- [204] Lee, B., Manteuffel, T.A., McCormick, S.F., and Ruge, J. (2000). Firstorder system least-squares for the Helmholtz equation. *SIAM J. Sci. Comput.*,

- 21**, 1927–49.
- [205] Lee, J.F. and Mittra, R. (1992). A note on the application of edge-elements for modeling 3-dimensional inhomogeneously-filled cavities. *IEEE Trans. Microwave Theory Tech.*, **40**, 1767–73.
  - [206] Lee, J.F., Sun, D.K., and Cendes, Z.J. (1991). Tangential vector finite elements for electromagnetic field computing. *IEEE Trans. Mag.*, **27**, 4032–5.
  - [207] Leis, R. (1988). *Initial Boundary Value Problems in Mathematical Physics*. Wiley, New York.
  - [208] Lenoir, M. (1986). Optimal isoparametric finite elements and error estimates for domains involving curved boundaries. *SIAM J. Numer. Anal.*, **23**, 562–80.
  - [209] Levillain, V. (1990). Coupling integral equation methods and finite volume elements for the resolution of time harmonic Maxwell's equations in three dimensional heterogeneous medium. Technical Report 222, Centre de Mathématiques Appliquées, Ecole Poly technique.
  - [210] Levillain, V. (1992). Eigenvalue approximation by mixed methods for resonant inhomogenous cavities with metallic boundaries. *Math. Comput.*, **58**, 11–20.
  - [211] Lin, Q. and Yan, N. (2000). Global superconvergence for Maxwell's equations. *Math. Comput.*, **229**, 159–76.
  - [212] Liu, J. and Jin, J.M. (2001). A novel hybridization of higher order finite element and boundary integral methods for electromagnetic scattering and radiation problems. *IEEE Trans. Antennas Propagat.*, **49**, 1794–806.
  - [213] Masmoudi, M. (1987). Numerical solution for exterior problems. *Numer. Math.*, **51**, 87–101.
  - [214] McCartin, B.J. and Dicello, J.F. (1989). Three dimensional finite difference frequency domain scattering computation using the control region approximation. *IEEE Trans. Mag.*, **25**, 3092–94.
  - [215] McLean, W. (2000). *Strongly Elliptic Systems and Boundary Integral Equations*. Cambridge University Press, Cambridge.
  - [216] Monk, P. (1992). Analysis of a finite element method for Maxwell's equations. *SIAM J. Numer. Anal.*, **29**, 714–29.
  - [217] Monk, P. (1992). A finite element method for approximating the timeharmonic Maxwell equations. *Numer. Math.*, **63**, 243–61.
  - [218] Monk, P. (1994). On the  $p$  and  $hp$  extension of Nédélec's curl conforming elements. *J. Comput. Appl. Math.*, **53**, 117–37.
  - [219] Monk, P. (1994). Superconvergence of finite element approximations to Maxwell's equations. *Numer. Meth. Partial Diff. Eqns.*, **10**, 793–812.
  - [220] Monk, P. (1995). The near field to far field transformation. *COMPEL*, **14**, 41–56.
  - [221] Monk, P. (1998). A posteriori error indicators for Maxwell's equations. *J. Comput. Appl. Math.*, **100**, 173–90.
  - [222] Monk, P. (2003). A simple proof of convergence for an edge element discretization

- of Maxwell's equations. To appear in *Computational Electromagnetics*, Proc. GAMM Workshop, Kiel (Germany), January 26–28, 2001 (C. Carstensen, S. Funken, W. Hackbusch, R.H.W. Hoppe, P. Monk; eds.), Lecture Notes in Computational Science and Engineering, Vol. 28, Springer, Berlin-Heidelberg-New York, 2003.
- [223] Monk, P. and Parrott, A.K. (1994). A dispersion analysis of finite element methods for Maxwell's equations. *SIAM J. Sci. Comput.*, **15**, 916–37.
- [224] Monk, P. and Parrott, A.K. (2001). Phase-accuracy comparisons and improved far-field estimates for 3-D edge elements on tetrahedral meshes. *J. Comput. Phys.*, **170**, 614–41.
- [225] Monk, P. and Süli, E. (1994). A convergence analysis of Yee's scheme on non-uniform grids. *SIAM J. Numer. Anal.*, **31**, 393–412.
- [226] Monk, P. and Süli, E. (1998). The adaptive computation of far field patterns by a posteriori error estimation of linear functionals. *SIAM J. Numer. Anal.*, **36**, 251–74.
- [227] Muñoz-Sola, R. (1997). Polynomial liftings on a tetrahedron and applications to the  $hp$  version of the finite element method in three dimensions. *SIAM J. Numer. Anal.*, **34**, 282–314.
- [228] Müller, C. (1969). *Foundations of the Mathematical Theory of Electromagnetic Waves*. Springer, Berlin.
- [229] Mur, G. (1981). Absorbing boundary conditions for the finite-difference approximation of the time-domain electromagnetic-field equations. *IEEE Trans. Electromag. Compatibility*, **23**, 377–82.
- [230] Mur, G. (1992). The finite-element modeling of three-dimensional timedomain electromagnetic fields in strongly inhomogeneous media. *IEEE Trans. Mag.*, **28**, 1130–3.
- [231] Mur, G. and de Hoop, A.T. (1985). A finite-element method for computing three-dimensional electromagnetic fields in inhomogeneous media. *IEEE Trans. Mag.*, **21**, 2188–91.
- [232] Natterer, F. and Wübbeling, F. (2001). *Mathematical Methods in Image Reconstruction*. SIAM, Philadelphia.
- [233] Nédélec, J.C. (1980). Mixed finite elements in  $\mathbb{R}^3$ . *Numer. Math.*, **35**, 315–41.
- [234] Nédélec, J.C. (1982). Eléments finis mixtes incompressibles pour l'équation de Stokes dans  $\mathbb{R}^3$ . *Numer. Math.*, **39**, 97–112.
- [235] Nédélec, J.C. (1986). A new family of mixed finite elements in  $\mathbb{R}^3$ . *Numer. Math.*, **50**, 57–81.
- [236] Nédélec, J.C. (2001). *Acoustic and Electromagnetic Equations*. Number 144 in Applied Mathematical Sciences. Springer, New York.
- [237] Nečas, J. (1967). *Les Méthodes Directes en Théorie Équations Elliptiques*. Masson, France.
- [238] Nicaise, S. (2001). Edge elements on anisotropic meshes and approximation of the Maxwell equations. *SIAM J. Numer. Anal.*, **39**, 784–816.
- [239] Nicolaides, R.A. (1972). On a class of finite elements generated by Lagrange

- interpolation. *SIAM J. Numer. Anal.*, **9**, 435–45.
- [240] Nicolaides, R.A. (1992). Direct discretization of planar div–curl problems. *SIAM J. Numer. Anal.*, **29**, 32–56.
- [241] Nicolaides, R.A. and D.Q.Wang (1998). Convergence analysis of a covolume scheme for Maxwell's equations in three dimensions. *Math. Comput.*, **67**, 947–63.
- [242] Norris, A. (1998). A direct inverse scattering method for imaging obstacles with unknown surface conditions. *IMA J. Appl. Math.*, **61**, 267–90.
- [243] Odeh, F. (1963). Uniqueness theorems for the Helmholtz equation in domains with finite boundaries. *J. Math. Mech.*, **12**, 857–67.
- [244] Oden, J.T. and Carey, G.F. (1983). *Finite Elements: Mathematical Aspects*, Volume IV. Prentice-Hall, Englewood-Cliffs.
- [245] Ōkaji, T. (2002). Strong unique continuation property for the time harmonic Maxwell equations. *J. Math. Soc. Japan*, **54**, 89–122.
- [246] Osborn, J.E. (1975). Spectral approximation for compact operators. *Math. Comput.*, **29**, 712–25.
- [247] Ozdemir, T. and Volakis, J.L. (1997). Triangular prisms for edge-based vector finite element analysis of conformal antennas. *IEEE Trans. Antennas Propagat.*, **45**, 788–97.
- [248] Paulus, M., Gay-Balmaz, P., and Martin, O.J.F. (2000). Accurate and efficient computation of the Green's tensor for stratified media. *Phys. Rev E*, **62**, 5797–807.
- [249] Peterson, A.F. and Baca, R.J. (1991). Error in the finite element discretization of the scalar Helmholtz equation over electrically large regions. *IEEE Microwave and Guided Wave Letters*, **1**, 219–21.
- [250] Petropoulis, P. (2002). An analytical study of the discrete perfectly matched layer for the time-domain Maxwell equations in cylindrical coordinate. To appear.
- [251] Petropoulos, P.G. (1998). On the termination of the perfectly matched layer with local absorbing boundary conditions. *J. Comput. Phys.*, **143**, 665–73.
- [252] Petropoulos, P.G. (2000). Reflectionless sponge layers as absorbing boundary conditions for the numerical solution of Maxwell's equations in rectangular, cylindrical and spherical coordinates. *SIAM J. Appl. Math.*, **60**, 1037–58.
- [253] Potthast, R. (2000). Stability estimates and reconstructions in inverse scattering using a singular source. *J. Comput. Appl. Math.*, **114**, 247–74.
- [254] Potthast, R. (2001). *Point Sources and Multipoles in Inverse Scattering Theory*. Chapman and Hall/CRC, London.
- [255] Rachowicz, W. and Demkowicz, L. (2000). An  $hp$ -adaptive finite element method for electromagnetics part I: aata structure and constrained approximation. *Comput. Meth. Appl. Mech. Eng.*, **187**, 307–35.
- [256] Rachowicz, W. and Demkowicz, L. (2000). A three-dimensional  $hp$ -adaptive finite element package for electromagnetics. Technical Report

- 00–04, TICAM, University of Texas.
- [257] Rachowicz, W. and Demkowicz, L. (2002). An  $hp$ -adaptive finite element method for electromagnetics - part II: a 3D implementation. *Int. J. Numer. Meth. Eng.*, **53**, 147–80.
- [258] Rao, S.M., Wilton, D.R., and Glisson, A.W. (1982). Electromagnetic scattering by surfaces of arbitrary shape. *IEEE Trans. Antennas Propagat.*, **30**, 409–18.
- [259] Rappaport, C.M. (1995). Reducing the computational domain for FDTD scattering simulation using the sawtooth anechoic chamber ABC. *IEEE Trans. Mag.*, **31**, 1546–9.
- [260] Raviart, P.A. and Thomas, J.M. (1977). A mixed finite element method for 2nd order elliptic problems. In *Mathematical Aspects of the Finite Element Method* (ed. A. Dold and B. Eckmann), Lecture Notes 606. Springer, London.
- [261] Raviart, P.A. and Thomas, J.M. (1977). Primal hybrid finite element methods for 2nd order elliptic equations. *Math. Comput.*, **31**, 391–413.
- [262] Reitzinger, S. and Schoberl, J. (2002). An algebraic multigrid method for finite element discretizations with edge elements. *Numer. Linear Algebra with Appl.*, **9**, 223–8.
- [263] Rokhlin, V. (1990). Rapid solution of integral equations of scattering theory in two dimensions. *J. Comput. Phys.*, **86**, 414–39.
- [264] Sacks, Z.S., Kingsland, D.M., Lee, R., and Lee, J.F. (1995). A perfectly matched anisotropic absorber for use as an absorbing boundary condition. *IEEE Trans. Antennas Propagat.*, **43**, 1460–3.
- [265] Santos, J.E. and Sheen, D. (2000). On the existence and uniqueness of Maxwell's equations in bounded domains with application to magnetotellurics. *Math. Meth. Appl. Sci.*, **10**, 615–28.
- [266] Saranen, J. (1982). On an inequality of Friedrichs'. *Math. Scand.*, **51**, 310–22.
- [267] Schatz, A.H. (1974). An observation concerning Ritz-Galerkin methods with indefinite bilinear forms. *Math. Comput.*, **28**, 959–62.
- [268] Schatz, A.H. and Wang, J. (1996). Some new error estimates for Ritz-Galerkin methods with minimal regularity assumptions. *Math. Comput.*, **65**, 19.
- [269] Schöberl, J. (1997). NETGEN – An advancing front 2D/3D-mesh generator based on abstract rules. *Comput. Visual. Sci.*, **1**, 41–52.
- [270] Schwab, C. (2000). *p- and hp-finite Element Methods. Theory and Applications in Solid and Fluid Mechanics*. Oxford University Press, Oxford.
- [271] Schwartz, M. (1972). *Principles of Electrodynamics*. Dover, New York.
- [272] Silvester, R.P. and Ferrari, R.L. (1996). *Finite element methods for electrical engineers* (3rd edn). Cambridge University Press, Cambridge.
- [273] Sommerfeld, A. (1949). *Partial Differential Equations in Physics*. Academic Press, New York.
- [274] Stratton, J.A. (1941). *Electromagnetic Theory*. McGraw-Hill, New York.

- [275] Szabo, B. and Babuska, I. (1991). *Finite Element Analysis*. Wiley, New York.
- [276] Teixeira, F.L. and Chew, W.C. (2000). Complex space approach to perfectly matched layers: a review and some new developments. *Int. J. Numer. Modelling-Electronic Network Devices and Fields*, **13**, 441–55.
- [277] Teixeira, F.L., Hwang, K.P., WC, W.C. Chew, and Jin, J.M. (2001). Conformal PML-FDTD schemes for electromagnetic field simulations: A dynamic stability study. *IEEE Trans. Antennas Propagat.*, **49**, 902–7.
- [278] Thompson, J.F., Warsi, Z.U.A., and Martin, C.W. (1985). *Numerical Grid Generation*. North-Holland, New York.
- [279] Tikhonov, A.N., Goncharsky, A.V., Stepanov, V.V., and Yagola, A.G. (1995). *Numerical Methods for the Solution of Illposed Problems*. Kluwer, Dordrecht.
- [280] Toselli, A. (2000). Overlapping Schwarz methods for Maxwell's equations in three dimensions. *Numer. Math.*, **86**, 733–52.
- [281] Toselli, A. (2001). An iterative substructuring method for Maxwell's equations in two dimensions. *Math. Comput.*, **70**, 935–49.
- [282] Trottenberg, U. (2001). *Multigrid*. Academic Press, San Diego.
- [283] Van Bladel, J. (1961). Some remarks on Green's dyadic for infinite space. *IRE Transactions on Antennas and Propagation*, **9**, 563–6.
- [284] Van Bladel, J. (1985). *Electromagnetic Fields*. Hemisphere, New York.
- [285] Vardapetyan, L. (1999). *hp Adaptive finite element method for electromagnetics with applications to waveguiding structures*. Ph. D. thesis, Graduate School of the University of Texas at Austin.
- [286] Vardapetyan, L. and Demkowicz, L. (1999). *hp*-Adaptive finite elements in electromagnetics. *Comput. Meth. Appl. Mech. Eng.*, **169**, 331–44.
- [287] Vogelsang, V. (1991). On the strong unique continuation principle for inequalities of Maxwell type. *Math. Ann.*, **289**, 285–95.
- [288] Wait, J.R. (1962). *Electromagnetic Waves in Stratified Media*. Macmillian, New York.
- [289] Wang, Y., Monk, P., and Szabo, B.A. (1996). Computing cavity mods using the  $p$ -version of the finite element method. *IEEE Trans. Mag.*, **32**, 1934–40.
- [290] Webb, J.P. and Forghani, B. (1993). Hierarchical scalar and vector tetrahedra. *IEEE Trans. Mag.*, **29**, 1495–8.
- [291] Webb, J.P. and Kanellopoulos, V.N. (1989). Absorbing boundary conditions for the finite element solution of the vector wave equation. *IEEE Microwave and Opt. Technol. Lett.*, **2**, 370–2.
- [292] Weber, C. (1980). A local compactness theorem for Maxwell's equations. *Math. Meth. Appl. Sci.*, **2**, 12–25.
- [293] Weber, C. (1981). Regularity theorems for Maxwell's equations. *Math. Meth. Appl. Sci.*, **3**, 523–6.
- [294] Wheeler, J.A. (1978). Permafrost thermal design for the trans-Alaska pipeline. In *Moving Boundary Problems* (ed. D. Wilson, A. Solomon, and

- P. Boggs), pp. 267–84. Academic Press, New York.
- [295] Whitney, H. (1957). *Geometric Integration Theory*. Princeton University Press, Princeton.
- [296] Wilcox, C. (1956). An expansion theorem for electromagnetic fields. *Comm. Pure Appl. Math.*, **9**, 115–34.
- [297] Wittgenstein, Ludwig (1961). *Tractatus Logico-Philosophicus*. Humanities Press, New York.
- [298] Wloka, J. (1987). *Partial Differential Equations*. Cambridge University Press, Cambridge.
- [299] Wu, J.Y. and Lee, R. (1997). The advantages of triangular and tetrahedral edge elements for electromagnetic modeling with the finite-element method. *IEEE Trans. Antennas Propagat.*, **45**, 1431–7.
- [300] Yang, B. and Hesthaven, J. S. (2000). Multidomain pseudospectral computation of Maxwell's equations in 3-D general curvilinear coordinates. *Appl. Numer. Math.*, **33**, 281–89.
- [301] Yee, K.S. (1966). Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media. *IEEE Trans. Antennas Propagat.*, **16**, 302–7.
- [302] Zhao, L. and Cangellaris, A.C. (1996). A general approach for developing unsplit-field time-domain implementations of perfectly matched layers for FDTD grid truncation. *IEEE Trans. Microwave Theory Tech.*, **44**, 2555–63.

# INDEX

- a posteriori* error estimate, 355; duality estimate, 359; numerical results, 363; residual estimate, 361
- ABC, *see* absorbing boundary condition
- absorbing boundary condition, 11, 365; error estimate, 366
- annihilator, 19
- arithmetic geometric mean inequality, 16
- assumptions on data; coefficients, 83; domain, 83; impedance, 84; source fields, 84
- asymptotic expansion of  $\Phi$ , 233
- Babuška–Brezzi condition; continuous, 22; discrete, 27
- backscattered RCS, 392
- barycentric coordinate, 109
- Bessel differential equation, 236; spherical, 239
- boundary component map, *see* Calderon operator
- boundary condition; impedance, 9; perfectly conducting, 9
- boundary inverse estimate, 152
- boundary projection  $P_\Sigma$ , 211
- boundary spaces, 150
- boundary to far field map  $\mathcal{B}$ , 419
- buried object, *see* scattering problem, layered medium
- Calderon Extension, 40
- Calderon operator; electric-to-magnetic, 249; exterior coercive  $G_e$ , 251; exterior electric-to-magnetic  $G_e$ , 249; exterior magnetic-to-electric  $G_m$ , 253; interior magnetic-to-electric  $G_i$ , 253; interior magnetic-to-electric  $G_i$ , 254; magnetic-to-electric, 252
- Cartesian coordinates, 425
- Cauchy–Schwarz inequality, 16
- cavity problem, 12; collective compactness, 180; discrete, 168; eigenvalues continuous, 13, 96 discrete, 195; ellipticized, 189 discrete, 191; error analysis via collective compactness, 176; error analysis via duality, 168; error estimate, 169, 187; existence, 95; numerical results, 188; scalar potential, 89; uniqueness, 92; variational, 83
- cavity resonances, *see* cavity problem, eigenvalues
- Cea's lemma, 25
- Clément interpolant, 147
- Clément macro-element, 147
- closed Hilbert space, 17
- closure, 17
- collectively compact, 32
- compact imbedding;  $X_0^*$  into  $(L^2(\Omega))^3$ , 286;  $X_0^*$  into  $(L^2(\Omega))^3$ , 268;  $X_0^*$  into  $(L^2(\Omega))^3$ , 87
- conditioning for small  $\varkappa$ , 193
- conductivity, 6
- conforming finite element, 105
- constitutive equations, 5
- continuity required of elements, 107
- continuous elements, danger, 191
- curl, 50
- curved domains, 209, 213; large elements, 214; method of Dubois, 210
- de Rham diagram; continuous, 65; discrete, 149
- Debye potential, 235
- degrees of freedom;  $H^1(\Omega)$  conforming hexahedra, 162 *hp*, 219 tetrahedra, 143, 209; curl conforming hexahedra, 158 *hp*, 220

- tetrahedra, 129, 205; definition, 102; divergence conforming hexahedra, 156  $hp$ , 222 tetrahedra, 119, 202; shorthand, 102
- dense subspace, 17
- density,  $X_b$  in  $X$ , 178
- Deny–Lions Theorem, 109
- dielectric, 6
- dipole source; explicit formula, 411; free space, 305; half space, 316; horizontal polarization, 325; vertical polarization, 322
- discrete compactness, 181, 292; of  $X_{0,p}$ , 183, 184
- discrete divergence, 192
- discrete divergence-free, 170; approximation by divergence-free, 173
- discrete Helmholtz decomposition, 170, 177
- dispersion error, *see* phase error
- distributional derivative, 37
- divergence, 50
- divergence theorem, 50
- domain, 37
- dot product, 2
- DtN map, *see* Calderon operator
- dual pairing, 19
- dual space, 19
- duality estimate, 174
- dyadic Green's function, *see* Green's dyadic
- edge element, 127, 158, 219; linear, 139; quadratic, 140
- eigenfunction, 24
- eigenvalue, 24
- element size parameters:  $b_K$  and  $\varrho_k$ , 112
- enhanced elements, 199
- Euclidean norm, 2
- far field equation, 397
- far field operator  $F$ , 397
- far field pattern, 233
- far field recovery, 386
- finite covering, 336
- finite element spaces; on curved domains, 216;  $S_{b,p}$ , 336;  $U_b$ , 145, 163;  $U_b^{(2)}$ , 209;  $U_{b,p}$ , 219;  $V_b$ , 134, 159;  $V_b^{(2)}$ , 207;  $V_{b,p-1}$ , 220;  $W_b$ , 124, 157;  $W_b^{(2)}$ , 204;  $W_{b,p-2}$ , 221;  $X_b$ , 168;  $X_{b,p}$ , 336;  $Z_b$ , 149, 164;  $Z_{b,p-3}$ , 222
- finite element, general definition, 101
- finite elements;  $H^1(\Omega)$  conforming  $hp$ , 218 hexahedra, 162 tetrahedra, 143, 209; curl conforming  $hp$ , 219 hexahedra, 158 tetrahedra, 126, 205; divergence conforming hexahedra, 155  $hp$ , 221 tetrahedra, 118, 202;  $L^2(\Omega)$  conforming hexahedra, 164 tetrahedra, 149; one dimensional, 101 estimates, 106
- first family of elements; hexahedra, 155; tetrahedra, 99
- Fourier space; error estimate, 289; inverse estimate, 290; on  $\Sigma$ , 289
- Fredholm alternative, 24
- Friedrichs inequality, 72, 88, 185
- function spaces; classical  $C^k(\Omega)$ , 36  $C_0^k(\Omega)$ , 36  $C^k\left(\vec{\Omega}\right)$ , 36  $L^p(\Omega)$ , 36; polynomial  $D_1(K)$ ,  $R_1(K)$ , 212  $D_k$ , 119  $P_k$ ,  $P_k'$ , 108  $P_k(\ell)$ ,  $P_k(f)$ , 108  $P_k'$ , 108  $Q_{l,m,n}$ , 109  $R_k$ , 128  $S_k$

128; Sobolev  $H^{\pm 1/2}(\partial\Omega)$ , 44  $H^1(\Omega)$ , 42  $H^s(\Omega)$ , 38  
 $H_0^s(\Omega)$ , 38  $L_{loc}^2(\Omega)$ , 45  $S$ , 85  $\hat{S}$  286  $\hat{S}$  265  $W^{s,p}(\Omega)$ , 37  
 $W_0^{s,p}(\Omega)$ , 38  $W^{s,p}(\partial\Omega)$  43; vector  $H_0(\text{curl}; \Omega)$ , 55  $H_0(\text{div}; \Omega)$ ,  
54  $H(\text{curl}; \Omega)$ , 55  $H(\text{div}; \Omega)$ , 52  $H_{imp}(\text{curl}; \Omega)$ , 69  $H_{loc}(\text{curl}; \mathbb{R}^3 \setminus \bar{\Omega})$ , 230  $H^{-1/2}(\text{Curl}; \gamma)$ , 59  $H^{-1/2}(\text{Curl}; \Gamma)$ , 59  
 $H^{-1/2}(\text{Div}; \Gamma)$ , 244  $H(\text{curl}; \Omega)$ , 55  $H(\text{Div}; \Gamma)$ , 244  $K_N(\Omega)$ , 67  $K_T(\Omega)$ , 67  $L_t^2(\partial\Omega)$ , 48  $X$ , 82  $X_0$ , 86  $\hat{X}_0$  286  $W_N$ ,  $W_p$   
 $X_N X_{N,0}$ ,  $X_p X_{T,0}$ , 71  $X^\sim$ , 263  $\hat{X}_0$ , 267  $Y(\Gamma)$ , 58, 410  
fundamental solution  $\Phi$  225  
Funk–Hecke formula, 241  
generalized Lax–Milgram lemma, 21  
generous overlap, 336  
geometric constraints on elements, 112  
gradient, 43  
Green's dyadic, 303; discrete, admissible, 307; layered medium, 321 first column, 325 third column, 322; perfectly conducting half space, 316  
Gårding inequality, 171  
health warning, 399  
Helmholtz decomposition, 65, 69, 86, 267, 286  
Herglotz wave function, 398, 414; approximation property, 415; characterization, 414; uniqueness, 414  
Hertz vector, 321  
Hilbert space, 16; compact, 23; relatively compact, 23  
Hilbert–Schmidt theorem, 24  
 $H$ -independent uniformity, 336  
Hodge operator, 172  
hp-finite elements, 217  
ill-posed/well-posed, 399  
imbedding, 40  
incident field, 9  
infinite element method, 370; discrete problem, 374  
inner product; boundary  $\langle \cdot, \cdot \rangle_s$ , 44  $\langle \cdot, \cdot \rangle$ , 82; volume  $(\cdot, \cdot)$ , 49  
integral identities, 50  
interface condition, 8  
interior cut, 65  
interpolant; definition  $\Pi_{clem}$ , 148  $\pi_h$ , 145, 163  $\pi_{h,p}$ , 219  $r_p$ , 134, 160, 207  $r_{h,p-1}$ , 220  $\omega_p$ , 124, 157, 204  $\omega_{h,p-2}$ , 222; error estimate  $\pi_p$ , 163  $\Pi_{clem}$ , 149  $\pi_p$ , 145, 164  $r_p$ , 136, 160, 208  $\omega_p$ , 124, 157, 204  
inverse problem, 394; linear sampling method, 397; uniqueness, 411  
Jacobi–Anger expansion, 241  
jump across a face  $[ \cdot ]_N$ , 359  
jump across a face  $[ \cdot ]_p$ , 358  
Laplace–Beltrami operator  $\delta_{\partial\Omega}$ , 49  
Lax–Milgram lemma, 20  
Legendre differential equation, 237; associated, 237  
Legendre function, associated, 238  
Legendre polynomials, 237  
Leis' variational method, 371  
linear sampling method,

- 397; implementation, 402; mathematics, 422; numerical results, 405
- Lipschitz domain, 38
- LSM, *see* linear sampling method
- matrix problem, 334
- Maxwell's equations; time dependent, 2; time harmonic, 7
- Mei series, 256, 259
- mesh parameter  $h$ , 112
- minimum rule, 218
- mixed problem, 22
- mixed reciprocity, 411
- Morozov discrepancy principle, 402
- NEA, *see* normalized echo area
- Neumann series, 23
- non-conforming elements, 200
- normal vector, 39
- normalized echo area, 392
- NtD map, *see* Calderon operator
- Ohm's law, 6
- operator; adjoint, 18; bounded, 18; collectively compact, 32; compact, 23; continuous, 18; dual, 19; linear, 18; norm, 18; nullspace, 18; pointwise convergent, 33; range, 18; self-adjoint, 24
- orthogonal complement, 17
- perfectly matched layer, 375; numerical results, 382; rectilinear, 377; spherical, 378; truncated domain, 380
- permeability, 5
- permeability, relative  $\mu_r$ , 6
- permittivity, 5
- permittivity, relative  $\epsilon_r$ , 6
- phase error, 344;  $\kappa$  dependence, 347; three dimensional edge elements, 351, 354
- plane wave, 9; polarization, 9
- PML, *see* perfectly matched layer
- Poincaré inequality, 46
- pointwise convergence, 178
- Poisson problem; Dirichlet boundary condition, 45; Neumann boundary condition, 46
- Poynting vector, 233
- preasymptotic convergence, 348
- preconditioned iteration, 335
- prismatic elements, 200
- projection  $P_\beta$ , 171, 275
- projection theorem, 17
- properties of  $H(\text{curl}; \Omega)$ , 55
- properties of  $H(\text{div}, \Omega)$ , 52
- pyramidal elements, 201
- quasi-uniform mesh on  $\Sigma$ , 152
- radar cross section, 392
- radiating solution, 230
- radiation condition; integral for layered medium, 12; Silver M\"ller, 10; Sommerfeld, 226, 240
- RCS, *see* radar cross section
- reciprocity, 418
- regular mesh, 116
- regularity; divergence-free projection, 182; elliptic problems, 45; Maxwell's equations, 69, 71, 283
- regularization, 399
- Rellich's lemma, 255
- Rellich's uniqueness theorem, 255
- residual based error estimator, 356
- resonance region, 1, 10
- Richardson iteration, 334
- Riesz theorem, 19
- Rodrigues' formula, 237
- scalar potential, 61
- Scalar product, 15
- scattered field, 9
- scattering problem; Babuška–Brezzi condition, 273; bounded scatterer, 13; convergence, 277; discrete, 275; discrete Babuška–Brezzi condition, 275; domain decomposition, 281; existence, 272, 289; half space, 315; layered medium, 14, 318; non-homogeneous, 272; overlapping scheme,

307; overlapping, discrete, 308; uniqueness, 256, 264,  
 288; variational, 263  
 Schwarz iteration, 337  
 second family of elements, tetrahedra, 202  
 series expansion for  $E^s$ , 246, 248  
 sesquilinear form; bounded, 20; coercive, 20; definition, 20;  
 $A$ , 265;  $a$ , 83;  $a_1, b_1$ , 265;  $a_2, b_2$ , 270;  $a_+$ , 89;  $\tilde{a}$ , 87;  $\tilde{b}$ , 89  
 singular system, 400  
 Sobolev imbedding theorem, 41  
 solution of the linear system, 333  
 speed of light, 5  
 spherical Bessel function; asymptotics, 239, 240  
 spherical Bessel functions, 239  
 spherical coordinates, 425  
 spherical harmonic, 236, 238  
 starlike, 56  
 Stoke's theorem, 52  
 Stratton–Chu formula, 228, 230, 304  
 strong convergence, 16  
 super-convergence, 188, 201  
 surface divergence  $\nabla_{\partial\Omega}$ ; 48  
 surface gradient  $\nabla_{\partial\Omega}$ , 48  
 surface scalar curl  $\nabla_{\partial\Omega} \times$ , 49  
 surface vector curl  $\vec{\nabla}_{\partial\Omega} \times$ , 49  
 Tikhonov regularization, 400  
 time harmonic field, 3  
 total field, 10  
 trace;  $\gamma 0$  for  $H^1(\Omega)$ , 43;  $\gamma t, \gamma T$  for  $H(\text{curl}; \Omega)$ , 57;  $\gamma n$  for  $H(\text{div}; \Omega)$ , 53  
 transformation; affine, 113; curl preserving, 77; divergence preserving, 79; for scalar functions, 77  
 unique continuation, 92  
 unisolvant element, 102  
 vector addition theorem, 245  
 vector potential, 63  
 vector spherical harmonics, 241  
 vector wave functions; exterior, 245; interior, 245  
 wavelength, 10, 344  
 wavenumber, 7  
 wavenumber dependence of error, 345  
 weak convergence, 16  
 Wesson's trick for orienting edges, 143  
 Whitney element, 139  
 Wilcox expansion, 366  
 Wronskian identity, 240  
 Z-coercivity, 21  
 Z-coercivity, 26