

Gary C. Cohen

Higher-Order Numerical Methods for Transient Wave Equations

Scientific
Computation



Springer

Scientific Computation

Editorial Board

J.-J. Chattot, Davis, CA, USA
P. Colella, Berkeley, CA, USA
R. Glowinski, Houston, TX, USA
M. Holt, Berkeley, CA, USA
Y. Hussaini, Tallahassee, FL, USA
P. Joly, Le Chesnay, France
H. B. Keller, Pasadena, CA, USA
D. I. Meiron, Pasadena, CA, USA
O. Pironneau, Paris, France
A. Quarteroni, Lausanne, Switzerland
J. Rappaz, Lausanne, Switzerland
R. Rosner, Chicago, IL, USA
J. H. Seinfeld, Pasadena, CA, USA
A. Szepessy, Stockholm, Sweden

Springer-Verlag Berlin Heidelberg GmbH



<http://www.springer.de/phys/>

Scientific Computation

A Computational Method in Plasma Physics

F. Bauer, O. Betancourt, P. Garabedian

Implementation of Finite Element Methods for Navier-Stokes Equations

F. Thomasset

Finite-Different Techniques for Vectorized Fluid Dynamics Calculations

Edited by D. Book

Unsteady Viscous Flows

D. P. Telionis

Computational Methods for Fluid Flow

R. Peyret, T. D. Taylor

Computational Methods in Bifurcation Theory and Dissipative Structures

M. Kubicek, M. Marek

Optimal Shape Design for Elliptic Systems

O. Pironneau

The Method of Differential Approximation

Yu. I. Shokin

Computational Galerkin Methods

C. A. J. Fletcher

Numerical Methods for Nonlinear Variational Problems

R. Glowinski

Numerical Methods in Fluid Dynamics

Second Edition M. Holt

Computer Studies of Phase Transitions and Critical Phenomena O. G. Mouritsen

Finite Element Methods in Linear Ideal Magnetohydrodynamics

R. Gruber, J. Rappaz

Numerical Simulation of Plasmas

Y. N. Dnestrovskii, D. P. Kostomarov

Computational Methods for Kinetic Models of Magnetically Confined Plasmas

J. Killeen, G. D. Kerbel, M. C. McCoy,

A. A. Mirin

Spectral Methods in Fluid Dynamics

Second Edition C. Canuto, M. Y. Hussaini,
A. Quarteroni, T. A. Zang

Computational Techniques

for Fluid Dynamics 1

Fundamental and General Techniques

Second Edition C. A. J. Fletcher

Computational Techniques

for Fluid Dynamics 2

Specific Techniques

for Different Flow Categories

Second Edition C. A. J. Fletcher

Methods for the Localization of Singularities in Numerical Solutions

of Gas Dynamics Problems

E. V. Vorozhtsov, N. N. Yanenko

Classical Orthogonal Polynomials of a Discrete Variable

A. F. Nikiforov, S. K. Suslov, V. B. Uvarov

Flux Coordinates and Magnetic Filed Structure: A Guide to a Fundamental Tool

of Plasma Theory

W. D. D'haeseleer, W. N. G. Hitchon,

J. D. Callen, J. L. Shohet

Monte Carlo Methods in Boundary Value Problems

K. K. Sabelfeld

The Least-Squares Finite Element Method

Theory and Applications in Computational
Fluid Dynamics and Electromagnetics

Bo-nan Jiang

Computer Simulation of Dynamic Phenomena

M. L. Wilkins

Grid Generation Methods

V. D. Liseikin

Radiation in Enclosures

A. Mbiock, R. Weber

Large Eddy Simulation for Incompressible Flows

An Introduction

P. Sagaut

Higher-Order Numerical Methods for Transient Wave Equations

G. C. Cohen

A selection of titles; for further information see:

Series homepage – <http://www.springer.de/phys/books/sc/>

Gary C. Cohen

Higher-Order Numerical Methods for Transient Wave Equations

With 100 Figures



Springer

Dr. Gary C. Cohen
INRIA
Domaine de Voluceau
Rocquencourt, B.P. 105
78153 Le Chesnay Cedex
France
gary.cohen@inria.fr

Library of Congress Cataloging-in-Publication Data

Cohen, Gary (Gary C.)
Higher-order numerical methods for transient wave equations / Gary C. Cohen.
p. cm. -- (Scientific computation, ISSN 1434-8322)
Includes bibliographical references and index.
ISBN 978-3-642-07482-0 ISBN 978-3-662-04823-8 (eBook)
DOI 10.1007/978-3-662-04823-8
1. Wave equation--Numerical solutions. 2. Numerical analysis. I. Title. II. Series.

QA927 .C54 2002
530.12'4--dc21

2001020983

ISSN 1434-8322
ISBN 978-3-642-07482-0

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag Berlin Heidelberg GmbH.

Violations are liable for prosecution under the German Copyright Law.

<http://www.springer.de>

© Springer-Verlag Berlin Heidelberg 2002
Originally published by Springer-Verlag Berlin Heidelberg New York in 2002
Softcover reprint of the hardcover 1st edition 2002

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Typesetting by the author using a Springer TeX macro package.
Cover design: *design & production* GmbH, Heidelberg

Printed on acid-free paper SPIN: 10790356 55/3141/mf - 5 4 3 2 1 0

To the Lubavitcher Rebbe,
to my father,
to my mother,
who are no longer in this
physical world but so deeply
present in my soul

Foreword

Problems involving the solution of wave-type equations are quite common in science and engineering. Let us mention, for example, electromagnetics, acoustics, elastodynamics, geophysics, but many other applications do exist (in fluid dynamics, for example). Dr. Gary Cohen has spent almost two decades, mostly at INRIA, developing computational methods for the solution of these wave problems. Thus, some of us thought that it was time to ask him to report the results of his own and his collaborators' investigations, in a volume of this Springer series. Dr. Cohen accepted the challenge, producing this very fine volume. Indeed, as one of the editors of these series, we are most pleased by both the clarity of the presentation and a wonderful blending of theoretical and computational considerations. Of course we also highly appreciate the fact that some very recent developments can be found in this book, such as:

- The use of high-order finite element approximations leading, nevertheless, to diagonal mass matrices (Chaps. 11–13).
- A well-written introduction to the theory of perfectly matched layers (PML) in order to construct transparent boundary conditions when simulating wave phenomena in unbounded physical domains.

There is more to compliment in this book, but the reader will discover these other gems by herself or himself. Indeed, we have no doubt that those readers will find Dr. Cohen's book extremely useful from both the theoretical and practical points of view.

To conclude on a personal note, let us say that we want to believe that the finite element graduate course that Dr. Cohen took from us at University Paris VI in the mid-1970s was one of his motivations to pursue a career in applied and computational mathematics. The present book is one of the great achievements of his career and hopefully not the last.

Houston, Texas, November 2000

R. Glowinski

Preface

This book is the result of more than 15 years of research at INRIA (Institut National de Recherche en Informatique et Automatique), in collaboration with French industries such as IFP, SNEA(P), Dassault-Aviation, EDF, CEA, etc. These companies have proposed challenging problems which have led to the development of a substantial number of the methods described in this book.

As the title indicates, this book is devoted to the numerical modeling of transient wave equations (also called, by engineers, wave equations in the time domain) rather than in the frequency domain. Both approaches have the common advantage of modeling wave propagation phenomena by solving partial differential equations (PDE) which are able to take into account a large range of physical phenomena such as scattering, layered media, semi-reflecting obstacles, anisotropy, etc. However, wave equations in the time domain can deal with a much wider range of applications than is possible in the frequency domain. This is the reason for the increasing interest in this approach over the last three decades. The first applications were proposed by geophysicists who used pulse sources which cannot be modeled by the frequency domain approach that requires harmonic sources. A little later, the electromagnetism community also started to use time domain methods because of the need to model more sophisticated radar sources containing a large range of frequencies.

For reasonably low frequencies, numerical approximation is the most convenient way to solve the PDEs which model wave propagation. The simplest approach is the finite difference method which is easy to implement and which uses regular grids that are particularly suitable for wave propagation. However, for more realistic configurations, the finite element method seems more attractive. Nevertheless, engineers and researchers were rather reluctant to use this approach because of its high cost. This barrier was broken during the last decade by the use of mass-lumping via Gauss-Lobatto quadrature rules and, more recently, by the use of mixed formulations of the wave equations.

I intend this book to be a useful tool for graduate students, researchers and engineers who want to become acquainted with numerical modeling of wave equations or would like to expand and update their knowledge of this subject. On the one hand, beginners in this area will find basic concepts

for approximating linear wave equations and classical tools for analyzing the resulting models. On the other hand, people already involved in this topic will find, besides a unified presentation of the subject, new methods such as mass-lumped (mixed) finite elements, as well as new results on classical and modern approximation schemes.

This book is divided into three parts. The first part is devoted to a presentation of the equations treated in the remainder of the book and the (minimal) mathematical ideas needed to understand the techniques of approximation and analysis of these equations.

The second part deals with finite difference approximation. Since this approximation is much easier to describe than the finite element method, it is used as a vehicle for describing the basic methods of analysis of numerical models such as the dispersion relation, reflection-transmission analysis, etc.

In the third part of the book, finite element methods with mass-lumping are constructed and the tools of analysis described in the second part are extended to this approach. Here we encounter additional difficulties such as numerical parasitic waves. In order to provide a complete presentation of the subject, we give, in the last chapter a survey of different methods for modeling unbounded domains with an emphasis on perfectly matched layers (PML). Nowadays, these seem to be the most general and the most efficient solution to this important problem.

Of course, such a book cannot be the work of only one person. Therefore, I shall end this foreword by thanking my colleagues, friends, students, etc., who worked with me and contributed to my research during the two last decades.

First, I would like to thank Alain Bamberger and Patrick Joly who motivated my interest in numerical modeling of wave equations. Alain left INRIA for IFP after one year of collaboration but Patrick remained by my side in the INRIA Ondes project which he leads. His clear and deep explanations based on his encyclopedic knowledge have always been a source of inspiration for me and my colleagues at INRIA.

A very special mention is merited by my friend Peter Monk who attracted me to the area of the approximation of the Maxwell equations and with whom I have had a fruitful collaboration for almost 10 years. As a true friend (i.e. a person you can call on when needed), he accepted the tedious task of reading my manuscript and correcting the English language of a French writer¹ (including this preface). His valuable suggestions and corrections were helpful to me.

I would also like to thank Dimitri Komatitsch for his references and his help in the area of geophysics.

¹ Once, an American researcher said to a French lecturer in an international conference: “You French people should give your talks in French so that, at least, French participants could understand”.

The Talmudic Sages assert: "I learnt a lot from my masters, a lot from my colleagues but from my pupils more than all of them". This is the reason why I must thank my students whose studies contributed so much to my understanding of the methods I developed: Nathalie Tordjman, Michel Barbiera, Alexandre Elmkies, Leila Rhaouti, Gary Cochard, and Sébastien Sonnenberg. In particular, I would like to thank Sandrine Fauqueux whose contribution was fundamental for the understanding of the mixed formulations of the acoustics and elastics equations and of their modeling in unbounded domains.

Last, but not least, I would like to thank Prof. Roland Glowinski who was my professor in my graduate classes and who was always by my side when needed. He has honored me by writing a foreword to this book which he suggested I write.

My preface would not be complete if I did not thank my wife Yaffa who supported me throughout the difficult task of writing this book and my children and grandchildren who gave me their smiles when I was tired.

Rocquencourt, August 2001

Gary C. Cohen

Contents

Part I. Basic Definitions and Properties

1. The Basic Equations	3
1.1 The Acoustics Equation	3
1.2 The Maxwell Equations	4
1.2.1 The 3D Case	4
1.2.2 The 2D Case	6
1.3 The Elastics System	7
1.3.1 General Formulation	7
1.3.2 The Isotropic Case	8
1.4 Boundary Conditions	11
1.4.1 The Wave Equation	11
1.4.2 The Maxwell Equations	12
1.4.3 The Elastics System	12
2. Functional Issues	15
2.1 Some Functional Spaces	15
2.1.1 Sobolev Spaces	15
2.1.2 Spaces $H(\mathbf{curl}, \Omega)$ and $H(\text{div}, \Omega)$	17
2.2 Variational Formulations	18
2.2.1 The Acoustics Equation	18
2.2.2 The Maxwell Equations	20
2.2.3 The Elastics System	22
2.3 Energy Identities	22
2.3.1 The Acoustics Equation	23
2.3.2 The Maxwell Equations	23
2.3.3 The Elastics System	24
3. Plane Wave Solutions	25
3.1 A General Solution of the Homogeneous Wave Equation	25
3.2 Application to the Maxwell Equations	26
3.2.1 The 3D Case	26
3.2.2 The 2D Case	28
3.3 Application to the Elastics System	29

Part II. Finite Difference Methods

4. Construction of the Schemes in Homogeneous Media	35
4.1 A Model Problem	35
4.2 Second-Order Approximation in Space	35
4.2.1 The 1D Case	35
4.2.2 The 2D Case	38
4.3 Fourth-Order Approximations in Space	41
4.3.1 First Approach: Global Approximation of Δ	41
4.3.2 Second Approach: Fourth-Order Approximations of the First-Order Operators	43
4.4 Approximation in Time	45
4.4.1 The Modified Equation Approach	46
4.4.2 Symmetric Schemes	48
4.5 Higher-Order Approximations in Space	50
4.5.1 First Approach	50
4.5.2 Second Approach	52
4.5.3 Extension to Higher Dimensions	54
4.6 Higher-Order Approximations in Time	56
4.6.1 The Modified Equation Approach	56
4.6.2 Symmetric Schemes	57
4.7 Extension to Systems	58
4.7.1 The Maxwell Equations	58
4.7.2 The Elastics System	60
4.7.3 Higher-Order Approximation in Time	62
5. The Dispersion Relation	65
5.1 Second-Order Schemes for the Wave Equation	65
5.1.1 Using Plane Wave Solutions	65
5.1.2 Computation by the Discrete Fourier Transform	66
5.1.3 Symbol of an Operator	68
5.2 Higher-Order Approximations in Space	69
5.2.1 The First Approach	69
5.2.2 The Second Approach	72
5.3 Approximation in Time	74
5.3.1 Second-Order Approximation in Time	74
5.3.2 The Modified Equation Approach	75
5.3.3 Symmetric Schemes	76
5.4 The Case of Systems	78
5.4.1 The Maxwell Equations	78
5.4.2 The Elastics System	80

6. Stability of the Schemes	83
6.1 General Presentation	83
6.2 Positivity of an Operator	84
6.3 Second-Order Approximation in Time	85
6.3.1 Second-Order Approximation in Space	86
6.3.2 A Basic Property	86
6.3.3 Application to Higher-Order Approximation in Space .	87
6.4 The Modified Equation Approach	89
6.4.1 Preliminary Results	89
6.4.2 Fourth-Order Approximation in Space: First Approach	90
6.4.3 Fourth-Order Approximation in Space: Second Approach	92
6.5 Symmetric Schemes	95
6.5.1 First Method	95
6.5.2 Second Method	97
6.6 The Case of Systems	97
6.7 A Numerical Illustration	99
7. Numerical Dispersion and Anisotropy	101
7.1 Phase and Group Velocities	101
7.2 The Concept of Numerical Dispersion	102
7.3 Order of the Numerical Dispersion	103
7.3.1 Schemes Semi-Discrete in Space	104
7.3.2 Fully Discrete Schemes: Second-Order in Time	104
7.3.3 Fully Discrete Schemes: The Modified Equation Approach	106
7.3.4 Fully Discrete Schemes: The Fourth-Order Symmetric Scheme	106
7.3.5 Error Committed on the Group Velocities	108
7.4 Change of Variables	109
7.5 Some Useful Properties of the Schemes	111
7.5.1 Relation between 1D and Higher Dimensions	111
7.5.2 Two Remarkable Schemes	113
7.6 Dispersion Curves	114
7.6.1 Second and Fourth-Order Schemes, Semi-Discrete in Space	114
7.6.2 Schemes, Second-Order in Time and Fourth-Order in Space	114
7.6.3 Schemes, Fourth-Order in Time and Space	115
7.6.4 Comparison with Higher-Order Schemes in Space	118
7.7 Isotropy Curves	118
7.8 The Elastics System	118

8. Construction of the Schemes in Heterogeneous Media	123
8.1 A General Framework	123
8.2 Second-Order Approximations	123
8.3 Higher-Order Approximations: First Approach	124
8.4 Higher-Order Approximations: Second Approach	126
8.5 The Case of Arakawa's Scheme	128
8.6 Approximation in Time	129
8.6.1 Second-Order Approximation in Time	129
8.6.2 The Heterogeneous Modified Equation	130
8.6.3 Expression of the Correction Term	131
8.7 Approximation of the Boundary Conditions	133
8.7.1 Second-Order Schemes	133
8.7.2 Higher-Order Schemes	134
8.7.3 Extension to Systems	136
9. Stability by Energy Techniques	137
9.1 Positivity of the Discrete Operators	137
9.1.1 Second-Order and Second Approach for Fourth-Order Approximations	137
9.1.2 Fourth-Order: First Approach	138
9.2 Stability Conditions	142
9.2.1 A General Framework	142
9.2.2 Computation of $\ A_h\ $ for the Second-Order Approximation	143
10. Reflection-Transmission Analysis	145
10.1 The 1D Case	145
10.1.1 The Continuous Problem	145
10.1.2 Second-Order Approximation	146
10.1.3 Fourth-Order: First Approach	147
10.1.4 A Numerical Study	153
10.2 The 2D Case	156
10.2.1 The Continuous Problem	156
10.2.2 Second-Order Scheme	158
10.2.3 A Numerical Experiment	162

Part III. Finite Element Methods

11. Mass-Lumping in 1D	169
11.1 Basic Approximations	169
11.1.1 Construction of Mass-Lumped Finite Elements	169
11.1.2 Approximation in Time	177
11.2 Dispersion Relations	179
11.2.1 P_2 Finite Elements	179
11.2.2 P_3 and Higher-Order Finite Elements	181

11.3 Stability Analysis	186
11.3.1 The Leapfrog Scheme	186
11.3.2 The Modified Equation Approach	189
11.3.3 Symmetric Schemes	191
11.4 Dispersion Analysis	192
11.4.1 Taylor Expansions	192
11.4.2 Dispersion Curves	193
11.5 Some Results on the Amplitudes	194
11.5.1 Error Estimates on the Physical Part of the Solution . .	200
11.5.2 Error Estimates on the Parasitic Part of the Solution .	202
11.6 Reflection-Transmission Analysis	202
11.6.1 FEM Approximation of the Heterogeneous Wave Equation	203
11.6.2 Taylor Expansion of the Wavenumber	203
11.6.3 Interface Between Two Elements	204
11.6.4 Interface at an Interior Point	205
11.6.5 Extension to Higher-Order Approximations	206
11.7 Taylor Expansions of the Eigenvectors	208
12. Spectral Elements	211
12.1 Construction of Quadrilateral and Hexahedral Finite Elements	211
12.1.1 Reference Spectral Elements	211
12.1.2 Extension to Quadrilateral Meshes	214
12.1.3 Extension to Hexahedral Meshes	220
12.2 Plane Wave Analysis of Regular Meshes	222
12.2.1 Decomposition of the Discrete Equations	222
12.2.2 Decomposition of the Eigenvalues and Eigenvectors . .	225
12.3 Some Analysis of Non-Regular Meshes in 2D	230
12.3.1 Dispersion Analysis	230
12.3.2 Numerical Study of the Stability	231
12.3.3 Numerical Study of the Accuracy	234
12.3.4 A Two-Layer Experiment	238
12.4 Triangular and Tetrahedral Meshes	244
12.4.1 The Basic Problem	244
12.4.2 A New Family of Triangular Elements	250
12.4.3 Tetrahedral Elements	254
12.4.4 Non-Conforming Triangular Elements	257
12.5 A Numerical Illustration	259
13. Mass-Lumped Mixed Formulations and Edge Elements . . .	261
13.1 Variational Extensions of the Yee Scheme	261
13.1.1 The Model Problem	261
13.1.2 A First Family of Hexahedral Edge Elements	262
13.1.3 The 2D Case	269

13.2 Efficient Edge Elements for the Maxwell Equations	271
13.2.1 Extension to Anisotropic Media and Complex Geometries	271
13.2.2 New Spaces of Approximation	273
13.2.3 Basis Functions and Degrees of Freedom	274
13.2.4 The Mass Integral	277
13.2.5 The Stiffness Integral	279
13.2.6 An Efficient Alternative	280
13.2.7 The 2D Case	283
13.2.8 A 2D Numerical Experiment	284
13.3 Triangular and Tetrahedral Edge Elements	284
13.3.1 Triangular Edge Elements	286
13.3.2 Tetrahedral Edge Elements	289
13.3.3 Spaces of Approximation	292
13.4 A New Formulation of Spectral Elements	294
13.4.1 A New Approximation of the Wave Equation	294
13.4.2 A Theorem of Equivalence	296
13.4.3 Extension to the Elastics System	301
14. Modeling Unbounded Domains	307
14.1 History of the Problem	307
14.2 Perfectly Matched Layers	308
14.2.1 Presentation of the Method	308
14.2.2 Construction of the PML in 2D	311
14.2.3 The Three-Dimensional Case	317
14.2.4 Finite Element Approximation	320
14.2.5 Approximation in Time	321
14.3 Numerical Illustrations	322
14.3.1 PML for the 2D Elastics System	322
14.3.2 Scattering by an Ogive of an Electromagnetic Wave	323
14.3.3 A “Foothills” Experiment	324
14.4 Computational Issues	329
14.4.1 Tetrahedra or Hexahedra?	329
14.4.2 Finite Element or Finite Difference Methods?	330
A. Appendix: Construction of a General $H(\mathbf{curl})$-Conforming Transform	331
A.1 A Local $H(\mathbf{curl}, \hat{K})$ -Conforming Isomorphism	331
A.1.1 Notation	331
A.1.2 A Local $H(\mathbf{curl}, \hat{K})$ -Conforming Isomorphism	332
A.2 A Global $H(\mathbf{curl}, \Omega)$ -Conforming Isomorphism	336
A.2.1 Notation	336
A.2.2 A Global $H(\mathbf{curl}, \Omega)$ -Conforming Isomorphism	337
Bibliography	341
Index	347

Part I

Basic Definitions and Properties

1. The Basic Equations

The equations that model wave propagation can be classified into three physical categories. The acoustics equation and the elastics system model mechanical waves in fluids and solids, respectively, and the Maxwell equations describe the propagation of electromagnetic waves (radio waves or light).

Several models for these three kinds of waves, which have various degrees of complexity, are given in the scientific literature. The purpose of this book is to construct numerical models for linear wave propagation in heterogeneous and sometimes even anisotropic media.

The three main models we are going to consider are given in the following sections. Throughout the book, we use the notation¹ $x \in \mathbb{R}$ or $\mathbf{x} \in \mathbb{R}^d$, $d = 2, 3$ to denote spatial position and $t \in \mathbb{R}^+$ to denote time. Moreover, in all the following equations, when used with vectors, the symbols \cdot and \times will denote the scalar and cross products. In particular, these products will be applied to vector valued differential operators.

1.1 The Acoustics Equation

Let u denote the acoustic pressure field and ρ the density of the medium in which the acoustic wave travels. In its scalar form, the acoustics equation² can be written as follows

$$\frac{1}{\kappa(\mathbf{x})} \frac{\partial^2 u}{\partial t^2}(\mathbf{x}, t) - \nabla \cdot \left(\frac{1}{\rho(\mathbf{x})} \nabla u(\mathbf{x}, t) \right) = f(\mathbf{x}, t), \quad (1.1)$$

where $\nabla = (\partial/\partial x_1, \dots, \partial/\partial x_d)^T$, and ρ and κ are strictly positive functions of position \mathbf{x} . This equation can be written as a first-order system by introducing the velocity \mathbf{v} as follows:

$$\frac{1}{\kappa(\mathbf{x})} \frac{\partial u}{\partial t}(\mathbf{x}, t) = \nabla \cdot \mathbf{v}(\mathbf{x}, t) + F(\mathbf{x}, t), \quad (1.2a)$$

$$\rho(\mathbf{x}) \frac{\partial \mathbf{v}}{\partial t}(\mathbf{x}, t) = \nabla u(\mathbf{x}, t), \quad (1.2b)$$

¹ Bold characters will indicate a vector of \mathbb{R}^d .

² Derived from the Euler equations.

where $F(\mathbf{x}, t) = \int_0^t f(\mathbf{x}, \tau) d\tau$.

The velocity c of sound wave propagation is given by the relation³

$$c(\mathbf{x}) = \sqrt{\frac{\kappa(\mathbf{x})}{\rho(\mathbf{x})}}. \quad (1.3)$$

In order to obtain well-posed equations, we must add the initial conditions

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}), \quad \frac{\partial u}{\partial t}(\mathbf{x}, 0) = u_1(\mathbf{x}) \quad (1.4)$$

to (1.1) and

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}), \quad \mathbf{v}(\mathbf{x}, 0) = \mathbf{v}_0(\mathbf{x}) \quad (1.5)$$

to (1.2a) and (1.2b) where u_0 , u_1 and \mathbf{v}_0 are given functions.

1.2 The Maxwell Equations

1.2.1 The 3D Case

We denote by \mathbf{E} , \mathbf{H} , \mathbf{D} and \mathbf{B} the electric and magnetic fields and the electric and magnetic inductions, respectively. The equations of electromagnetism considered in this book can be written as

$$\frac{\partial \mathbf{D}}{\partial t}(\mathbf{x}, t) - \nabla \times \mathbf{H}(\mathbf{x}, t) = -\mathbf{J}(\mathbf{x}, t), \quad (1.6a)$$

$$\frac{\partial \mathbf{B}}{\partial t}(\mathbf{x}, t) + \nabla \times \mathbf{E}(\mathbf{x}, t) = 0, \quad (1.6b)$$

$$\mathbf{D} = \underline{\underline{\epsilon}}(\mathbf{x})\mathbf{E}, \quad (1.6c)$$

$$\mathbf{B} = \underline{\underline{\mu}}(\mathbf{x})\mathbf{H}, \quad (1.6d)$$

where $\underline{\underline{\epsilon}}$ and $\underline{\underline{\mu}}$ are the symmetric, positive definite, dielectric permittivity and magnetic permeability matrices depending on space which model anisotropic media⁴, and \mathbf{J} is the current density.

Moreover, the fields \mathbf{D} and \mathbf{B} satisfy the two relations

$$\operatorname{div} \mathbf{D} = \rho, \quad (1.7a)$$

$$\operatorname{div} \mathbf{B} = 0, \quad (1.7b)$$

where ρ is the charge density⁵.

³ A more appropriate term would be “celerity” but “velocity” is widely used in literature.

⁴ The double underline indicates a matrix or a tensor.

⁵ When \mathbf{D} and \mathbf{E} and \mathbf{B} and \mathbf{H} are not explicitly related one to the other as in (1.6c) and (1.6d), the Maxwell system is composed of equations (1.6a), (1.6b), (1.7a) and (1.7b).

Of course, the isotropic case is obtained when $\underline{\varepsilon} = \varepsilon I_3$ and $\underline{\mu} = \mu I_3$, where ε and μ are strictly positive scalar functions and I_3 is the identity matrix of \mathbb{R}^3 .

We shall use (1.6c) and (1.6d) to eliminate \mathbf{B} and \mathbf{D} from the Maxwell system. Hence, in the remainder of the book we shall generally deal with equations involving \mathbf{E} and \mathbf{H} .

The initial conditions for the Maxwell equations are

$$\mathbf{E}(\mathbf{x}, 0) = \mathbf{E}_0(\mathbf{x}), \quad \mathbf{H}(\mathbf{x}, 0) = \mathbf{H}_0(\mathbf{x}). \quad (1.8)$$

By combining (1.6a)–(1.6d), we obtain the second-order forms of the Maxwell system:

$$\underline{\varepsilon}(\mathbf{x}) \frac{\partial^2 \mathbf{E}}{\partial t^2}(\mathbf{x}, t) + \nabla \times (\underline{\mu}^{-1}(\mathbf{x}) \nabla \times \mathbf{E}(\mathbf{x}, t)) = -\mathbf{j}(\mathbf{x}, t), \quad (1.9)$$

$$\underline{\mu}(\mathbf{x}) \frac{\partial^2 \mathbf{H}}{\partial t^2}(\mathbf{x}, t) + \nabla \times (\underline{\varepsilon}^{-1}(\mathbf{x}) \nabla \times \mathbf{H}(\mathbf{x}, t)) = \mathbf{J}'(\mathbf{x}, t), \quad (1.10)$$

where $\mathbf{j} = \partial \mathbf{J} / \partial t$ and $\mathbf{J}' = \nabla \times (\underline{\varepsilon}^{-1} \mathbf{J})$.

Now, if we assume that the waves propagate in a homogeneous isotropic medium (vacuum for instance) far enough from its source, so that $\mathbf{J} = 0$ and $\rho = 0$, (1.9) and (1.10) can be rewritten as

$$\frac{\partial^2 \mathbf{E}}{\partial t^2}(\mathbf{x}, t) + \frac{1}{\varepsilon \mu} \nabla \times \nabla \times \mathbf{E}(\mathbf{x}, t) = 0, \quad (1.11)$$

$$\frac{\partial^2 \mathbf{H}}{\partial t^2}(\mathbf{x}, t) + \frac{1}{\varepsilon \mu} \nabla \times \nabla \times \mathbf{H}(\mathbf{x}, t) = 0, \quad (1.12)$$

by using the fact that $\nabla \times \nabla \times \mathbf{V} = \mathbf{grad}(\operatorname{div} \mathbf{V}) - \Delta \mathbf{V}$ and taking into account relations (1.7a)–(1.7b), we obtain:

$$\frac{\partial^2 \mathbf{E}}{\partial t^2}(\mathbf{x}, t) - \frac{1}{\varepsilon \mu} \Delta \mathbf{E}(\mathbf{x}, t) = 0, \quad (1.13)$$

$$\frac{\partial^2 \mathbf{H}}{\partial t^2}(\mathbf{x}, t) - \frac{1}{\varepsilon \mu} \Delta \mathbf{H}(\mathbf{x}, t) = 0. \quad (1.14)$$

In other words, each component of \mathbf{E} and \mathbf{H} satisfies the wave equation with a velocity equal to $1/\sqrt{\varepsilon \mu}$.

More generally, the velocity c in a non-homogeneous isotropic medium is defined by

$$c^2(\mathbf{x}) \varepsilon(\mathbf{x}) \mu(\mathbf{x}) = 1. \quad (1.15)$$

1.2.2 The 2D Case

The Transverse-Magnetic (TM) Case. By considering an electric field polarized in the plane (x_1, x_2) and a magnetic field polarized in the direction orthogonal to this plane and assuming that both fields are independent of x_3 , we obtain

$$\underline{\varepsilon}(\mathbf{x}) \frac{\partial \mathbf{E}}{\partial t}(\mathbf{x}, t) - \mathbf{curl} H(\mathbf{x}, t) = -\mathbf{J}(\mathbf{x}, t), \quad (1.16a)$$

$$\mu(\mathbf{x}) \frac{\partial H}{\partial t}(\mathbf{x}, t) + \mathbf{curl} \mathbf{E}(\mathbf{x}, t) = 0, \quad (1.16b)$$

where $\mathbf{curl} H = (\partial H / \partial x_2, -\partial H / \partial x_1)^T$ and $\mathbf{curl} \mathbf{E} = \partial E_y / \partial x_1 - \partial E_x / \partial x_2$, and its two second-order versions

$$\underline{\varepsilon}(\mathbf{x}) \frac{\partial^2 \mathbf{E}}{\partial t^2}(\mathbf{x}, t) + \mathbf{curl} \left(\frac{1}{\mu(\mathbf{x})} \mathbf{curl} \mathbf{E}(\mathbf{x}, t) \right) = -\mathbf{j}(\mathbf{x}, t), \quad (1.17)$$

$$\mu(\mathbf{x}) \frac{\partial^2 H}{\partial t^2}(\mathbf{x}, t) + \mathbf{curl} \left(\underline{\varepsilon}^{-1}(\mathbf{x}) \mathbf{curl} H(\mathbf{x}, t) \right) = J'(\mathbf{x}, t), \quad (1.18)$$

where \mathbf{j} and J' are defined as in (1.9) and (1.10).

The Transverse-Electric (TE) Case. In this case, the polarizations of the electric and magnetic fields are switched. We obtain:

$$\underline{\varepsilon}(\mathbf{x}) \frac{\partial E}{\partial t}(\mathbf{x}, t) - \mathbf{curl} \mathbf{H}(\mathbf{x}, t) = -J(\mathbf{x}, t), \quad (1.19a)$$

$$\underline{\mu}(\mathbf{x}) \frac{\partial \mathbf{H}}{\partial t}(\mathbf{x}, t) + \mathbf{curl} \mathbf{E}(\mathbf{x}, t) = 0, \quad (1.19b)$$

where \mathbf{curl} and \mathbf{curl} are defined as for the TM case, and its two second-order versions

$$\underline{\varepsilon}(\mathbf{x}) \frac{\partial^2 E}{\partial t^2}(\mathbf{x}, t) + \mathbf{curl} \left(\underline{\mu}^{-1}(\mathbf{x}) \mathbf{curl} \mathbf{E}(\mathbf{x}, t) \right) = -j(\mathbf{x}, t), \quad (1.20)$$

$$\underline{\mu}(\mathbf{x}) \frac{\partial^2 \mathbf{H}}{\partial t^2}(\mathbf{x}, t) + \mathbf{curl} \left(\frac{1}{\underline{\varepsilon}(\mathbf{x})} \mathbf{curl} \mathbf{H}(\mathbf{x}, t) \right) = \mathbf{J}'(\mathbf{x}, t). \quad (1.21)$$

Remarks

1. The right-hand side \mathbf{J} can take the form $\mathbf{J} = \mathbf{J}_1 + \underline{\sigma} \mathbf{E}$, where $\underline{\sigma}$ is the conductivity. This additional term introduces physical damping. In this case, we can only obtain a second-order formulation in \mathbf{E} in the heterogeneous case (i.e. when $\underline{\varepsilon}$ and $\underline{\mu}$ depend on \mathbf{x}).

2. In the isotropic case ($\underline{\varepsilon} = \varepsilon I_2$, $\underline{\mu} = \mu I_2$, I_2 identity of \mathbb{R}^2), (1.18) and (1.21) can be written as the two following second-order wave equations⁶:

$$\mu(\mathbf{x}) \frac{\partial^2 H}{\partial t^2}(\mathbf{x}, t) - \nabla \cdot \left(\frac{1}{\varepsilon(\mathbf{x})} \nabla H(\mathbf{x}, t) \right) = J'(\mathbf{x}, t). \quad (1.22)$$

$$\varepsilon(\mathbf{x}) \frac{\partial^2 E}{\partial t^2}(\mathbf{x}, t) - \nabla \cdot \left(\frac{1}{\mu(\mathbf{x})} \nabla E(\mathbf{x}, t) \right) = -j(\mathbf{x}, t), \quad (1.23)$$

3. Some authors call “transverse-magnetic” the formulation given in (1.19a) and (1.19b) and “transverse-electric” the formulation given in (1.16a) and (1.16b).

1.3 The Elastics System

1.3.1 General Formulation

Let $\mathbf{v} \in \mathbb{R}^d$ denote the displacement vector and $\underline{\underline{\tau}}$ the stress tensor for the elastic medium. Then the general formulation of the elastics system in a non-homogeneous, anisotropic medium is:

$$\rho(\mathbf{x}) \frac{\partial^2 \mathbf{v}}{\partial t^2}(\mathbf{x}, t) - \mathbf{div} \underline{\underline{\tau}}(\mathbf{x}, t) = \mathbf{f}(\mathbf{x}, t), \quad (1.24a)$$

$$\underline{\underline{\tau}}(\mathbf{x}, t) = \underline{\underline{C}}(\mathbf{x}) \underline{\underline{\varepsilon}}(\mathbf{v})(\mathbf{x}, t) \quad (\text{Hooke's law}). \quad (1.24b)$$

Now, if $(i, j, k, \ell) \in \{1, \dots, d\}^4$, we have

$$\underline{\underline{\tau}} = (\boldsymbol{\tau}_1, \dots, \boldsymbol{\tau}_d), \quad (1.25a)$$

$$\boldsymbol{\tau}_i = (\tau_{i1}, \dots, \tau_{id})^T, \quad (1.25b)$$

$$\mathbf{div} \underline{\underline{\tau}} = (\nabla \cdot \boldsymbol{\tau}_1, \dots, \nabla \cdot \boldsymbol{\tau}_d)^T, \quad (1.25c)$$

$$\varepsilon_{ij} \mathbf{v} = \frac{1}{2} \left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right) \quad (\text{strain tensor}), \quad (1.25d)$$

$$(\underline{\underline{C}} \underline{\underline{\varepsilon}})_{ij} = \sum_{k=1}^d \sum_{\ell=1}^d C_{ijkl} \varepsilon_{k\ell}. \quad (1.25e)$$

Moreover, $\underline{\underline{C}}$ is a cyclic symmetric tensor, i.e. $C_{ijkl} = C_{k\ell ij} = C_{jik\ell}$. For this reason, the number of independent coefficients of $\underline{\underline{C}}$ is equal to 6 for $d = 2$ and 21 for $d = 3$. This symmetry of $\underline{\underline{C}}$ implies the symmetry of the stress tensor $\underline{\underline{\tau}}$.

Here, the initial conditions can be written as

$$\mathbf{v}(\mathbf{x}, 0) = \mathbf{v}_0(\mathbf{x}), \quad \frac{\partial \mathbf{v}}{\partial t}(\mathbf{x}, 0) = \mathbf{v}_1(\mathbf{x}). \quad (1.26)$$

⁶ Similar but very complicated forms can also be obtained for the anisotropic case.

1.3.2 The Isotropic Case

When one deals with an isotropic medium, $\underline{\underline{C}}$ can be written as

$$C_{ijkl} = \lambda \delta_{ij} \delta_{kl} + \mu (\delta_{ik} \delta_{jl} + \delta_{il} \delta_{jk}), \quad (1.27)$$

where δ_{jl} is the Kronecker symbol.

So, Hooke's law assumes the well-known form

$$\tau_{ij} = \lambda \delta_{ij} \sum_{k=1}^d \varepsilon_{kk} + 2\mu \varepsilon_{ij}, \quad (1.28)$$

where λ and μ are Lamé's coefficients.

In this case, the system (1.24a) and (1.24b) can be written as

$$\rho(\mathbf{x}) \frac{\partial^2 \mathbf{v}}{\partial t^2}(\mathbf{x}, t) - \mathbf{div} \underline{\underline{\tau}}(\mathbf{x}, t) = \mathbf{f}(\mathbf{x}, t), \quad (1.29a)$$

$$\underline{\underline{\tau}}_D(\mathbf{x}, t) = \mathcal{A}(\mathbf{x}) \mathbf{v}, \quad (1.29b)$$

$$\underline{\underline{\tau}}_{ND}(\mathbf{x}, t) = \mathcal{B}(\mathbf{x}) \mathbf{v}, \quad (1.29c)$$

where

$$\underline{\underline{\tau}}_D = (\tau_{ii})_{1 \leq i \leq d}, \quad \underline{\underline{\tau}}_{ND} = (\tau_{ij})_{1 \leq i < j \leq d},$$

and \mathcal{A} and \mathcal{B} are the following matrix differential operators.

In 3D.

$$\mathcal{A} = \begin{pmatrix} (\lambda + 2\mu) \frac{\partial}{\partial x_1} & \lambda \frac{\partial}{\partial x_2} & \lambda \frac{\partial}{\partial x_3} \\ \lambda \frac{\partial}{\partial x_1} & (\lambda + 2\mu) \frac{\partial}{\partial x_2} & \lambda \frac{\partial}{\partial x_3} \\ \lambda \frac{\partial}{\partial x_1} & \lambda \frac{\partial}{\partial x_2} & (\lambda + 2\mu) \frac{\partial}{\partial x_3} \end{pmatrix}, \quad (1.30)$$

$$\mathcal{B} = \begin{pmatrix} \mu \frac{\partial}{\partial x_2} & \mu \frac{\partial}{\partial x_1} & 0 \\ \mu \frac{\partial}{\partial x_3} & 0 & \mu \frac{\partial}{\partial x_1} \\ 0 & \mu \frac{\partial}{\partial x_3} & \mu \frac{\partial}{\partial x_2} \end{pmatrix}. \quad (1.31)$$

In 2D.

$$\mathcal{A} = \begin{pmatrix} (\lambda + 2\mu) \frac{\partial}{\partial x_1} & \lambda \frac{\partial}{\partial x_2} \\ \lambda \frac{\partial}{\partial x_1} & (\lambda + 2\mu) \frac{\partial}{\partial x_2} \end{pmatrix}, \quad (1.32)$$

$$\mathcal{B} = \begin{pmatrix} \mu \frac{\partial}{\partial x_2} & \mu \frac{\partial}{\partial x_1} \end{pmatrix}. \quad (1.33)$$

This leads to the following expressions of (1.24a) and (1.24b) in terms of \mathbf{v} only:

In 3D.

$$\begin{aligned} \rho \frac{\partial^2 v_1}{\partial t^2} - \frac{\partial}{\partial x_1} \left((\lambda + 2\mu) \frac{\partial v_1}{\partial x_1} + \lambda \frac{\partial v_2}{\partial x_2} + \lambda \frac{\partial v_3}{\partial x_3} \right) \\ - \frac{\partial}{\partial x_2} \left(\mu \left(\frac{\partial v_1}{\partial x_2} + \frac{\partial v_2}{\partial x_1} \right) \right) \end{aligned} \quad (1.34a)$$

$$- \frac{\partial}{\partial x_3} \left(\mu \left(\frac{\partial v_1}{\partial x_3} + \frac{\partial v_3}{\partial x_1} \right) \right) = f_1,$$

$$\begin{aligned} \rho \frac{\partial^2 v_2}{\partial t^2} - \frac{\partial}{\partial x_1} \left(\mu \left(\frac{\partial v_1}{\partial x_2} + \frac{\partial v_2}{\partial x_1} \right) \right) \\ - \frac{\partial}{\partial x_2} \left(\lambda \frac{\partial v_1}{\partial x_1} + (\lambda + 2\mu) \frac{\partial v_2}{\partial x_2} + \lambda \frac{\partial v_3}{\partial x_3} \right) \end{aligned} \quad (1.34b)$$

$$- \frac{\partial}{\partial x_3} \left(\mu \left(\frac{\partial v_2}{\partial x_3} + \frac{\partial v_3}{\partial x_2} \right) \right) = f_2,$$

$$\begin{aligned} \rho \frac{\partial^2 v_3}{\partial t^2} - \frac{\partial}{\partial x_1} \left(\mu \left(\frac{\partial v_1}{\partial x_3} + \frac{\partial v_3}{\partial x_1} \right) \right) \\ - \frac{\partial}{\partial x_2} \left(\mu \left(\frac{\partial v_3}{\partial x_2} + \frac{\partial v_2}{\partial x_3} \right) \right) \end{aligned} \quad (1.34c)$$

$$+ \frac{\partial}{\partial x_3} \left(\lambda \frac{\partial v_1}{\partial x_1} + \lambda \frac{\partial v_2}{\partial x_2} + (\lambda + 2\mu) \frac{\partial v_3}{\partial x_3} \right) = f_3.$$

In 2D.

$$\begin{aligned} \rho \frac{\partial^2 v_1}{\partial t^2} - \frac{\partial}{\partial x_1} \left((\lambda + 2\mu) \frac{\partial v_1}{\partial x_1} + \lambda \frac{\partial v_2}{\partial x_2} \right) \\ - \frac{\partial}{\partial x_2} \left(\mu \left(\frac{\partial v_1}{\partial x_2} + \frac{\partial v_2}{\partial x_1} \right) \right) = f_1, \end{aligned} \quad (1.35a)$$

$$\begin{aligned} \rho \frac{\partial^2 v_2}{\partial t^2} - \frac{\partial}{\partial x_1} \left(\mu \left(\frac{\partial v_1}{\partial x_2} + \frac{\partial v_2}{\partial x_1} \right) \right) \\ - \frac{\partial}{\partial x_2} \left(\lambda \frac{\partial v_1}{\partial x_1} + (\lambda + 2\mu) \frac{\partial v_2}{\partial x_2} \right) = f_2. \end{aligned} \quad (1.35b)$$

In the homogeneous case, (1.34a)–(1.34c) can be written in the following compact form:

$$\rho \frac{\partial^2 \mathbf{v}}{\partial t^2} = \mu \Delta \mathbf{v} + (\lambda + \mu) \nabla(\nabla \cdot \mathbf{v}), \quad (1.36)$$

where $\Delta \mathbf{v} = (\Delta v_j)_{j=1..3}$.

Now, let us consider a decomposition of the displacement vector \mathbf{v} of the form

$$\mathbf{v} = \nabla \varphi + \nabla \times \boldsymbol{\psi}. \quad (1.37)$$

We obtain, after inserting this decomposition into (1.36):

$$\rho \frac{\partial^2}{\partial t^2} (\nabla \varphi + \nabla \times \boldsymbol{\psi}) = \mu \Delta (\nabla \varphi + \nabla \times \boldsymbol{\psi}) + (\lambda + \mu) \nabla(\nabla \cdot (\nabla \varphi + \nabla \times \boldsymbol{\psi})). \quad (1.38)$$

Since $\nabla \cdot \nabla \varphi = \Delta \varphi$ and $\nabla \cdot (\nabla \times \boldsymbol{\psi}) = 0$, we obtain, after rearranging the terms of (1.38):

$$\nabla \left[\rho \frac{\partial^2 \varphi}{\partial t^2} - (\lambda + 2\mu) \Delta \varphi \right] + \nabla \times \left[\rho \frac{\partial^2 \boldsymbol{\psi}}{\partial t^2} - \mu \Delta \boldsymbol{\psi} \right] = 0. \quad (1.39)$$

So, φ and $\boldsymbol{\psi}$ each satisfy one wave equation with a different velocity. These potentials actually correspond to two physical waves:

- The P -wave (or *pressure*⁷ wave) whose velocity is

$$V_P = \sqrt{\frac{\lambda + 2\mu}{\rho}}. \quad (1.40)$$

⁷ Also called *primary* or *longitudinal* wave.

- The S -wave or (*shear*⁸ wave) whose velocity is

$$V_S = \sqrt{\frac{\mu}{\rho}}. \quad (1.41)$$

Obviously, we have:

$$V_P^2 \geq 2V_S^2. \quad (1.42)$$

From the physical point of view, the P -wave corresponds to the propagation of a displacement parallel to the direction of propagation and the S -wave to the propagation of a distortion (described by the curl operator) in a plane orthogonal to the direction of propagation.

1.4 Boundary Conditions

The wave equations given in this chapter were all written in \mathbb{R}^d . However, in practice, the models are defined in bounded domains at the boundary of which some physical conditions must be written. These boundary conditions can be of different sorts but we provide here the most classical ones. The modeling of unbounded domains by boundary conditions will be treated in a separate chapter at the end of this book.

1.4.1 The Wave Equation

For the wave equation, the most classical conditions are

- *The Dirichlet condition:*

$$u = g(\mathbf{x}, t), \quad (1.43)$$

which models, when $g = 0$, a soft boundary which can be, for instance, the surface of contact of a liquid with the air.

- *The Neumann condition:*

$$\frac{\partial u}{\partial n} = g(\mathbf{x}, t), \quad (1.44)$$

where \mathbf{n} is the unit outward normal to the boundary, which models, when $g = 0$ a rigid boundary which can be, for instance, the wall of a container.

For both conditions, the function g can be a source located on the boundary. When $g = 0$, both conditions provide perfectly reflecting boundaries.

More seldom, one can use impedance boundary conditions:

⁸ Also called *secondary* or *transverse* wave.

$$\frac{\partial u}{\partial t} + \alpha \frac{\partial u}{\partial n} = 0, \quad (1.45)$$

which model semi-reflecting boundaries. This kind of condition appears, in particular, in the treatment of unbounded domains by absorbing boundary conditions (cf. Chap. 14).

1.4.2 The Maxwell Equations

The simplest and most classical boundary condition for the Maxwell equation is the perfectly conducting boundary condition:

$$\mathbf{E} \times \mathbf{n} = 0, \quad (1.46)$$

where \mathbf{n} is the outward normal unit, which is also a reflecting boundary condition. Equation (1.46) means that the tangential component of \mathbf{E} is equal to 0.

In isotropic media, this condition can also take the form:

$$\mathbf{H} \cdot \mathbf{n} = 0. \quad (1.47)$$

For these equations, the impedance-like conditions are more frequent. In particular, the Silver-Müller condition [18]:

$$\sqrt{\varepsilon\mu} \left(\frac{\partial \mathbf{E}}{\partial t} \times \mathbf{n} \right) \times \mathbf{n} - (\nabla \times \mathbf{E}) \times \mathbf{n} = 0 \quad (1.48)$$

is a basic condition of physics which models a partially absorbing boundary in the time domain.

1.4.3 The Elastics System

From the elastics system, two kinds of boundary conditions are classically used [3]:

- *Displacement boundary conditions:*

$$v_j = g_j(\mathbf{x}, t), \quad j = 1..d, \quad (1.49)$$

for which the components of the displacement are prescribed on the boundary.

- *Traction boundary conditions:*

$$\tau_{jk} n_j = g_k(\mathbf{x}, t), \quad j = 1..d, k = 1..d, \quad (1.50)$$

which provide, when $g_k = 0 \forall k = 1..d$, a free surface condition which models the interface of a solid with the vacuum (or the air). In particular,

this is the condition which models the surface of the earth in geophysics. This condition generates a surface wave called the *Rayleigh wave* whose velocity V_R is given by the following equation:

$$\left(2 - \frac{V_R^2}{V_S^2}\right)^2 - 4 \left(1 - \frac{V_R^2}{V_P^2}\right)^{\frac{1}{2}} \left(1 - \frac{V_R^2}{V_S^2}\right)^{\frac{1}{2}} = 0, \quad (1.51)$$

where V_P and V_S are defined as in (1.40) and (1.41). One can show that we have $0 < V_R < V_S$.

A detailed study of this wave can be found in [3].

From the theoretical point of view, one can find a general proof of existence of this class of equations in, among others, [44, 79, 81] and a broad discussion of the homogeneous case in [113]. For the physical point of view, one can consult [3, 54, 56, 77], for instance, and [9] for a general description of elastic waves in anisotropic media.

2. Functional Issues

Of course, the purpose of this chapter is not to provide an exhaustive theory of functional analysis. We are only going to give the basic notions which will enable us to better understand the finite element approximations. Although these approximations can be introduced in an intuitive way, their definition based on functional spaces enables us to obtain a wider and more rigorous view of these methods.

2.1 Some Functional Spaces

The definition of functional spaces is based on the concept of distributions, which extends the notion of derivative to a much larger class of functions than by the classical differential calculus. An important application of this theory for us is the computation of the derivative of any continuous function, even not derivable. The presentation, even when summarized, of this theory is beyond of the framework of this book and the reader who has no knowledge of it could consult [43, 100, 101, 113] for instance. So, in the following, all the derivatives that we shall use will be taken in the distributions sense.

2.1.1 Sobolev Spaces

The first and basic functional spaces used in the finite element theory are the Sobolev spaces H^m . Let us give their definitions for an open set Ω of \mathbb{R}^d (in our case, we shall take $1 \leq d \leq 3$) whose boundary is $\partial\Omega$. As a first step, we define the general partial differential operator for a scalar function u of \mathbb{R}^d ($\boldsymbol{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$)

$$D^\alpha = \frac{\partial^p}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}, \quad (2.1)$$

where $p \in \mathbb{N}^*$ and

$$\alpha = \left\{ (\alpha_1, \dots, \alpha_d) \in \mathbb{N}^d \text{ such that } |\alpha| = \sum_{j=1}^d \alpha_j = p \right\}.$$

Then, the Sobolev space H^m is defined as follows

$$H^m(\Omega) = \left\{ u \in L^2(\Omega) \text{ such that } \forall \alpha \text{ such that } |\alpha| \leq m, D^\alpha u \in L^2(\Omega) \right\}, \quad (2.2)$$

where, of course, $L^2(\Omega)$ is the space of functions whose square is integrable over Ω .

In our case, the most useful Sobolev space is

$$H^1(\Omega) = \left\{ u \in L^2(\Omega) \text{ such that } \forall j = 1..d, \frac{\partial u}{\partial x_j} \in L^2(\Omega) \right\}. \quad (2.3)$$

One can define a value (trace) of a function of $H^1(\Omega)$ on $\partial\Omega$ but not a value of its derivative. The trace is in a subspace of $L^2(\partial\Omega)$. This property enables us to define the following subspace of $H^1(\Omega)$:

$$H_0^1(\Omega) = \left\{ u \in H^1(\Omega) \text{ such that } u = 0 \text{ on } \partial\Omega \right\}. \quad (2.4)$$

This subspace is the appropriate framework for solving problems with homogeneous Dirichlet boundary conditions.

The main property of $H^1(\Omega)$ for us is contained in the theorem below that we give without proof:

Theorem 1. *Let Ω_1 and Ω_2 be two subsets of Ω such that $\bar{\Omega} = \bar{\Omega}_1 \cup \bar{\Omega}_2$, $\Omega_1 \cap \Omega_2 = \emptyset$ and $\bar{\Omega}_1 \cap \bar{\Omega}_2 = \Gamma$ and u a function such that $u_1 = u|_{\Omega_1} \in H^1(\Omega_1)$, $u_2 = u|_{\Omega_2} \in H^1(\Omega_2)$. Then $u \in H^1(\Omega)$ if and only if $u_1 = u_2$ on Γ .*

An immediate consequence of this theorem is:

Corollary 1. *Let $\{\bar{\Omega}_j\}_{j=1}^N$ be a partition of $\bar{\Omega}$ and u a real function defined on Ω . If $\forall j = 1..N$, $u_j = u|_{\Omega_j} \in H^1(\Omega_j)$ and $\forall \ell = 1..N, \forall m = 1..N$ such that $\bar{\Omega}_\ell \cap \bar{\Omega}_m = \Gamma_{\ell,m}$, $u_\ell = u_m$ on $\Gamma_{\ell,m}$, then $u \in H^1(\Omega)$.*

This corollary, which is a more general form of Theorem 1, has an important feature in terms of finite element approximation. It justifies the approximation of $H^1(\Omega_j)$ by Lagrangian finite elements. These elements are such that, on a mesh $\mathcal{M} = \bigcup K_j$, the restriction of an approximated function to K_j is a polynomial¹ and only the continuity is required at the interfaces of the mesh (cf. Chaps. 11 and 12).

Remarks

1. The equality $u_1 = u_2$ holds almost everywhere on Γ . This implies that for $d = 1$, $H^1(\Omega) \subset C^0(\Omega)$ but, for $d > 1$, we have no relation between

¹ Or derived from a polynomial.

$H^1(\Omega)$ and $C^0(\Omega)$ since the solution can be discontinuous at some points when $d = 2$ and on some curves when $d = 3$. This shows that the notion of derivability is applied here to a much larger class of functions than $C^1(\Omega)$ ².

2. The spaces $H^m(\Omega)$ are actually a small part of the general Sobolev spaces which can be defined by using L^p spaces and for $m \in \mathbb{R}$ but their definition is far beyond the needs of this book. A compact presentation of these spaces with more references can be found in [43].

2.1.2 Spaces $H(\mathbf{curl}, \Omega)$ and $H(\mathbf{div}, \Omega)$

The Sobolev spaces described in the previous section required that all the derivatives of the functions should be in $L^2(\Omega)$. For some equations, such as the Maxwell ones, it is useful to define functional spaces with less requirements.

The basic space for the Maxwell equations is, of course, the space in which one can define the curl of a vectorial function. In 3D, this space can be written as

$$H(\mathbf{curl}, \Omega) = \{\mathbf{u} \in [L^2(\Omega)]^3 \text{ such that } \nabla \times \mathbf{u} \in [L^2(\Omega)]^3\}. \quad (2.5)$$

As for $H^1(\Omega)$, one can define a tangential trace of a function of $H(\mathbf{curl}, \Omega)$ on $\partial\Omega$. However, this trace is not in $L^2(\partial\Omega)$ but in a subspace of the space of distributions. This property enables us to define the following subspace of $H(\mathbf{curl}, \Omega)$ (\mathbf{n} denotes the unit outward normal to the boundary):

$$H_0(\mathbf{curl}, \Omega) = \{\mathbf{u} \in H(\mathbf{curl}, \Omega) \text{ such that } \mathbf{u} \times \mathbf{n} = 0\}, \quad (2.6)$$

where \mathbf{n} is the unit outward normal to $\partial\Omega$, which corresponds to a medium with a perfectly conducting boundary condition. The equality $\mathbf{u} \times \mathbf{n} = 0$ is given in the distributions sense.

Of course, the electric and magnetic fields \mathbf{E} and \mathbf{H} belong to $H(\mathbf{curl}, \Omega)$.

Another useful space for the Maxwell equations is the following space in which are the electric and magnetic inductions \mathbf{D} and \mathbf{B} :

$$H(\mathbf{div}, \Omega) = \{\mathbf{u} \in [L^2(\Omega)]^3 \text{ such that } \nabla \cdot \mathbf{u} \in L^2(\Omega)\}. \quad (2.7)$$

One can define a normal trace of a function of $H(\mathbf{div}, \Omega)$ which is also in a subspace of the space of distributions.

² We recall that $C^0(\Omega)$ is the space of continuous functions on Ω and $C^n(\Omega)$, the space of functions whose derivatives are continuous to the n th-order.

In 2D, $H(\mathbf{curl}, \Omega)$ can be defined in two ways which depend on the character of the function. For a vectorial function, we have

$$H(\mathbf{curl}, \Omega) = \{\mathbf{u} \in [L^2(\Omega)]^2 \text{ such that } \mathbf{curl}\mathbf{u} \in L^2(\Omega)\}, \quad (2.8)$$

where \mathbf{curl} is defined as in Sect. 1.2.2.

For a scalar function, the space is based on the definition of \mathbf{curl} in Sect. 1.2.2:

$$H(\mathbf{curl}, \Omega) = \{u \in L^2(\Omega) \text{ such that } \mathbf{curl}u \in [L^2(\Omega)]^2\}. \quad (2.9)$$

The definition of \mathbf{curl} shows that, in this case, we actually have $H(\mathbf{curl}, \Omega) = H^1(\Omega)$.

As for $H^1(\Omega)$, we have the following important property of the functions of all the above spaces:

Theorem 2. Let Ω_1 and Ω_2 be two subsets of Ω such that $\bar{\Omega} = \bar{\Omega}_1 \cup \bar{\Omega}_2$, $\Omega_1 \cap \Omega_2 = \emptyset$ and $\bar{\Omega}_1 \cap \bar{\Omega}_2 = \Gamma$ and \mathbf{u} a function such that $\mathbf{u}_1 = \mathbf{u}|_{\Omega_1} \in H(\mathbf{curl}, \Omega_1)$ (respectively $\in H(\mathbf{div}, \Omega_1)$), $\mathbf{u}_2 = \mathbf{u}|_{\Omega_2} \in H(\mathbf{curl}, \Omega_2)$ (respectively $\in H(\mathbf{div}, \Omega_2)$). Then $\mathbf{u} \in H(\mathbf{curl}, \Omega)$ (respectively $\in H(\mathbf{div}, \Omega)$) if and only if $\mathbf{u}_1 \times \mathbf{n} = \mathbf{u}_2 \times \mathbf{n}$ (respectively $\mathbf{u}_1 \cdot \mathbf{n} = \mathbf{u}_2 \cdot \mathbf{n}$) in the sense of distributions on Γ , where \mathbf{n} denotes a unit normal to Γ .

On this theorem will be based the approximation of $H(\mathbf{curl}, \Omega)$ by the so-called *edge elements* which are continuous in each K_j of a mesh \mathcal{M} and whose tangential components only are continuous through the interfaces. In the same way, $H(\mathbf{div}, \Omega)$ will be approximated by elements with continuous normal components through the interfaces (cf. Chap. 13).

2.2 Variational Formulations

2.2.1 The Acoustics Equation

On the basis of the functional spaces defined in the previous section, one can define *variational formulations* of the wave equations which are the first step of their finite element approximations. Let us first define it for the acoustics equation.

If we multiply the following acoustics equation:

$$\eta \frac{\partial^2 u}{\partial t^2} - \nabla \cdot (\gamma \nabla u) = f \quad (2.10)$$

by a function $v \in H^1(\Omega)$ and we integrate by parts the *stiffness integral*, i.e. the integral corresponding to the *stiffness term* $\nabla \cdot (\gamma \nabla u)$, we obtain:

$$\frac{d^2}{dt^2} \int_{\Omega} \eta uv \, d\mathbf{x} + \int_{\Omega} \gamma \nabla u \cdot \nabla v \, d\mathbf{x} - \int_{\partial\Omega} \gamma v \frac{\partial u}{\partial n} \, d\Gamma = \int_{\Omega} fv \, d\mathbf{x}. \quad (2.11)$$

Let us now suppose that we have a Neumann condition on the boundary $\partial\Omega$, i.e. $\partial u/\partial n = g$. We derive from (2.10) the following *variational problem*:

Find u such that $u(., t) \in H^1(\Omega)$ and

$$\begin{aligned} & \frac{d^2}{dt^2} \int_{\Omega} \eta uv \, d\mathbf{x} + \int_{\Omega} \gamma \nabla u \cdot \nabla v \, d\mathbf{x} \\ &= \int_{\partial\Omega} \gamma gv \, d\Gamma + \int_{\Omega} fv \, d\mathbf{x}, \quad \forall v \in H^1(\Omega). \end{aligned} \quad (2.12)$$

For the homogeneous Dirichlet condition $u = 0$, the variational problem is

Find u such that $u(., t) \in H_0^1(\Omega)$ and

$$\frac{d^2}{dt^2} \int_{\Omega} \eta uv \, d\mathbf{x} + \int_{\Omega} \gamma \nabla u \cdot \nabla v \, d\mathbf{x} = \int_{\Omega} fv \, d\mathbf{x}, \quad \forall v \in H_0^1(\Omega) \quad (2.13)$$

and, for the impedance condition given in (1.45), we obtain:

Find u such that $u(., t) \in H^1(\Omega)$ and

$$\left\{ \begin{array}{l} \frac{d^2}{dt^2} \int_{\Omega} \eta uv \, d\mathbf{x} + \int_{\Omega} \gamma \nabla u \cdot \nabla v \, d\mathbf{x} = - \int_{\partial\Omega} \frac{1}{\alpha} \gamma v \frac{\partial u}{\partial t} \, d\Gamma + \int_{\Omega} fv \, d\mathbf{x}, \\ \forall v \in H^1(\Omega). \end{array} \right. \quad (2.14)$$

Of course, one must add the initial conditions defined in (1.4) to (2.12)–(2.14).

By using density properties of functional spaces, one shows that (2.12)–(2.14) are equivalent to the acoustics equation (2.10) with the different boundary conditions when the solution is sought as a distribution. These formulations are also called *weak formulations* of the acoustics equation because their solution is sought in a space whose functions are required to be only once derivable whereas the stiffness term uses second derivatives.

Remarks

1. The use of g and of the time derivative of u in the boundary integrals in (2.12) and (2.14) avoids the presence of the normal derivative of u which cannot be defined for a function of $H^1(\Omega)$.
2. The functional frame for non-homogeneous Dirichlet conditions is more difficult to define and is widely treated in [81].

2.2.2 The Maxwell Equations

Let us consider the following 3D Maxwell equations:

$$\underline{\varepsilon} \frac{\partial \mathbf{E}}{\partial t} - \nabla \times \mathbf{H} = -\mathbf{J}, \quad (2.15a)$$

$$\underline{\mu} \frac{\partial \mathbf{H}}{\partial t} + \nabla \times \mathbf{E} = 0. \quad (2.15b)$$

We first multiply the first equation by $\varphi \in H(\mathbf{curl}, \Omega)$ and the second one by $\psi \in [L^2(\Omega)]^3$. Then, after integrating by parts the stiffness integral of (2.15a) (which corresponds to $\nabla \times \mathbf{H}$), we obtain:

$$\begin{aligned} \frac{d}{dt} \int_{\Omega} \underline{\varepsilon} \mathbf{E} \cdot \varphi d\mathbf{x} - \int_{\Omega} \mathbf{H} \cdot (\nabla \times \varphi) d\mathbf{x} = \\ \int_{\partial\Omega} [\mathbf{n} \times (\mathbf{H} \times \mathbf{n})] \cdot (\varphi \times \mathbf{n}) d\Gamma - \int_{\Omega} \mathbf{J} \varphi d\mathbf{x}, \end{aligned} \quad (2.16a)$$

$$\frac{d}{dt} \int_{\Omega} \underline{\mu} \mathbf{H} \cdot \psi d\mathbf{x} + \int_{\Omega} (\nabla \times \mathbf{E}) \cdot \psi d\mathbf{x} = 0. \quad (2.16b)$$

Since $[\mathbf{n} \times (\mathbf{H} \times \mathbf{n})] \cdot (\varphi \times \mathbf{n}) = \mathbf{H} \cdot (\varphi \times \mathbf{n})$, for the perfectly conducting boundary condition (1.46), we obtain the following variational problem:

Find \mathbf{E} and \mathbf{H} such that $\mathbf{E}(., t) \in H_0(\mathbf{curl}, \Omega)$ and $\mathbf{H}(., t) \in [L^2(\Omega)]^3$ and

$$\begin{aligned} \frac{d}{dt} \int_{\Omega} \underline{\varepsilon} \mathbf{E} \cdot \varphi d\mathbf{x} - \int_{\Omega} \mathbf{H} \cdot (\nabla \times \varphi) d\mathbf{x} = - \int_{\Omega} \mathbf{J} \varphi d\mathbf{x}, \\ \forall \varphi \in H_0(\mathbf{curl}, \Omega), \end{aligned} \quad (2.17a)$$

$$\frac{d}{dt} \int_{\Omega} \underline{\mu} \mathbf{H} \cdot \psi d\mathbf{x} + \int_{\Omega} (\nabla \times \mathbf{E}) \cdot \psi d\mathbf{x} = 0, \quad \forall \psi \in [L^2(\Omega)]^3. \quad (2.17b)$$

Of course, one must add to (2.17a) and (2.17b), the initial conditions defined in (1.8).

Here also, the magnetic field \mathbf{H} is sought in $[L^2(\Omega)]^3$ which is larger than $H(\mathbf{curl}, \Omega)$ to which \mathbf{H} belongs.

Another variational formulation can be derived from the second-order formulation of the Maxwell equation given in (1.9) for instance. If we multiply this equation by $\varphi \in H(\mathbf{curl}, \Omega)$ and we integrate by parts the stiffness term, we obtain similarly:

$$\begin{aligned} \frac{d^2}{dt^2} \int_{\Omega} \underline{\underline{\varepsilon}} \cdot \mathbf{E} \cdot \boldsymbol{\varphi} dx + \int_{\Omega} \underline{\underline{\mu}}^{-1} (\nabla \times \mathbf{E}) \cdot (\nabla \times \boldsymbol{\varphi}) dx = \\ - \int_{\partial\Omega} [\mathbf{n} \times (\underline{\underline{\mu}}^{-1} (\nabla \times \mathbf{E}) \times \mathbf{n})] \cdot (\boldsymbol{\varphi} \times \mathbf{n}) d\Gamma - \int_{\Omega} \mathbf{j} \cdot \boldsymbol{\varphi} dx, \end{aligned} \quad (2.18)$$

which provides, for the perfectly conducting boundary condition:

Find \mathbf{E} such that $\mathbf{E}(., t) \in H_0(\mathbf{curl}, \Omega)$ and

$$\left\{ \begin{array}{l} \frac{d^2}{dt^2} \int_{\Omega} \underline{\underline{\varepsilon}} \cdot \mathbf{E} \cdot \boldsymbol{\varphi} dx + \int_{\Omega} \underline{\underline{\mu}}^{-1} (\nabla \times \mathbf{E}) \cdot (\nabla \times \boldsymbol{\varphi}) dx = - \int_{\Omega} \mathbf{j} \cdot \boldsymbol{\varphi} dx, \\ \forall \boldsymbol{\varphi} \in H_0(\mathbf{curl}, \Omega). \end{array} \right. \quad (2.19)$$

In this form, one can take into account the Silver-Müller condition defined in (1.48), in the isotropic case, as follows:

Find \mathbf{E} such that $\mathbf{E}(., t) \in H(\mathbf{curl}, \Omega)$ and

$$\left\{ \begin{array}{l} \frac{d^2}{dt^2} \int_{\Omega} \varepsilon \mathbf{E} \cdot \boldsymbol{\varphi} dx + \int_{\Omega} \mu^{-1} (\nabla \times \mathbf{E}) \cdot (\nabla \times \boldsymbol{\varphi}) dx \\ + \frac{d}{dt} \int_{\partial\Omega} \sqrt{\frac{\varepsilon}{\mu}} (\mathbf{E} \times \mathbf{n}) \cdot (\boldsymbol{\varphi} \times \mathbf{n}) d\Gamma = - \int_{\Omega} \mathbf{j} \cdot \boldsymbol{\varphi} dx, \\ \forall \boldsymbol{\varphi} \in H(\mathbf{curl}, \Omega). \end{array} \right. \quad (2.20)$$

Remark

One can also take $\mathbf{E}(., t) \in [L^2(\Omega)]^3$ and $\mathbf{H}(., t) \in H(\mathbf{curl}, \Omega)$ in order to obtain a variational formulation of (2.15a) and (2.15b). In this case, the integration by parts is made on (2.15b) and the boundary integral is

$$\int_{\partial\Omega} [\mathbf{n} \times (\mathbf{E} \times \mathbf{n})] \cdot (\boldsymbol{\psi} \times \mathbf{n}) d\Gamma, \quad \boldsymbol{\psi} \in H(\mathbf{curl}, \Omega).$$

This boundary integral vanishes when $\mathbf{E} \times \mathbf{n} = 0$ in the variational formulation and so, \mathbf{H} must be sought in $H(\mathbf{curl}, \Omega)$ instead of $H_0(\mathbf{curl}, \Omega)$. Then, the resulting variational problem for the perfectly conducting boundary condition can be written as

Find \mathbf{E} and \mathbf{H} such that $\mathbf{E}(., t) \in [L^2(\Omega)]^3$ and $\mathbf{H}(., t) \in H(\mathbf{curl}, \Omega)$ and

$$\begin{aligned} \frac{d}{dt} \int_{\Omega} \underline{\underline{\varepsilon}} \cdot \mathbf{E} \cdot \boldsymbol{\varphi} dx - \int_{\Omega} (\nabla \times \mathbf{H}) \cdot \boldsymbol{\varphi} dx = - \int_{\Omega} \mathbf{J} \cdot \boldsymbol{\varphi} dx, \\ \forall \boldsymbol{\varphi} \in [L^2(\Omega)]^3, \end{aligned} \quad (2.21a)$$

$$\frac{d}{dt} \int_{\Omega} \underline{\underline{\mu}} \cdot \mathbf{H} \cdot \boldsymbol{\psi} dx + \int_{\Omega} \mathbf{E} \cdot (\nabla \times \boldsymbol{\psi}) dx = 0, \quad \forall \boldsymbol{\psi} \in H(\mathbf{curl}, \Omega). \quad (2.21b)$$

Actually, the perfectly conducting boundary condition behaves as a Dirichlet condition versus \mathbf{E} and a Neumann condition versus \mathbf{H} .

A similar variational formulation can be obtained for (1.10).

2.2.3 The Elastics System

In order to construct the variational formulation of the elastics system, let us first introduce the following notations:

If $\underline{\underline{u}} = (\mathbf{u}_1, \dots, \mathbf{u}_d)$, where $\mathbf{u}_j = (u_{j,1}, \dots, u_{j,d})$ is a d -dimensional tensor and $\underline{\underline{v}} = (\mathbf{v}_1, \dots, \mathbf{v}_d)^T$, where $\mathbf{v}_j = (v_{j,1}, \dots, v_{j,d})^T$, we denote:

$$\underline{\underline{u}} : \underline{\underline{v}} = \sum_{j=1}^d \mathbf{u}_j \cdot \mathbf{v}_j. \quad (2.22)$$

With these notations, if φ is a function of $[H^1(\Omega)]^d$, after multiplying the stiffness term of (1.24a) and integrating by parts and by taking into account the symmetric character of $\underline{\underline{\tau}}$, we obtain:

$$\int_{\Omega} \mathbf{div}_{\underline{\underline{\tau}}} \cdot \varphi \, d\mathbf{x} = - \int_{\Omega} \underline{\underline{\tau}} : \underline{\underline{\varepsilon}}(\varphi) \, d\mathbf{x} + \int_{\partial\Omega} \underline{\underline{\tau}} \mathbf{n} \cdot \varphi \, d\Gamma. \quad (2.23)$$

So, by multiplying (1.24a) by φ and taking into account (2.23) and (1.24b), we obtain, for a free surface condition (i.e. $\underline{\underline{\tau}} \mathbf{n} = 0$), the following variational problem:

Find \mathbf{v} such that $\mathbf{v}(., t) \in [H^1(\Omega)]^d$ and

$$\frac{d^2}{dt^2} \int_{\Omega} \rho \mathbf{v} \cdot \varphi \, d\mathbf{x} + \int_{\Omega} \underline{\underline{C}} \underline{\underline{\varepsilon}}(\mathbf{v}) : \underline{\underline{\varepsilon}}(\varphi) \, d\mathbf{x} = \int_{\Omega} \mathbf{f} \cdot \varphi \, d\mathbf{x}, \quad \forall \varphi \in [H^1(\Omega)]^d, \quad (2.24)$$

to which one must add the initial conditions defined in (1.26).

2.3 Energy Identities

From the variational formulations, one can derive *energy identities*. These identities are the basic features of the wave equations which ensure their well-posedness and, more concretely, the stability of the solutions.

2.3.1 The Acoustics Equation

Let us consider the variational formulation of the acoustics equation obtained in (2.13) in which we set $f = 0$. If we set $v = \partial u / \partial t$ (which is legitimate since $\forall t \geq 0$, $(\partial u / \partial t)(\cdot, t) \in H_0^1(\Omega)$), we obtain:

$$\frac{d^2}{dt^2} \int_{\Omega} \eta u \frac{\partial u}{\partial t} dx + \int_{\Omega} \gamma \nabla u \cdot \nabla \frac{\partial u}{\partial t} dx = 0, \quad (2.25)$$

which can be rewritten as

$$\frac{d}{dt} \mathcal{E}(u) = \frac{d}{dt} \left\{ \frac{1}{2} \int_{\Omega} \eta \left| \frac{\partial u}{\partial t} \right|^2 dx + \frac{1}{2} \int_{\Omega} \gamma |\nabla u|^2 dx \right\} = 0. \quad (2.26)$$

Equation (2.26) means that, for a given initial solution u_0 , we have $\forall t \geq 0$, $\mathcal{E}(u) = \mathcal{E}(u_0)$, which is the principle of *energy conservation* for the wave equation. The same identity can also be obtained for the homogeneous Neumann condition ($g = 0$ in (2.12)). Actually, both Dirichlet and Neumann homogeneous conditions, which are perfectly reflecting conditions, ensure that no dissipation of the waves occurs in the domain Ω . This feature is given by the energy conservation principle. Of course, the positivity of \mathcal{E} is fundamental.

2.3.2 The Maxwell Equations

Let us set $\boldsymbol{\varphi} = \mathbf{E}$ and $\boldsymbol{\psi} = \mathbf{H}$ in (2.17a) and (2.17b), where we have set $\mathbf{J} = 0$. We obtain:

$$\frac{d}{dt} \int_{\Omega} \underline{\underline{\varepsilon}} \mathbf{E} \cdot \mathbf{E} dx - \int_{\Omega} \mathbf{H} \cdot (\nabla \times \mathbf{E}) dx = 0, \quad (2.27a)$$

$$\frac{d}{dt} \int_{\Omega} \underline{\underline{\mu}} \mathbf{H} \cdot \mathbf{H} + \int_{\Omega} (\nabla \times \mathbf{E}) \cdot \mathbf{H} dx = 0. \quad (2.27b)$$

Since $\underline{\underline{\varepsilon}}$ and $\underline{\underline{\mu}}$ are both symmetric, definite, positive, there exists two matrices $\tilde{\underline{\underline{\varepsilon}}}$ and $\tilde{\underline{\underline{\mu}}}$ such that $\underline{\underline{\varepsilon}} = \tilde{\underline{\underline{\varepsilon}}}^* \tilde{\underline{\underline{\varepsilon}}}$ and $\underline{\underline{\mu}} = \tilde{\underline{\underline{\mu}}}^* \tilde{\underline{\underline{\mu}}}$ (the * symbol indicates the transposed matrix). So, by combining (2.27a) and (2.27b), we obtain the following energy identity for the Maxwell equations:

$$\frac{d}{dt} \mathcal{E}(\mathbf{E}, \mathbf{H}) = \frac{d}{dt} \left\{ \int_{\Omega} |\tilde{\underline{\underline{\varepsilon}}} \mathbf{E}|^2 dx + \int_{\Omega} |\tilde{\underline{\underline{\mu}}} \mathbf{H}|^2 dx \right\} = 0. \quad (2.28)$$

In the same way, by setting $\boldsymbol{\varphi} = \partial \mathbf{E} / \partial t$ in (2.19) (with $\mathbf{j} = 0$), we obtain, as for the wave equation:

$$\frac{d}{dt} \mathcal{E}(\mathbf{E}) = \frac{d}{dt} \left\{ \frac{1}{2} \int_{\Omega} \left| \tilde{\underline{\underline{\varepsilon}}} \frac{\partial \mathbf{E}}{\partial t} \right|^2 dx + \frac{1}{2} \int_{\Omega} |\tilde{\underline{\underline{\mu}}}^{*-1} \nabla \times \mathbf{E}|^2 dx \right\} = 0. \quad (2.29)$$

In the case of the Silver-Müller condition given in (2.20), we obtain:

$$\begin{aligned}\frac{d}{dt}\mathcal{E}(\mathbf{E}) &= \frac{d}{dt}\left\{\frac{1}{2}\int_{\Omega}\left|\varepsilon^{\frac{1}{2}}\frac{\partial\mathbf{E}}{\partial t}\right|^2 d\mathbf{x} + \frac{1}{2}\int_{\Omega}|\mu^{-\frac{1}{2}}\nabla\times\mathbf{E}|^2 d\mathbf{x}\right\} \\ &= -\int_{\partial\Omega}\left(\frac{\varepsilon}{\mu}\right)^{\frac{1}{2}}\left|\frac{\partial\mathbf{E}}{\partial t}\times\mathbf{n}\right|^2 d\Gamma.\end{aligned}\quad (2.30)$$

Relation (2.30) shows that, contrary to the perfectly conducting boundary condition, the energy is decreasing here. This means that the waves vanish from the domain. In fact, they are absorbed by the boundary condition.

2.3.3 The Elastics System

Although using vector-valued unknowns, the elastics system can be treated as the acoustics equation in order to obtain energy identities. By setting $\varphi = \partial\mathbf{v}/\partial t$ in (2.24), we obtain:

$$\frac{d^2}{dt^2}\int_{\Omega}\rho\mathbf{v}\cdot\frac{\partial\mathbf{v}}{\partial t}d\mathbf{x} + \int_{\Omega}\underline{\underline{C}}\underline{\underline{\varepsilon}}(\mathbf{v}):\underline{\underline{\varepsilon}}\left(\frac{\partial\mathbf{v}}{\partial t}\right)d\mathbf{x} = 0, \forall\varphi\in[H^1(\Omega)]^d. \quad (2.31)$$

By using the linearity of $\underline{\underline{\varepsilon}}$, (2.31) can be rewritten as

$$\frac{d}{dt}\mathcal{E}(\mathbf{v}) = \frac{d}{dt}\left\{\frac{1}{2}\int_{\Omega}\rho\left|\frac{\partial\mathbf{v}}{\partial t}\right|^2 d\mathbf{x} + \frac{1}{2}\int_{\Omega}\underline{\underline{C}}\underline{\underline{\varepsilon}}(\mathbf{v}):\underline{\underline{\varepsilon}}(\mathbf{v})d\mathbf{x}\right\} = 0. \quad (2.32)$$

The positivity of \mathcal{E} is derived from the positivity of $\underline{\underline{C}}$.

Remark

One can easily check that the energy identities can be written for subspaces of the functional spaces. In particular, they hold for finite element subspaces and then provide natural conditions of stability for these approximations.

Part II

Finite Difference Methods

Introduction

This part will deal with the construction and the analysis of finite difference methods. These methods were the first to be used to solve the wave equation and remain very popular in the engineering community. Their main qualities are the simplicity of their construction and implementation and the fact that regular grids are very well-suited to wave propagation. From an educational point of view, we prefer to introduce the basic principles of the analysis of numerical methods for the wave equation via these methods because of their simplicity.

The first finite difference method used for the wave equation was based on centered second-order approximations of the second-order derivative in time and of the Laplace operator. This approximation provides all the good properties of stability and dispersion needed for the wave equation. Unfortunately it does not have enough accuracy to be able to model large domains in which the waves propagate on distances of several tens, even hundreds, of wavelengths. Even though, at that time, computers were not yet capable of solving such problems, geophysicists already had in mind the need for them. The first paper on a fourth-order method for the homogeneous wave equation was published by Alford et al. [4] in 1974. This paper dealt with a 9-point scheme in space for the 2D wave equation but was still second-order in time. The rush to higher-order difference methods began around 1985. From this time, an intensive activity in this subject, both in acoustics and elastics, appeared in the world of geophysics [5, 14, 16, 28, 34, 35, 42, 69, 73, 95, 102, 105, 117, 119]. The main part of the effort was to obtain higher-order in space because most higher-order methods in time were unstable. Only Dablain [42] presented a stable higher-order method in time based on the modified equation approach.

The world of electromagnetism was much less active for two main reasons. First, frequency domain equations seemed more appropriate for radar simulation. Secondly, the complexity of the geometries involved required a finite element approach. However, in the time domain, the second-order Yee scheme [120] remains widely used today and can be considered a reference method. Some other finite difference methods were also developed for the transient Maxwell equations [47, 70, 109]. A wide presentation of finite difference methods in time domain Maxwell equations can be found in [110].

We shall focus our presentation on fourth-order approximations in 2D in homogeneous and heterogeneous media, which contain the main difficulties encountered in higher-order methods versus second-order methods. The basic techniques of analysis and properties of the approximations will also be given, first for the wave equation whose simple character is very convenient. Moreover, all the schemes described below will be centered, since uncentered schemes generate numerical dissipation for wave equations which satisfy a principle of energy conservation.

In the fourth chapter, we shall construct different kinds of centered higher-order approximations in space and time for the homogeneous wave equations. In the fifth chapter, we shall give the basis of plane wave analysis by defining the dispersion relation of the schemes. The sixth and seventh chapters will be devoted to the analysis of stability and accuracy on the basis of the dispersion relation. Finally, the three remaining chapters will provide guidelines to construct schemes in heterogeneous media and to analyze them.

3. Plane Wave Solutions

3.1 A General Solution of the Homogeneous Wave Equation

Let us consider the homogeneous wave equation¹

$$\frac{\partial^2 u}{\partial t^2} - c^2 \Delta u = 0 \quad (3.1)$$

and the direct and inverse *Fourier transforms* in space

$$\mathcal{F}_x u = \hat{u}(\mathbf{k}) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} u(\mathbf{x}) e^{-i\mathbf{k}\cdot\mathbf{x}} d\mathbf{x}, \quad (3.2a)$$

$$\mathcal{F}_x^{-1} \hat{u} = u(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \hat{u}(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{x}} d\mathbf{k}, \quad (3.2b)$$

where $i^2 = -1$, $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{k} \in \mathbb{R}^d$. By applying the Fourier transform in space to (3.1), we obtain the following ODE:

$$\frac{d^2 \hat{u}}{dt^2} + c^2 |\mathbf{k}|^2 \hat{u} = 0, \quad (3.3)$$

whose solution is of the form

$$\hat{u}(\mathbf{k}, t) = A(\mathbf{k}) e^{ic|\mathbf{k}|t} + B(\mathbf{k}) e^{-ic|\mathbf{k}|t}. \quad (3.4)$$

On the other hand, by applying to (3.3) the Fourier transform in time $\mathcal{F}_t : L^2(\mathbb{R}) \rightarrow L^2(\mathbb{R})$ where

$$\mathcal{F}_t u = \hat{u}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} u(t) e^{-i\omega t} dt, \quad (3.5)$$

¹ With a right-hand side, the correct equation would actually be

$$\frac{1}{c^2} \frac{\partial^2 u}{\partial t^2} - \Delta u = f$$

but the formulation given below is equivalent and easier to manipulate when $f = 0$.

we obtain the *dispersion relation* of (3.1):

$$\omega^2 = c^2 |\mathbf{k}|^2. \quad (3.6)$$

Now, by assuming $\omega > 0$, the solution can be written as

$$\hat{u}(\mathbf{k}, t) = A(\mathbf{k}) e^{i\omega t} + B(\mathbf{k}) e^{-i\omega t}. \quad (3.7)$$

The inverse Fourier transform in space applied to \hat{u} provides the following form of the solution of (3.3)

$$u(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} [A(\mathbf{k}) e^{i(\omega t + \mathbf{k} \cdot \mathbf{x})} + B(\mathbf{k}) e^{i(-\omega t + \mathbf{k} \cdot \mathbf{x})}] d\mathbf{k}. \quad (3.8)$$

Equation (3.8) shows that the solution of the homogeneous wave equation can be expressed as a continuous superposition of the plane waves

$$e^{i(\omega t + \mathbf{k} \cdot \mathbf{x})}, \quad (3.9)$$

whose amplitudes are $A(\mathbf{k})$ and $B(\mathbf{k})$. Hence the study of properties of the solutions of the wave equation can be carried out by considering the plane wave solution defined in (3.9).

Remarks

1. ω is the *pulsation* and \mathbf{k} the *wave vector* which indicates the direction of propagation of the plane wave. Obviously, $\omega/|\mathbf{k}|$ is the velocity of the propagated wave.
2. This result given in the case of the scalar wave equation, can be extended to the other equations.

3.2 Application to the Maxwell Equations

3.2.1 The 3D Case

Now, let us look for a plane wave solution of the homogeneous anisotropic Maxwell equations

$$\frac{\partial \mathbf{D}}{\partial t} - \nabla \times \mathbf{H} = 0, \quad (3.10a)$$

$$\frac{\partial \mathbf{B}}{\partial t} + \nabla \times \mathbf{E} = 0, \quad (3.10b)$$

$$\mathbf{D} = \underline{\underline{\epsilon}_0} \mathbf{E}, \quad (3.10c)$$

$$\mathbf{B} = \underline{\underline{\mu}_0} \mathbf{H}, \quad (3.10d)$$

of the form

$$\mathbf{E} = \mathbf{E}_0 e^{i(\omega t + \mathbf{k} \cdot \mathbf{x})}, \quad (3.11a)$$

$$\mathbf{H} = \mathbf{H}_0 e^{i(\omega t + \mathbf{k} \cdot \mathbf{x})}, \quad (3.11b)$$

$$\mathbf{D} = \mathbf{D}_0 e^{i(\omega t + \mathbf{k} \cdot \mathbf{x})}, \quad (3.11c)$$

$$\mathbf{B} = \mathbf{B}_0 e^{i(\omega t + \mathbf{k} \cdot \mathbf{x})}. \quad (3.11d)$$

By inserting (3.11a)–(3.11d) into (3.10a)–(3.10d), we obtain the following relations

$$\omega \mathbf{D}_0 - \mathbf{k} \times \mathbf{H}_0 = 0, \quad (3.12a)$$

$$\omega \mathbf{B}_0 + \mathbf{k} \times \mathbf{E}_0 = 0, \quad (3.12b)$$

$$\mathbf{D}_0 = \underline{\underline{\varepsilon}}_0 \mathbf{E}_0, \quad (3.12c)$$

$$\mathbf{B}_0 = \underline{\underline{\mu}}_0 \mathbf{H}_0. \quad (3.12d)$$

If $\omega \neq 0$, (3.12a)–(3.12b) show that \mathbf{D} is orthogonal to \mathbf{H} and \mathbf{k} and \mathbf{B} is orthogonal to \mathbf{E} and \mathbf{k} . The case $\omega = 0$ will be discussed below.

On the other hand, since $\underline{\underline{\varepsilon}}_0$ and $\underline{\underline{\mu}}_0$ are symmetric positive definite, we can write $\underline{\underline{\varepsilon}}_0 = \tilde{\varepsilon}_0^* \tilde{\varepsilon}_0$ and $\underline{\underline{\mu}}_0 = \tilde{\mu}_0^* \tilde{\mu}_0$. By multiplying (3.12c) and (3.12d) by \mathbf{H}_0 and \mathbf{E}_0 respectively, we obtain, thanks to these decompositions

$$\tilde{\varepsilon}_0 \mathbf{E}_0 \cdot \tilde{\varepsilon}_0 \mathbf{H}_0 = 0, \quad (3.13a)$$

$$\tilde{\mu}_0 \mathbf{E}_0 \cdot \tilde{\mu}_0 \mathbf{H}_0 = 0. \quad (3.13b)$$

A similar process provides

$$\tilde{\varepsilon}_0 \mathbf{E}_0 \cdot \tilde{\varepsilon}_0 \mathbf{k} = 0, \quad (3.14a)$$

$$\tilde{\mu}_0 \mathbf{H}_0 \cdot \tilde{\mu}_0 \mathbf{k} = 0, \quad (3.14b)$$

which shows that the vector $\tilde{\varepsilon}_0 \mathbf{E}$ is orthogonal to $\tilde{\varepsilon}_0 \mathbf{H}$ and $\tilde{\varepsilon}_0 \mathbf{k}$ and the vector $\tilde{\mu}_0 \mathbf{H}$ is orthogonal to $\tilde{\mu}_0 \mathbf{k}$ and $\tilde{\mu}_0 \mathbf{E}$.

Now, by eliminating \mathbf{E}_0 , \mathbf{D}_0 and \mathbf{B}_0 in (3.12a)–(3.12d), we obtain the following dispersion relation:

$$\omega^2 \mathbf{H}_0 = -\underline{\underline{\mu}}_0^{-1} \left(\mathbf{k} \times \left(\underline{\underline{\varepsilon}}_0^{-1} (\mathbf{k} \times \mathbf{H}_0) \right) \right). \quad (3.15)$$

In other words, there are three velocities which are the square roots of the eigenvalues of matrix M_0 defined by

$$M_0 \mathbf{H}_0 = -\frac{1}{|\mathbf{k}|^2} \underline{\underline{\mu}}_0^{-1} \left(\mathbf{k} \times \left(\underline{\underline{\varepsilon}}_0^{-1} (\mathbf{k} \times \mathbf{H}_0) \right) \right). \quad (3.16)$$

Obviously, 0 is an eigenvalue of M_0 . The corresponding eigenvector is colinear to \mathbf{k} . So $\omega = 0$ provides a stationary mode for which the field \mathbf{H} is parallel to

\mathbf{k} . This stationary solution can be written as the gradient of a scalar potential ψ . The other two modes lead to two dispersion relations which correspond to two waves with different polarizations and different velocities.

In the isotropic case, one can easily check that the electric and magnetic fields are orthogonal and are both orthogonal to \mathbf{k} . The stationary mode still exists but M_0 has a double eigenvalue. Actually, when $\omega \neq 0$, the problem can be written as

$$\omega^2 \mathbf{H}_0 = \frac{1}{\varepsilon_0 \mu_0} |\mathbf{k}|^2 \mathbf{H}_0. \quad (3.17)$$

Equation (3.17) leads to the following dispersion relation

$$\varepsilon_0 \mu_0 \omega^2 = |\mathbf{k}|^2, \quad (3.18)$$

which shows that the velocity c of the waves is equal to $1/\sqrt{\varepsilon_0 \mu_0}$.

Remark

A similar development can be carried out for the electric field. In the isotropic case the same dispersion relation is obtained.

3.2.2 The 2D Case

In the 2D case, we shall present the TM polarization defined in (1.16a) and (1.16b) but an equivalent study can be carried out for the TE polarization.

So, let us consider the 2D Maxwell system in a homogeneous anisotropic medium:

$$\underline{\underline{\varepsilon}}_0 \frac{\partial \mathbf{E}}{\partial t} - \mathbf{curl} \mathbf{H} = 0, \quad (3.19a)$$

$$\mu_0 \frac{\partial \mathbf{H}}{\partial t} + \mathbf{curl} \mathbf{E} = 0, \quad (3.19b)$$

for which we are looking for a plane wave solution derived from (3.11a)–(3.11d).

Let us set:

$$\underline{\underline{\varepsilon}}_0^{-1} = \begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix}. \quad (3.20)$$

After inserting this solution into (3.19a) and (3.19b) and eliminating \mathbf{E}_0 , we obtain the following dispersion relation

$$\mu_0 \omega^2 = \gamma k_1^2 + \alpha k_2^2 - 2\beta k_1 k_2, \quad (3.21)$$

where k_1 and k_2 are the components of the vector \mathbf{k} .

Now, if we set $k_1 = |\mathbf{k}|^2 \cos \theta$ and $k_2 = |\mathbf{k}|^2 \sin \theta$, we obtain²

$$c = \frac{\omega}{|\mathbf{k}|} = \sqrt{\frac{1}{\mu_0} (\alpha \sin^2 \theta + \gamma \cos^2 \theta - 2\beta \sin \theta \cos \theta)}, \quad (3.22)$$

which provides an equation in polar coordinates of the velocity c . This equation indicates the anisotropy of c . For instance, when $\mu_0 = 1$, $\alpha = 32/31$, $\beta = -4/31$ and $\gamma = 16/31$ (which provides a matrix whose inverse is the positive definite matrix A defined in Sect. 13.2.8), we obtain the curve drawn in Fig. 3.1.

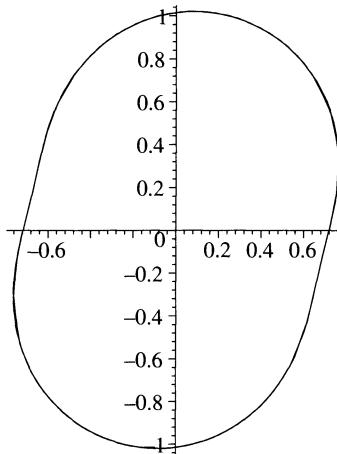


Fig. 3.1. An example of an anisotropy curve of the velocity for the 2D Maxwell system

3.3 Application to the Elastics System

We are looking for a plane wave solution of (1.36) when Lamé's coefficients are constant. The plane wave solution is

$$\mathbf{v} = \mathbf{v}_0 e^{i(\omega t + \mathbf{k} \cdot \mathbf{x})}. \quad (3.23)$$

By inserting (1.6d) into (1.36), we obtain the dispersion relation

$$(\mu |\mathbf{k}|^2 - s) \mathbf{v}_0 + (\lambda + \mu)(\mathbf{k} \cdot \mathbf{v}_0) \mathbf{k} = 0, \quad (3.24)$$

where $s = \rho \omega^2$.

² The positivity of the expression under the square root comes from the positive character of the matrix $\underline{\underline{\varepsilon}}_0^{-1}$.

Equation (3.24) has two classes of solutions:

1. $\mathbf{v}_0 \parallel \mathbf{k} \Rightarrow s = (\lambda + 2\mu)|\mathbf{k}|^2.$

2. $\mathbf{v}_0 \perp \mathbf{k} \Rightarrow s = \mu|\mathbf{k}|^2.$

These two classes provide two kinds of waves with two different dispersion relations:

1. Waves parallel to the direction of propagation whose dispersion relation is

$$\omega_P^2 = \frac{\lambda + 2\mu}{\rho} |\mathbf{k}|^2. \quad (3.25)$$

2. Waves perpendicular to the direction of propagation whose dispersion relation is

$$\omega_S^2 = \frac{\mu}{\rho} |\mathbf{k}|^2. \quad (3.26)$$

So, by setting $V_P = \sqrt{(\lambda + 2\mu)/\rho}$ and $V_S = \sqrt{\mu/\rho}$, we obtain the P -wave and S -wave defined in (1.40) and (1.41).

4. Construction of the Schemes in Homogeneous Media

4.1 A Model Problem

We want to approximate the wave equation (1.1) with the standard initial conditions (1.4). As a first step, we shall present schemes in \mathbb{R}^N so that we do not have to deal with boundary conditions. Moreover, in this section, κ and ρ will be assumed not to depend on position. So we shall consider the wave equation

$$\frac{\partial^2 u}{\partial t^2} - c^2 \Delta u = 0, \quad (4.1)$$

where c is defined as in (1.3).

4.2 Second-Order Approximation in Space

Before presenting higher-order schemes, we shall introduce some ideas about second-order schemes which will be useful later.

4.2.1 The 1D Case

Let \mathcal{M}_h denote a uniform grid of \mathbb{R} with a step equal to h . For any function u , we shall denote $u_\ell = u(\ell h)$. On this grid, we want to construct a centered second-order approximation of the operator $\partial^2/\partial x^2$. By taking into account symmetries of the problem, this approximation can be written as¹:

$$\left(\frac{\partial^2 u}{\partial x^2} \right)_{h,\ell}^{(2)} = Au_\ell + a(u_{\ell+1} + u_{\ell-1}) = \frac{\partial^2 u}{\partial x^2}(x_\ell) + O(h^2), \quad (4.2)$$

where $\left(\frac{\partial^2 u}{\partial x^2} \right)_{h,\ell}^{(2)}$ is a second-order centered approximation of $\frac{\partial^2}{\partial x^2}$ at $x_\ell = \ell h$.

¹ In all the following, the upper index in parenthesis will indicate the order of approximation.

Now, let us write the Taylor expansion² of u at $x_{\ell-1}$ and $x_{\ell+1}$:

$$\begin{aligned} u(x_{\ell+1}) &= u(x_\ell) + h \frac{\partial u}{\partial x}(x_\ell) + \frac{h^2}{2} \frac{\partial^2 u}{\partial x^2}(x_\ell) + \frac{h^3}{6} \frac{\partial^3 u}{\partial x^3}(x_\ell) \\ &\quad + \frac{h^4}{24} \frac{\partial^4 u}{\partial x^4}(x_\ell) + O(h^6), \end{aligned} \quad (4.3a)$$

$$\begin{aligned} u(x_{\ell-1}) &= u(x_\ell) - h \frac{\partial u}{\partial x}(x_\ell) + \frac{h^2}{2} \frac{\partial^2 u}{\partial x^2}(x_\ell) - \frac{h^3}{6} \frac{\partial^3 u}{\partial x^3}(x_\ell) \\ &\quad + \frac{h^4}{24} \frac{\partial^4 u}{\partial x^4}(x_\ell) + O(h^6). \end{aligned} \quad (4.3b)$$

By inserting (4.3a) and (4.3b) into (4.2), we obtain:

$$\begin{aligned} \frac{\partial^2 u}{\partial x^2}(x_\ell) + O(h^2) &= Au_\ell + a(u_{\ell+1} + u_{\ell-1}) \\ &= (A + 2a)u_\ell + ah^2 \frac{\partial^2 u}{\partial x^2}(x_\ell) \\ &\quad + a \frac{h^4}{12} \frac{\partial^4 u}{\partial x^4}(x_\ell) + O(h^6). \end{aligned} \quad (4.4)$$

So, in order to obtain a second-order approximation of $\partial^2 u / \partial x^2(x_\ell)$, A and a must satisfy

$$\left\{ \begin{array}{l} A + 2a = 0, \\ ah^2 = 1, \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} A = -\frac{2}{h^2}, \\ a = \frac{1}{h^2}, \end{array} \right. \quad (4.5)$$

which provides

$$\left(\frac{\partial^2 u}{\partial x^2} \right)_{h,\ell}^{(2)} = \frac{1}{h^2} (u_{\ell+1} - 2u_\ell + u_{\ell-1}) = \frac{\partial^2 u}{\partial x^2} + \frac{h^2}{12} \frac{\partial^4 u}{\partial x^4} + O(h^4). \quad (4.6)$$

Now, in order to obtain the basic properties of this approximation we introduce the staggered l^2 normed spaces³

$$V_0^1 = \{(v_\ell), \quad \ell \in \mathbb{Z} | \quad h \sum_{\ell \in \mathbb{Z}} |v_\ell|^2 < \infty\}, \quad (4.7)$$

² We shall write all the terms of this expansion but, in practice, only the even terms must be considered for even operators and odd terms for odd operators. The other terms vanish for obvious symmetry reasons.

³ We recall that l^2 is the space of sequences $(u_k)_{k \in \mathbb{Z}}$ such that $\sum_{k \in \mathbb{Z}} |u_k|^2 < \infty$.

$$V_{\frac{1}{2}}^1 = \{(v_{\ell+\frac{1}{2}}), \quad \ell \in \mathbb{Z} \mid h \sum_{\ell \in \mathbb{Z}} |v_{\ell+\frac{1}{2}}|^2 < \infty\}, \quad (4.8)$$

with the scalar products $(\cdot, \cdot)_0$ and $(\cdot, \cdot)_{\frac{1}{2}}$ derived from the norms and the discrete operators

$$D_{1,h}^{(2)} : V_0^1 \rightarrow V_{\frac{1}{2}}^1 \text{ such that } (D_{1,h}^{(2)} v)_{\ell+\frac{1}{2}} = \frac{v_{\ell+1} - v_\ell}{h}, \quad (4.9)$$

$$\tilde{D}_{1,h}^{(2)} : V_{\frac{1}{2}}^1 \rightarrow V_0^1 \text{ such that } (\tilde{D}_{1,h}^{(2)} v)_\ell = \frac{v_{\ell+\frac{1}{2}} - v_{\ell-\frac{1}{2}}}{h}. \quad (4.10)$$

With this notation and by denoting \mathcal{D}^* the adjoint of a discrete operator \mathcal{D} for the l^2 scalar product, we obtain

Proposition 1. $\tilde{D}_{1,h}^{(2)} = - (D_{1,h}^{(2)})^*$ and $(\frac{\partial^2 u}{\partial x^2})_h^{(2)} = - (D_{1,h}^{(2)})^* \circ D_{1,h}^{(2)} u$.

Proof. Let be $u \in V_{\frac{1}{2}}^1$ and $v \in V_0^1$.

By definition

$$\begin{aligned} (\tilde{D}_{1,h}^{(2)} u, v)_0 &= h \sum_{\ell \in \mathbb{Z}} \frac{u_{\ell+\frac{1}{2}} - u_{\ell-\frac{1}{2}}}{h} v_\ell \\ &= h \sum_{\ell \in \mathbb{Z}} \frac{u_{\ell+\frac{1}{2}} v_\ell}{h} - \sum_{\ell \in \mathbb{Z}} \frac{u_{\ell+\frac{1}{2}} v_{\ell+1}}{h} \\ &= h \sum_{\ell \in \mathbb{Z}} u_{\ell+\frac{1}{2}} \frac{v_\ell - v_{\ell+1}}{h} = (u, -D_{1,h}^{(2)} v)_{\frac{1}{2}}. \end{aligned}$$

Hence $\tilde{D}_{1,h}^{(2)} = - (D_{1,h}^{(2)})^*$.

On the other hand,

$$\begin{aligned} \left(\frac{\partial^2 u}{\partial x^2}\right)_{h,\ell}^{(2)} &= \frac{1}{h^2} (u_{\ell+1} - 2u_\ell + u_{\ell-1}) \\ &= \frac{\frac{u_{\ell+1} - u_\ell}{h} - \frac{u_\ell - u_{\ell-1}}{h}}{h} \\ &= \left(\tilde{D}_{1,h}^{(2)} \circ D_{1,h}^{(2)} u\right)_\ell. \end{aligned} \quad \diamond$$

Proposition 1 shows that the second-order approximation of $\partial^2/\partial x^2$ can be written as the composition of second-order approximations of the first-order derivative on staggered grids.

4.2.2 The 2D Case

We now extend this study to the 2D case. Since the space variables play a symmetric role, we shall take the same space-step h in both directions. Let

$$M_0^h = \{(\ell h, mh), (\ell, m) \in \mathbb{Z}^2\}, \quad M_{\frac{1}{2}}^h = \{((\ell + \frac{1}{2})h, mh), (\ell, m) \in \mathbb{Z}^2\},$$

$$M_{\frac{1}{2}}^h = \{(\ell h, (m + \frac{1}{2})h), (\ell, m) \in \mathbb{Z}^2\},$$

be three grids of \mathbb{R}^2 with a space-step h and let us define the staggered l^2 normed spaces

$$V_0^2 = \{(v_{\ell,m}), (\ell, m) \in \mathbb{Z}^2 \mid h^2 \sum_{(\ell,m) \in \mathbb{Z}^2} |v_{\ell,m}|^2 < \infty\}, \quad (4.11)$$

$$\begin{aligned} V_{\frac{1}{2}}^2 &= \{(v_{\ell+\frac{1}{2},m}, v_{\ell,m+\frac{1}{2}}), (\ell, m) \in \mathbb{Z}^2 \mid \\ &\quad h^2 \sum_{(\ell,m) \in \mathbb{Z}^2} (|v_{\ell+\frac{1}{2},m}|^2 + |v_{\ell,m+\frac{1}{2}}|^2) < \infty\}, \end{aligned} \quad (4.12)$$

with the corresponding scalar products (as in 1D).

For any function u , we denote $u_{p,q} = u(ph, qh)$, $(p, q) \in \mathbb{Z}^2$.

Since $\Delta = \partial^2/\partial x_1^2 + \partial^2/\partial x_2^2$, its second-order approximation $\Delta_h^{(2)}$ to $(\partial^2/\partial x_1^2)_h^{(2)} + (\partial^2/\partial x_2^2)_h^{(2)}$. So,

$$\begin{aligned} (\Delta_h^{(2)} u)_{\ell,m} &= \frac{1}{h^2} [(u_{\ell+1,m} - 2u_{\ell,m} + u_{\ell-1,m}) \\ &\quad + (u_{\ell,m+1} - 2u_{\ell,m} + u_{\ell,m-1})] \\ &= \frac{1}{h^2} (-4u_{\ell,m} + u_{\ell+1,m} + u_{\ell-1,m} \\ &\quad + u_{\ell,m+1} + u_{\ell,m-1}). \end{aligned} \quad (4.13)$$

As in the 1D case, we introduce the discrete operators

$$D_{2,h}^{(2)} : V_0^2 \rightarrow V_{\frac{1}{2}}^2 \text{ such that } (D_{2,h}^{(2)} v)_{\ell+\frac{1}{2},m+\frac{1}{2}} = \begin{pmatrix} \frac{v_{\ell+1,m} - v_{\ell,m}}{h} \\ \frac{v_{\ell,m+1} - v_{\ell,m}}{h} \end{pmatrix}, \quad (4.14)$$

$$\tilde{D}_{2,h}^{(2)} : V_{\frac{1}{2}}^1 \rightarrow V_0^2 \text{ such that } \left(\tilde{D}_{2,h}^{(2)} v \right)_{\ell,m} = \frac{v_{\ell+\frac{1}{2},m} - v_{\ell-\frac{1}{2},m}}{h} + \frac{v_{\ell,m+\frac{1}{2}} - v_{\ell,m-\frac{1}{2}}}{h} \quad (4.15)$$

and we have the following result:

Proposition 2. $\tilde{D}_{2,h}^{(2)} = - \left(D_{2,h}^{(2)} \right)^*$ and $\Delta_h^{(2)} u = - \left(D_{2,h}^{(2)} \right)^* \circ D_{2,h}^{(2)} u$.

Proof. For any $u \in V_{\frac{1}{2}}^2$ et $v \in V_0^2$, we have

$$\begin{aligned} (\tilde{D}_{2,h}^{(2)} u, v)_0 &= h^2 \sum_{(\ell,m) \in \mathbb{Z}^2} \frac{u_{\ell+\frac{1}{2},m} - u_{\ell-\frac{1}{2},m}}{h} v_{\ell,m} + \frac{u_{\ell,m+\frac{1}{2}} - u_{\ell,m-\frac{1}{2}}}{h} v_{\ell,m} \\ &= h^2 \sum_{(\ell,m) \in \mathbb{Z}^2} u_{\ell+\frac{1}{2},m} \frac{v_{\ell,m} - v_{\ell+1,m}}{h} + u_{\ell,m+\frac{1}{2}} \frac{v_{\ell,m} - v_{\ell,m+1}}{h} \\ &= \left(u, - \left(D_{2,h}^{(2)} \right)^* v \right)_{\frac{1}{2}}, \end{aligned}$$

which provides the first relation.

On the other hand,

$$\begin{aligned} \left(\Delta_h^{(2)} u \right)_{\ell,m} &= \frac{\frac{u_{\ell+1,m} - u_{\ell,m}}{h} - \frac{u_{\ell,m} - u_{\ell-1,m}}{h}}{h} \\ &\quad + \frac{\frac{u_{\ell,m+1} - u_{\ell,m}}{h} - \frac{u_{\ell,m} - u_{\ell,m-1}}{h}}{h} \\ &= \left(\tilde{D}_{2,h}^{(2)} \circ D_{2,h}^{(2)} u \right)_{\ell,m} = - \left(\left(D_{2,h}^{(2)} \right)^* \circ D_{2,h}^{(2)} u \right)_{\ell,m}. \quad \diamond \end{aligned}$$

A more general class of second-order finite difference approximations is given by introducing diagonal terms which leads to the following 9-point approximation:

$$\begin{aligned} &\left(\check{\Delta}_h^{(2)} u \right)_{\ell,m} \\ &= \frac{1}{h^2} [A u_{\ell,m} + a(u_{\ell+1,m} + u_{\ell-1,m} + u_{\ell,m+1} + u_{\ell,m-1}) \\ &\quad + b(u_{\ell+1,m+1} + u_{\ell-1,m+1} + u_{\ell+1,m-1} + u_{\ell-1,m-1})]. \end{aligned} \quad (4.16)$$

By writing the (2D) Taylor expansions of the terms involved in (4.16), we obtain

$$\begin{aligned}
& h^2 \left(\check{\Delta}_h^{(2)} u \right)_{\ell,m} \\
&= (A + 4a + 4b)u(\ell h, mh) + h^2(a + 2b)\Delta u(\ell h, mh) \\
&\quad + h^4 \left[\left(\frac{a}{12} + \frac{b}{6} \right) \left(\frac{\partial^4 u}{\partial x_1^4}(\ell h, mh) + \frac{\partial^4 u}{\partial x_2^4}(\ell h, mh) \right) \right. \\
&\quad \left. + b \frac{\partial^4 u}{\partial x_1^2 \partial x_2^2}(\ell h, mh) \right] + O(h^6).
\end{aligned} \tag{4.17}$$

So, it is necessary and sufficient to have

$$A + 4a + 4b = 0, \tag{4.18a}$$

$$a + 2b = 1, \tag{4.18b}$$

to obtain a second-order approximation.

Equations (4.18a) and (4.18b) provide a one-parameter family of second-order approximations. Among them, we obtain our classical approximation (4.13) by setting $b = 0$ and its $\pi/4$ -rotated version for $a = 0$. Another interesting approximation (called Arakawa's scheme) is obtained by looking for a truncation error proportional to $\Delta^2 u(\ell h, mh)$, i.e.

$$\frac{b}{2} = \frac{a}{12} + \frac{b}{6}. \tag{4.19}$$

By combining (4.18a), (4.18b) and (4.19), we obtain

$$A = -\frac{10}{3}, \quad a = \frac{2}{3}, \quad b = \frac{1}{6}. \tag{4.20}$$

Such an approximation is called maxi-isotropic because the dependence of its error of approximation on the direction of propagation of the wave is minimal.

Remarks

1. $D_{2,h}^{(2)}$ and $\tilde{D}_{2,h}^{(2)}$ are second-order finite difference approximations of the gradient and the divergence. So, as for Δ , we can write $\Delta_h^{(2)} = \text{div}_h^{(2)} \circ \text{grad}_h^{(2)}$.
2. The properties given in Propositions 1 and 2 will lead to positivity properties of the discrete operators⁴. Actually they can be generalized to any centered approximation of Δ based on staggered approximations of the divergence and the gradient. These properties will also be very useful in the construction of discrete operators in heterogeneous media.

⁴ This is a necessary condition of the stability of the schemes in time since $-\Delta$ is a positive operator.

3. Arakawa's scheme is wrongly considered as a fourth-order scheme. Actually, it is fourth-order only for solving the equation

$$\Delta u = 0 \quad (4.21)$$

(since the second-order truncation error in (4.17) is proportional to a power of Δ) but, as soon as one has a right-hand side f (a fortiori when the equation contains $\partial^2 u / \partial t^2$), it loses this fourth-order character.

4.3 Fourth-Order Approximations in Space

Unlike second-order approximations, the fourth-order approximations of the Laplace operator Δ are not the same if we search for a global approximation of Δ [34, 35], or for the composition of fourth-order approximations of the divergence and the gradient [102]. Each approach has its own advantages and drawbacks. For this reason, we shall deal with these two fundamental approaches in the following.

4.3.1 First Approach: Global Approximation of Δ

In 1D, the discrete Laplace operator can be written as

$$\left(\frac{\partial^2 u}{\partial x^2} \right)_{h,\ell}^{(4)} = Au_\ell + a(u_{\ell+1} + u_{\ell-1}) + b(u_{\ell+2} + u_{\ell-2}). \quad (4.22)$$

By combining the Taylor expansions⁵ of u at the points involved in (4.22)

$$\begin{aligned} a \times u_{\ell+1} &= u_\ell + \frac{h^2}{2} u''_\ell + \frac{h^4}{24} u_\ell^{(iv)} + \frac{h^6}{720} u_\ell^{(vi)} + O(h^7), \\ a \times u_{\ell-1} &= u_\ell + \frac{h^2}{2} u''_\ell + \frac{h^4}{24} u_\ell^{(iv)} + \frac{h^6}{720} u_\ell^{(vi)} + O(h^7), \\ b \times u_{\ell+2} &= u_\ell + 2h^2 u''_\ell + \frac{2h^4}{3} u_\ell^{(iv)} + \frac{4h^6}{45} u_\ell^{(vi)} + O(h^7), \\ b \times u_{\ell-2} &= u_\ell + 2h^2 u''_\ell + \frac{2h^4}{3} u_\ell^{(iv)} + \frac{4h^6}{45} u_\ell^{(vi)} + O(h^7), \end{aligned}$$

we obtain

$$\left\{ \begin{array}{l} A + 2a + 2b = 0, \\ a + 4b = \frac{1}{h^2}, \\ a + 16b = 0, \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} A = -\frac{5}{2h^2}, \\ a = \frac{4}{3h^2}, \\ b = -\frac{1}{12h^2}. \end{array} \right. \quad (4.23)$$

⁵ We shall only write the even terms, as indicated in footnote 2.

So, by inserting (4.23) into (4.22), we obtain the 5-point approximation:

$$\begin{aligned}
 \left(\frac{\partial^2 u}{\partial x^2} \right)_{h,\ell}^{(4)} &= \frac{1}{h^2} \left(-\frac{1}{12} u_{\ell-2} + \frac{4}{3} u_{\ell-1} \right. \\
 &\quad \left. - \frac{5}{2} u_\ell + \frac{4}{3} u_{\ell+1} - \frac{1}{12} u_{\ell+2} \right) \\
 &= \frac{4}{3} \frac{1}{h^2} (u_{\ell+1} - 2u_\ell + u_{\ell-1}) \\
 &\quad - \frac{1}{3} \frac{1}{4h^2} (u_{\ell+2} - 2u_\ell + u_{\ell-2}) \\
 &= \frac{4}{3} \left(\frac{\partial^2 u}{\partial x^2} \right)_{h,\ell}^{(2)} - \frac{1}{3} \left(\frac{\partial^2 u}{\partial x^2} \right)_{2h,\ell}^{(2)}.
 \end{aligned} \tag{4.24}$$

In 2D, we combine the two 1D derivatives of Δ and we obtain the 9-point approximation:

$$\begin{aligned}
 \left(\Delta_h^{(4)} u \right)_{\ell,m} &= \frac{1}{h^2} (-5u_{\ell,m} \\
 &\quad + \frac{4}{3} (u_{\ell+1,m} + u_{\ell-1,m} + u_{\ell,m+1} + u_{\ell,m-1}) \\
 &\quad - \frac{1}{12} (u_{\ell+2,m} + u_{\ell-2,m} + u_{\ell,m+2} + u_{\ell,m-2})),
 \end{aligned} \tag{4.25}$$

which provides

$$\begin{aligned}
 \left(\Delta_h^{(4)} u \right)_{\ell,m} &= \frac{4}{3} \frac{1}{h^2} (-4u_{\ell,m} + u_{\ell+1,m} + u_{\ell-1,m} \\
 &\quad + u_{\ell,m+1} + u_{\ell,m-1}) \\
 &\quad - \frac{1}{3} \frac{1}{4h^2} (-4u_{\ell,m} + u_{\ell+2,m} + u_{\ell-2,m} \\
 &\quad + u_{\ell,m+2} + u_{\ell,m-2}) \\
 &= \frac{4}{3} \left(\Delta_h^{(2)} u \right)_{\ell,m} - \frac{1}{3} \left(\Delta_{2h}^{(2)} u \right)_{\ell,m}.
 \end{aligned} \tag{4.26}$$

Remarks

1. We could a priori set $(\partial^2 u / \partial x^2)_{h,\ell}^{(4)} = \lambda (\partial^2 u / \partial x^2)_{h,\ell}^{(2)} + \mu (\partial^2 u / \partial x^2)_{2h,\ell}^{(2)}$ and derive the coefficients from the Taylor expansions of $(\partial^2 u / \partial x^2)_{h,\ell}^{(2)}$

and $(\partial^2 u / \partial x^2)_{2h,\ell}^{(2)}$. This method will be used to determine higher-order approximations.

- Relations (4.24) and (4.25) show that this first approximation can be written as a combination of two second-order approximations on grids with space-steps h and $2h$. This combination will be very useful for the extension of the schemes to heterogeneous media. Unfortunately, it contains a negative coefficient which will be a problem for the positivity of the discrete operators and, consequently, for the stability analysis of the schemes arising from such approximations, in particular in heterogeneous media.

4.3.2 Second Approach: Fourth-Order Approximations of the First-Order Operators

For this point of view, we first obtain, by using Taylor expansions, staggered fourth-order approximations of $\partial/\partial x$:

$$\left(D_{1,h}^{(4)} u\right)_{\ell+\frac{1}{2}} = \left(\frac{\partial u}{\partial x}\right)_{h,\ell+\frac{1}{2}}^{(4)} = \frac{1}{24h}(u_{\ell-1} - 27u_\ell + 27u_{\ell+1} - u_{\ell+2}), \quad (4.27)$$

$$\left(\tilde{D}_{1,h}^{(4)} u\right)_\ell = \left(\frac{\partial u}{\partial x}\right)_{h,\ell}^{(4)} = \frac{1}{24h}(u_{\ell-\frac{3}{2}} - 27u_{\ell-\frac{1}{2}} + 27u_{\ell+\frac{1}{2}} - u_{\ell+\frac{3}{2}}). \quad (4.28)$$

These discrete operators are the fourth-order versions of the second order operators defined in (4.9) and (4.10). So, we can write

$$\begin{aligned} \left(\frac{\partial^2 u}{\partial x^2}\right)_{h,\ell}^{(4)} &= \left(\tilde{D}_{1,h}^{(4)} \circ D_{1,h}^{(4)} u\right)_\ell \\ &= \frac{1}{h^2} \left(\frac{1}{576} u_{\ell-3} - \frac{3}{32} u_{\ell-2} + \frac{87}{64} u_{\ell-1} - \frac{365}{144} u_\ell \right. \\ &\quad \left. + \frac{87}{64} u_{\ell+1} - \frac{3}{32} u_{\ell+2} + \frac{1}{576} u_{\ell+3} \right). \end{aligned} \quad (4.29)$$

We obtain a 7-point scheme about which we have, as for second-order, the following proposition:

Proposition 3. $\tilde{D}_{1,h}^{(4)} = -\left(D_{1,h}^{(4)}\right)^*$ and $\left(\frac{\partial^2 u}{\partial x^2}\right)_h^{(4)} = -\left(D_{1,h}^{(4)}\right)^* \circ D_{1,h}^{(4)} u$.

The proof is the same as in the second-order case.

By writing that $\nabla = (\partial/\partial x_1, \partial/\partial x_2)^T$, we obtain the 2D discrete operators:

$D_{2,h}^{(4)} : V_0^2 \rightarrow V_{\frac{1}{2}}^2$ such that:

$$\begin{aligned} & \left(D_{2,h}^{(4)} u \right)_{\ell+\frac{1}{2}, m+\frac{1}{2}} = \\ & \left(\frac{1}{24h} (u_{\ell-1,m} - 27u_{\ell,m} + 27u_{\ell+1,m} - u_{\ell+2,m}), \right. \\ & \left. \frac{1}{24h} (u_{\ell,m-1} - 27u_{\ell,m} + 27u_{\ell,m+1} - u_{\ell,m+2}) \right), \end{aligned} \quad (4.30)$$

$\tilde{D}_{2,h}^{(4)} : V_{\frac{1}{2}}^1 \rightarrow V_0^2$ such that:

$$\begin{aligned} & \left(\tilde{D}_{2,h}^{(4)} u \right)_{\ell,m} = \\ & \frac{1}{24h} (u_{\ell-\frac{3}{2},m} - 27u_{\ell-\frac{1}{2},m} + 27u_{\ell+\frac{1}{2},m} - u_{\ell+\frac{3}{2},m}) \\ & + \frac{1}{24h} (u_{\ell,m-\frac{3}{2}} - 27u_{\ell,m-\frac{1}{2}} + 27u_{\ell,m+\frac{1}{2}} - u_{\ell,m+\frac{3}{2}}), \end{aligned} \quad (4.31)$$

which are the approximations of the gradient and the divergence respectively. By combining these two operators, we obtain the following 13-point approximation of Δ :

$$\begin{aligned} \left(\Delta_h^{(4')} u \right)_{\ell,m} &= \left(\tilde{D}_{2,h}^{(4)} \circ D_{2,h}^{(4)} u \right)_{\ell,m} \\ &= \frac{1}{h^2} \left(-\frac{365}{72} u_{\ell,m} \right. \\ &\quad \left. + \frac{87}{64} (u_{\ell+1,m} + u_{\ell-1,m} + u_{\ell,m+1} + u_{\ell,m-1}) \right. \\ &\quad \left. - \frac{3}{32} (u_{\ell+2,m} + u_{\ell-2,m} + u_{\ell,m+2} + u_{\ell,m-2}) \right. \\ &\quad \left. + \frac{1}{576} (u_{\ell+3,m} + u_{\ell-3,m} + u_{\ell,m+3} + u_{\ell,m-3}) \right). \end{aligned} \quad (4.32)$$

As in the 1D case, we have the following result:

Proposition 4. $\tilde{D}_{2,h}^{(4)} = - \left(D_{2,h}^{(4)} \right)^*$ and $\Delta_h^{(4')} = - \left(D_{2,h}^{(4)} \right)^* \circ D_{2,h}^{(4)}$

The proof is obtained in the same way as for the second-order approximation.

Although more expensive than the approximations obtained with the first approach, these new approximations have better positivity and, hence, stability properties than the first ones. This is because of their decomposition into first-order discrete operators.

Our study shows that, unlike the second-order approximations for which the global approximation of Δ gives $\Delta_h^{(2)} = \text{div}_h^{(2)} \circ \mathbf{grad}_h^{(2)}$, in the fourth-order case, $\Delta_h^{(4)} \neq \text{div}_h^{(4)} \circ \mathbf{grad}_h^{(4)}$. The equality is only obtained by the second approach in Proposition 4.

Remark

The second approach provides also a fourth-order approximation on staggered grids (see Sect. 4.7.1) of the wave equation written as a first-order system:

$$\frac{\partial u}{\partial t} = c^2 \text{div} v, \quad (4.33a)$$

$$\frac{\partial v}{\partial t} = \mathbf{grad} u. \quad (4.33b)$$

Although more expensive for finite difference approximations, this formulation has some advantages, as we shall see in Chap. 13.

4.4 Approximation in Time

The most popular scheme in time is the leapfrog scheme given by

$$\frac{u_h^{n+1} - 2u_h^n + u_h^{n-1}}{\Delta t^2} = c^2 \Delta_h^{(4)} u_h^n, \quad (4.34)$$

which is second-order accurate in time

However, since a fourth-order approximation in space is more accurate, it is legitimate to search for fourth-order approximations in time, a priori centered, since uncentered schemes would be dissipative. The first idea would be to approximate the second-order derivative in time of (4.1) by a 5-point or 7-point approximation defined in (4.24) and (4.29). The scheme derived from such approximations could be written, for a 5-point scheme, for instance

$$\frac{1}{\Delta t^2} \left(-\frac{1}{12} u_h^{n+2} + \frac{4}{3} u_h^{n+1} - \frac{5}{2} u_h^n + \frac{4}{3} u_h^{n-1} - \frac{1}{12} u_h^{n-2} \right) = c^2 \Delta_h^{(4)} u_h^n. \quad (4.35)$$

Unfortunately, such approximations and, more generally, any higher-order approximation⁶ of the time derivative lead to unconditionally unstable schemes [62].

⁶ A priori, even uncentered.

4.4.1 The Modified Equation Approach

A very efficient alternative, introduced by Dablain in 1986 [42], is the modified equation approach. This approach is constructed as follows:

The Taylor expansion of the leapfrog scheme can be written as

$$\frac{u^{n+1} - 2u^n + u^{n-1}}{\Delta t^2} = \left(\frac{\partial^2 u}{\partial t^2} \right)^n + \frac{\Delta t^2}{12} \left(\frac{\partial^4 u}{\partial t^4} \right)^n + O(\Delta t^4). \quad (4.36)$$

By using the continuous wave equation (4.1), we obtain

$$\frac{\partial^4 u}{\partial t^4} = \frac{\partial^2}{\partial t^2} \left(\frac{\partial^2 u}{\partial t^2} \right) = \frac{\partial^2}{\partial t^2} (c^2 \Delta u) = c^2 \Delta \left(\frac{\partial^2 u}{\partial t^2} \right) = c^4 \Delta^2 u. \quad (4.37)$$

So,

$$\frac{u^{n+1} - 2u^n + u^{n-1}}{\Delta t^2} - \frac{c^4 \Delta t^2}{12} \Delta^2 u^n = \left(\frac{\partial^2 u}{\partial t^2} \right)^n + O(\Delta t^4), \quad (4.38)$$

which finally provides

$$\frac{u_h^{n+1} - 2u_h^n + u_h^{n-1}}{\Delta t^2} - \frac{c^4 \Delta t^2}{12} (\Delta^2)_h u_h^n - c^2 \Delta_h u_h^n = 0. \quad (4.39)$$

Equation (4.39) is a fourth-order approximation in time of the homogeneous wave equation.

Now, the natural way to approximate the biharmonic operator Δ^2 would be to use a fourth-order approximation in space. However, such an approximation would be very expensive since its stencil⁷ contains 17 points in 2D (instead of 9 points) and 37 points in 3D (instead of 13 points). Fortunately, the presence of Δt^2 before Δ^2 implies that a second-order approximation of this operator provides a fourth-order approximation of (4.39) since

$$h^2 \Delta t^2 \leq \frac{1}{2} (h^4 + \Delta t^4). \quad (4.40)$$

Therefore, convergence in $h^2 \Delta t^2$ implies a convergence in $h^4 + \Delta t^4$.

So, the appropriate approximations of Δ^2 is

– in 1D:

$$\left(\frac{\partial^4 u_h}{\partial x^4} \right)_{h,\ell}^{(2)} = \frac{1}{h^4} (6u_\ell - 4(u_{\ell+1} + u_{\ell-1}) + u_{\ell+2} + u_{\ell-2}), \quad (4.41)$$

⁷ We recall that the stencil of a discrete operator is the set of the terms occurring in its computation at a given point. For instance, $\Delta_h^{(2)}$ in 2D has a 5-point stencil.

– in 2D:

$$\begin{aligned}
 & \left((\Delta^2)_h^{(2)} u_h \right)_{\ell,m} = \\
 & \frac{1}{h^4} (20u_{\ell,m} - 8(u_{\ell+1,m} + u_{\ell-1,m} + u_{\ell,m+1} + u_{\ell,m-1}) \\
 & + u_{\ell+2,m} + u_{\ell-2,m} + u_{\ell,m+2} + u_{\ell,m-2} \\
 & + 2(u_{\ell+1,m+1} + u_{\ell-1,m+1} + u_{\ell+1,m-1} + u_{\ell-1,m-1})). \tag{4.42}
 \end{aligned}$$

Stencils of schemes obtained with this approach are given in Fig. 4.1.

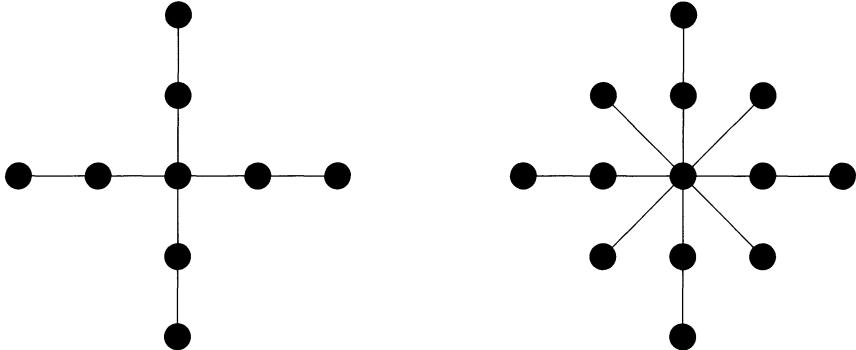


Fig. 4.1. Stencils of the schemes, second-order in time (left) and with the modified equation approach (right) with a fourth-order approximation in space (first approach)

Remarks

1. One can easily check that

$$(\Delta^2)_h^{(2)} = \left(\Delta_h^{(2)} \right)^2. \tag{4.43}$$

2. As we shall see later, Δt is always taken proportional to h , so that we actually obtain: $h^2 \Delta t^2 = 1/2 (h^4 + \Delta t^4)$.
3. $\Delta_h + c^2 \Delta t^2 / 12 (\Delta^2)_h$ is actually an approximation of Δ which tends to Δ with h and Δt . In particular, when Δt is proportional to h , this approximation tends to Δ with h .

4. Another possibility would be to use $(\Delta_h^{(4)})^2$ (square of the fourth-order discrete operator) instead of $(\Delta^2)_h^{(2)}$ (second-order approximation of the biharmonic operator) in (4.39). This point of view will be necessary and efficient for finite element methods but, in our case, it provides very dispersive⁸ schemes.
5. As we shall see in Chap. 6 and 7, this scheme has remarkable properties of stability and accuracy.
6. For the equation with a right-hand side

$$\frac{\partial^2 u}{\partial t^2} - c^2 \Delta u = f, \quad (4.44)$$

one can neglect f when it has a compact support in (4.37) without loss of accuracy and write the modified equation as

$$\frac{u_h^{n+1} - 2u_h^n + u_h^{n-1}}{\Delta t^2} - \frac{c^4 \Delta t^2}{12} (\Delta^2)_h u_h^n - c^2 \Delta_h u_h^n = f^n. \quad (4.45)$$

7. A general class of 25-point schemes, based on the modified equation approach and the first approach in space, was studied in [28]. As for second-order, this class provides a $\pi/4$ -rotated version of our scheme and a fourth-order version of Arakawa's scheme defined in (4.20).

4.4.2 Symmetric Schemes

Another alternative is provided by symmetric schemes whose general form is⁹

$$\frac{1}{\Delta t^2} \sum_{j=-p}^p a_{|j|} u_h^{n+j} = c^2 \Delta_h \left(\sum_{j=-q}^q b_{|j|} u_h^{n+j} \right). \quad (4.46)$$

Obviously, one must have $q < p$ to obtain explicit schemes.

In order to obtain fourth-order schemes, it is necessary to have $p \geq 2$ and $q \geq 1$. The minimal schemes are of course given by $p = 2$ and $q = 1$.

A convenient way to obtain the coefficients for a given order is to apply the scheme to the ODE $u_{tt} = u$ which has as a solution $u = \exp(t)$. This method leads to the study of the order in Δt of the expression [62]

⁸ The numerical dispersion will be defined in Chap. 7.

⁹ For the equation with a right-hand side (4.44) the scheme will be written as

$$\frac{1}{\Delta t^2} \sum_{j=-p}^p a_{|j|} u_h^{n+j} = c^2 \Delta_h \left(\sum_{j=-q}^q b_{|j|} u_h^{n+j} \right) + \sum_{j=-q}^q b_{|j|} f_h^{n+j}.$$

$$\sum_{j=-p}^p a_{|j|} e^{j\Delta t} - c^2 \Delta t^2 \sum_{j=-q}^q b_{|j|} e^{j\Delta t} = 0. \quad (4.47)$$

After taking into account the fourth-order character of the schemes, one obtains a one-parameter family of stable schemes given by

$$\begin{aligned} & \frac{1}{\Delta t^2} \left(-\frac{\theta+2}{6} u_h^{n+2} + \frac{1+2\theta}{3} u_h^{n+1} - \theta u_h^n \right. \\ & \left. + \frac{1+2\theta}{3} u_h^{n-1} - \frac{\theta+2}{6} u_h^{n-2} \right) = \\ & c^2 \Delta_h \left(\frac{5+2\theta}{12} u_h^{n+1} + \frac{1-2\theta}{6} u_h^n + \frac{5+2\theta}{12} u_h^{n-1} \right). \end{aligned} \quad (4.48)$$

Optimal accuracy and minimal stability is obtained for $\theta = 0$.

Of course, the main drawback of such methods is the additional storage needed. One must also carefully handle the initial time steps.

Remarks

- For $\theta = 0$, the scheme given in (4.48) can be written as the fourth-order staggered approximation of the system (4.33a) and (4.33b):

$$\frac{u_h^{n+2} - u_h^{n-1}}{3\Delta t} = c \operatorname{div}_h \mathbf{v}_h^{n+\frac{1}{2}}, \quad (4.49a)$$

$$\frac{\mathbf{v}_h^{n+\frac{3}{2}} - \mathbf{v}_h^{n+\frac{1}{2}}}{\Delta t} = c \frac{1}{12} \operatorname{grad}_h (5u_h^{n+2} + 2u_h^{n+1} + 5u_h^n). \quad (4.49b)$$

We shall see in Chap. 4 that such an approach can be very useful for finite element approximations¹⁰.

- Another (not staggered) symmetric approximation of (4.33a) and (4.33b) is provided by

$$\frac{1}{\Delta t} \sum_{j=-p}^p a_j u_h^{n+j} = c \operatorname{div}_h \left(\sum_{j=-p+1}^{p-1} b_j \mathbf{v}_h^{n+j} \right), \quad (4.50a)$$

¹⁰ More general schemes for (4.33a) and (4.33b) of the form

$$\frac{a_1 u_h^{n+2} + a_2 u_h^{n+1} - a_2 u_h^n - a_1 u_h^{n-1}}{\Delta t} = c \operatorname{div}_h (a_3 \mathbf{v}_h^{n+\frac{3}{2}} + a_4 \mathbf{v}_h^{n+\frac{1}{2}} + a_3 \mathbf{v}_h^{n-\frac{1}{2}}),$$

$$\frac{a_5 \mathbf{v}_h^{n+\frac{5}{2}} + a_6 \mathbf{v}_h^{n+\frac{3}{2}} - a_6 \mathbf{v}_h^{n+\frac{1}{2}} - a_5 \mathbf{v}_h^{n+\frac{5}{2}}}{\Delta t} = c \operatorname{grad}_h (a_7 u_h^{n+2} + a_8 u_h^{n+1} + a_6 u_h^n),$$

do not have good stability properties and also need much more storage than the scheme given in (4.49a) and (4.49b)

$$\frac{1}{\Delta t} \sum_{j=-p}^p \hat{a}_j \mathbf{v}_h^{n+j} = c \mathbf{grad}_h \left(\sum_{j=-p}^p \hat{b}_j u_h^{n+j} \right), \quad (4.50b)$$

where $a_{-k} = -a_k$, $\hat{a}_{-k} = -\hat{a}_k$, $b_{-k} = b_k$ and $\hat{b}_{-k} = \hat{b}_k$ for $-p \leq k \leq p$. For symmetry reasons, we have $a_0 = \hat{a}_0 = 0$.

A fourth-order scheme in time is obtained for $p = 2$, $a_2 = \hat{a}_2 = 1$, $a_1 = -2 \cos \theta$, $\hat{a}_2 = -2 \cos \hat{\theta}$ and

$$b_1 = 8/3 - 2/3 \cos \theta - 4b_2, \quad (4.51a)$$

$$b_0 = -4/3 - 8/3 \cos \theta + 6b_2, \quad (4.51b)$$

$$\hat{b}_1 = 8/3 - 2/3 \cos \hat{\theta} - 4\hat{b}_2, \quad (4.51c)$$

$$\hat{b}_0 = -4/3 - 8/3 \cos \hat{\theta} + 6\hat{b}_2. \quad (4.51d)$$

The values $b_2 = 0$, $\theta = \pi/4$, $\hat{\theta} = \pi/2$ and $\hat{b}_2 = 0.376$ of the parameters seem to be quasi-optimal¹¹.

4.5 Higher-Order Approximations in Space

Higher-order approximations of $\partial^2/\partial x^2$ can be written in the form

$$\left(\frac{\partial^2 u}{\partial x^2} \right)_{h,\ell}^{(2k)} = \frac{1}{h^2} \sum_{i=-p}^p \alpha_{|i|} u_{\ell+i}, \quad (4.52)$$

where $p = k$ for the first approach and $p = 2k - 1$ for the second one.

4.5.1 First Approach

In the first approach, (4.52) can be written as the following sum of second-order approximations:

$$\left(\frac{\partial^2}{\partial x^2} \right)_h^{(2k)} = \sum_{j=1}^k \lambda_j \left(\frac{\partial^2}{\partial x^2} \right)_{jh}^{(2)}, \quad (4.53)$$

where

$$\left(\frac{\partial^2}{\partial x^2} \right)_{jh,\ell}^{(2)} = \frac{u_{\ell+j} - 2u_\ell + u_{\ell-j}}{(jh)^2}. \quad (4.54)$$

¹¹ This kind of approximation was given to me by P. Chartier and J. Erhel from Irisa.

Since

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{h,\ell}^{(2)} = 2 \sum_{i=0}^p \frac{h^{2i}}{(2i+2)!} \frac{\partial^{2i+2} u_\ell}{\partial x^{2i+2}} + O(h^{2p+2}), \quad (4.55)$$

we obtain, after replacing h by jh in (4.55) and inserting the results into (4.53):

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{h,\ell}^{(2k)} = 2 \sum_{i=0}^k \left(\frac{h^{2i}}{(2i+2)!} \frac{\partial^{2i+2} u_\ell}{\partial x^{2i+2}} \sum_{j=1}^{k+1} \lambda_j j^{2i} \right) + O(h^{2k+2}). \quad (4.56)$$

So, in order to obtain the coefficients of $(\partial^2/\partial x^2)_h^{(2k)}$, one must solve the system

$$\begin{aligned} \lambda_1 + \lambda_2 + \dots + \lambda_{k+1} &= 1, \\ \lambda_1 + 2^2 \lambda_2 + \dots + (k+1)^2 \lambda_{k+1} &= 0, \\ \lambda_1 + 2^4 \lambda_2 + \dots + (k+1)^4 \lambda_{k+1} &= 0, \\ &\dots, \\ \lambda_1 + 2^{2k} \lambda_2 + \dots + (k+1)^{2k} \lambda_{k+1} &= 0. \end{aligned} \quad (4.57)$$

By setting $\alpha_j = j^2$, $j = 1, k+1$, one can see that the matrix of this system is a Vandermonde matrix. So Cramer's rule gives the solution:

$$\lambda_j = \frac{\prod_{1 \leq \ell < m \leq k+1} (\beta_\ell - \beta_m)}{\prod_{1 \leq \ell < m \leq k+1} (\alpha_\ell - \alpha_m)} = \frac{\prod_{\ell=1, \ell \neq j}^{k+1} \alpha_\ell}{\prod_{\ell=1, \ell \neq j}^{k+1} (\alpha_\ell - \alpha_j)}, \quad (4.58)$$

where $\beta_j = 0$ and $\beta_\ell = \alpha_\ell$ and $\beta_m = \alpha_m$ for $\ell \neq j$ and, on the other hand, $m \neq j$.

For a $2k$ th-order approximation, the relations between α_i and λ_i are obviously:

$$\left\{ \begin{array}{l} \alpha_0 = -2 \sum_{i=1}^k \frac{\lambda_j}{j^2}, \\ \alpha_j = \frac{\lambda_j}{j^2} \quad \forall j \text{ such that } 1 \leq j \leq k. \end{array} \right. \quad (4.59)$$

In Table 4.1, we give the coefficients λ_i obtained by this method for different approximations. One can see that these coefficients are alternately positive and negative for all the orders. This will be troublesome for the proof of the positivity of such approximations.

Table 4.1. The coefficients λ_j for approximations from 2nd-order to 14th-order

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7
$O(h^2)$	1						
$O(h^4)$	$\frac{4}{3}$	$-\frac{1}{3}$					
$O(h^6)$	$\frac{3}{2}$	$-\frac{3}{5}$	$\frac{1}{10}$				
$O(h^8)$	$\frac{8}{5}$	$-\frac{4}{5}$	$\frac{8}{35}$	$-\frac{1}{35}$			
$O(h^{10})$	$\frac{5}{3}$	$-\frac{20}{21}$	$\frac{5}{14}$	$-\frac{5}{63}$	$\frac{1}{126}$		
$O(h^{12})$	$\frac{12}{7}$	$-\frac{15}{14}$	$\frac{10}{21}$	$-\frac{1}{7}$	$\frac{2}{77}$	$-\frac{1}{462}$	
$O(h^{14})$	$\frac{7}{4}$	$-\frac{7}{6}$	$\frac{7}{12}$	$-\frac{7}{33}$	$\frac{7}{132}$	$-\frac{7}{858}$	$\frac{1}{1716}$

4.5.2 Second Approach

As for 2nd and 4th-order approximations, a $2k$ th-order approximation of $\frac{\partial^2}{\partial x^2}$ is, for the second approach, of the form:

$$\left(\frac{\partial^2}{\partial x^2} \right)_h^{(2k')} = \tilde{D}_{1,h}^{(2k)} \circ D_{1,h}^{(2k)}, \quad (4.60)$$

where

$$\left(D_{1,h}^{(2k)} u \right)_{\ell+\frac{1}{2}} = \frac{1}{h} \sum_{i=1}^k \alpha'_i (u_{\ell+i} - u_{\ell-i+1}), \quad (4.61a)$$

$$\left(\tilde{D}_{1,h}^{(2k)} u \right)_\ell = \frac{1}{h} \sum_{i=1}^k \alpha'_i (u_{\ell+i-\frac{1}{2}} - u_{\ell-i+\frac{1}{2}}). \quad (4.61b)$$

Similarly to the first approach, we can write:

$$D_{1,h}^{(2k)} = \sum_{j=1}^k \lambda'_j D_{1,jh}^{(2)} \quad \text{and} \quad \tilde{D}_{1,h}^{(2k)} = \sum_{j=1}^k \lambda'_j \tilde{D}_{1,jh}^{(2)}, \quad (4.62)$$

where

$$\left(D_{1,jh}^{(2)} u\right)_{\ell+\frac{1}{2}} = \frac{u_{\ell+j} - u_{\ell-j+1}}{(2j-1)h}, \quad (4.63a)$$

$$\left(\tilde{D}_{1,jh}^{(2k)} u\right)_\ell = \frac{u_{\ell+j-\frac{1}{2}} - u_{\ell-j+\frac{1}{2}}}{(2j-1)h}. \quad (4.63b)$$

The Taylor expansion of $\tilde{D}_{1,h}^{(2)}$ at $x = \ell h$ is

$$\left(\tilde{D}_{1,h}^{(2)} u\right)_\ell = \sum_{i=0}^p \frac{h^{2i}}{(2i+1)!2^{2i}} \frac{\partial^{2i+1} u_\ell}{\partial x^{2i+1}} + O(h^{2p+2}). \quad (4.64)$$

The Taylor expansion of $D_{1,h}^{(2)}$ is deduced from (4.64) by replacing ℓ by $\ell + \frac{1}{2}$. By replacing $h/2$ by $(2j-1)h/2$ in (4.64) and by inserting the different Taylor expansions obtained by this process into (4.62), we obtain:

$$\left(\tilde{D}_{1,h}^{(2)} u\right)_\ell = \sum_{i=0}^k \left(\frac{h^{2i}}{(2i+1)!2^{2i}} \frac{\partial^{2i+1} u_\ell}{\partial x^{2i+1}} \sum_{j=1}^{k+1} \lambda'_j (2j-1)^{2i} \right) + O(h^{2k+2}), \quad (4.65)$$

which leads to the following system in λ' :

$$\begin{aligned} \lambda'_1 + \lambda'_2 + \dots + \lambda'_{k+1} &= 1, \\ \lambda'_1 + 3^2 \lambda'_2 + \dots + (2k+1)^2 \lambda'_{k+1} &= 0, \\ \lambda'_1 + 3^4 \lambda'_2 + \dots + (2k+1)^4 \lambda'_{k+1} &= 0, \\ &\dots, \\ \lambda'_1 + 3^{2k} \lambda'_2 + \dots + (2k+1)^{2k} \lambda'_{k+1} &= 0. \end{aligned} \quad (4.66)$$

By setting $\alpha_j = (2j-1)^2$, $j = 1..k+1$, one can see that the matrix of this system is also a Vandermonde matrix. The solutions of this system for k from 0 to 6 are given in Table 4.2.

In this case, we have:

$$\alpha'_j = \frac{\lambda'_j}{2j-1} \quad \forall j \text{ such that } 1 \leq j \leq k. \quad (4.67)$$

Table 4.2. The coefficients λ'_j for approximations from 2nd-order to 12th-order

	λ'_1	λ'_2	λ'_3	λ'_4	λ'_5	λ'_6
$O(h^2)$	1					
$O(h^4)$	$\frac{9}{8}$	$-\frac{1}{8}$				
$O(h^6)$	$\frac{75}{64}$	$-\frac{25}{128}$	$\frac{3}{128}$			
$O(h^8)$	$\frac{1225}{1024}$	$-\frac{245}{1024}$	$\frac{49}{1024}$	$-\frac{5}{1024}$		
$O(h^{10})$	$\frac{19\,845}{16\,384}$	$-\frac{2205}{8192}$	$\frac{567}{8192}$	$-\frac{405}{32\,768}$	$\frac{35}{32\,768}$	
$O(h^{12})$	$\frac{160\,083}{131\,072}$	$-\frac{38\,115}{131\,072}$	$\frac{22\,869}{262\,144}$	$-\frac{5445}{262\,144}$	$\frac{847}{262\,144}$	$-\frac{63}{262\,144}$

4.5.3 Extension to Higher Dimensions

Since $\Delta = \sum_{i=1}^d \frac{\partial^2}{\partial x_i^2}$ in dimension d , the higher-order approximations of Δ in 2D can be written as

$$\Delta_h^{(2k)} u_{\ell,m} = \frac{1}{h^2} \sum_{i=-p}^p \alpha_{|i|} u_{\ell+i,m} + \frac{1}{h^2} \sum_{i=-p}^p \alpha_{|i|} u_{\ell,m+i} \quad (4.68)$$

and, similarly, in 3D:

$$\begin{aligned} \Delta_h^{(2k)} u_{\ell,m,n} &= \frac{1}{h^2} \sum_{i=-p}^p \alpha_{|i|} u_{\ell+i,m,n} \\ &+ \frac{1}{h^2} \sum_{i=-p}^p \alpha_{|i|} u_{\ell,m+i,n} \\ &+ \frac{1}{h^2} \sum_{i=-p}^p \alpha_{|i|} u_{\ell,m,n+i}. \end{aligned} \quad (4.69)$$

This implies that the coefficients of u in each direction are the same as the 1D coefficients except for $i = 0$ for which it is equal to $N\alpha_0$.

For the second approach we have

$$\begin{aligned} \left(D_{2,h}^{(2k)} u \right)_{\ell+\frac{1}{2}, m+\frac{1}{2}} &= \left(\frac{1}{h} \sum_{i=1}^k \alpha'_i (u_{\ell+i,m} - u_{\ell-i+1,m}), \right. \\ &\quad \left. \frac{1}{h} \sum_{i=1}^k \alpha'_i (u_{\ell,m+i} - u_{\ell,m-i+1}) \right), \end{aligned} \quad (4.70)$$

$$\begin{aligned} \left(\tilde{D}_{2,h}^{(2k)} u \right)_{\ell,m} &= \frac{1}{h} \sum_{i=1}^k \alpha'_i (u_{\ell+i-\frac{1}{2},m} - u_{\ell-i+\frac{1}{2},m}) \\ &\quad + \frac{1}{h} \sum_{i=1}^k \alpha'_i (u_{\ell,m+i-\frac{1}{2}} - u_{\ell,m-i+\frac{1}{2}}) \end{aligned} \quad (4.71)$$

and, in 3D:

$$\begin{aligned} \left(D_{3,h}^{(2k)} u \right)_{\ell+\frac{1}{2}, m+\frac{1}{2}, n+\frac{1}{2}} &= \left(\frac{1}{h} \sum_{i=1}^k \alpha'_i (u_{\ell+i,m,n} - u_{\ell-i+1,m,n}), \right. \\ &\quad \left. \frac{1}{h} \sum_{i=1}^k \alpha'_i (u_{\ell,m+i,n} - u_{\ell,m-i+1,n}), \right. \\ &\quad \left. \frac{1}{h} \sum_{i=1}^k \alpha'_i (u_{\ell,m,n+i} - u_{\ell,m,n-i+1}) \right), \end{aligned} \quad (4.72)$$

$$\begin{aligned} \left(\tilde{D}_{3,h}^{(2k)} u \right)_{\ell,m,n} &= \frac{1}{h} \sum_{i=1}^k \alpha'_i (u_{\ell+i-\frac{1}{2},m,n} - u_{\ell-i+\frac{1}{2},m,n}) \\ &\quad + \frac{1}{h} \sum_{i=1}^k \alpha'_i (u_{\ell,m+i-\frac{1}{2},n} - u_{\ell,m-i+\frac{1}{2},n}) \\ &\quad + \frac{1}{h} \sum_{i=1}^k \alpha'_i (u_{\ell,m,n+i-\frac{1}{2}} - u_{\ell,m,n-i+\frac{1}{2}}). \end{aligned} \quad (4.73)$$

Here, the coefficients are exactly the 1D coefficients in each direction.

Remarks

1. The second approach leads to much larger stencils than the first one. In any dimension, the ratio between the numbers of points involved in the stencils for the two approaches is $(4k - 1)/(2k + 1)$.

2. One can find other approaches to obtain highly accurate schemes in [69] and [117]. In particular, Holberg's approach [69] was very popular among geophysicists when it appeared. This approach is based on the computation of the coefficients of the approximation by a minimization of a cost function derived from the error committed on the velocity (numerical dispersion). This, which leads to very large stencils, is a high-accurate rather than high-order approach.

4.6 Higher-Order Approximations in Time

4.6.1 The Modified Equation Approach

A general Taylor expansion of the leapfrog approximation of $\partial^2 u / \partial t^2$ is:

$$\frac{u^{n+1} - 2u^n + u^{n-1}}{\Delta t^2} = 2 \sum_{j=1}^k \frac{\Delta t^{2j-2}}{(2j)!} \left(\frac{\partial^{2j} u}{\partial t^{2j}} \right)^n + O(\Delta t^{2k}). \quad (4.74)$$

By iterating the process described in (4.36)–(4.39), we obtain the $2k$ th-order modified equation:

$$\frac{u^{n+1} - 2u^n + u^{n-1}}{\Delta t^2} - c^2 \Delta_h u^n - 2 \sum_{j=2}^k \frac{c^{2j} \Delta t^{2j-2}}{(2j)!} (\Delta^j)_h u^n = 0. \quad (4.75)$$

If we assume that $\Delta t = \alpha h$ (which is natural), we see that, in order to obtain a global $2k$ -th-order of approximation, we must approximate $(\Delta^j)_h$ to the $2(k-j+1)$ th-order.

In 1D, these approximations all have a $(2k+1)$ -point stencil and are all obtained by solving a $k+1$ linear systems with the same Vandermonde-like matrix and different right-hand sides. One can even prove that the the stability condition (CFL¹²) condition for all these approximations is equal to 1 [6].

In higher dimensions, the construction of such approximations is more difficult since it involves multidimensional Taylor expansions. Moreover, such approximations are not unique even when their stencils are a subset of a $(2k+1)^d$ stencil. However, one can find a “minimal” stencil for each operator $(\Delta^j)_h$ for which we can conjecture that the CFL remains constant for any order of approximation.

For instance, we give below the form of the minimal scheme, sixth-order in time and space, for this approach in 2D. The general formula of this 29-point scheme can be written as

¹² These letters are the initials of Courant Friedrichs Levy.

$$\begin{aligned}
& \frac{u^{n+1} - 2u^n + u^{n-1}}{\Delta t^2} \\
&= Au_{\ell,m}^n + a(u_{\ell+1,m}^n + u_{\ell-1,m}^n + u_{\ell,m+1}^n + u_{\ell,m-1}^n) \\
&\quad + b(u_{\ell+2,m}^n + u_{\ell-2,m}^n + u_{\ell,m+2}^n + u_{\ell,m-2}^n) \\
&\quad + c(u_{\ell+3,m}^n + u_{\ell-3,m}^n + u_{\ell,m+3}^n + u_{\ell,m-3}^n) \\
&\quad + d(u_{\ell+1,m+1}^n + u_{\ell-1,m+1}^n + u_{\ell+1,m-1}^n + u_{\ell-1,m-1}^n) \\
&\quad + e(u_{\ell+2,m+2}^n + u_{\ell-2,m+2}^n + u_{\ell+2,m-2}^n + u_{\ell-2,m-2}^n),
\end{aligned} \tag{4.76}$$

where

$$\begin{aligned}
A &= -\frac{49}{9h^2} + \frac{175\Delta t^2}{72h^4} - \frac{11\Delta t^4}{72h^6}, \quad a = \frac{3}{2h^2} - \frac{71\Delta t^2}{72h^4} + \frac{23\Delta t^4}{360h^6}, \\
b &= -\frac{3}{20h^2} + \frac{25\Delta t^2}{144h^4} - \frac{13\Delta t^4}{720h^6}, \quad c = \frac{1}{90h^2} - \frac{\Delta t^2}{72h^4} + \frac{\Delta t^4}{360h^6}, \\
d &= \frac{2\Delta t^2}{9h^4} - \frac{\Delta t^4}{90h^6}, \quad e = -\frac{\Delta t^2}{288h^4} + \frac{\Delta t^4}{1440h^6}.
\end{aligned} \tag{4.77}$$

Remark

In 1D, the modified equation approach provides a characteristics scheme for any order when used with the maximal CFL [6, 35].

4.6.2 Symmetric Schemes

For $\theta = 0$, the $2k$ th-order version of the symmetric scheme described in (4.48) can be written as

$$\frac{u_h^{n+k} - u_h^{n+k-1} + u_h^{n-k+1} - u_h^{n-k}}{(2k-1)\Delta t^2} = \Delta_h \left(\sum_{j=-k+1}^{k-1} a_{|i|} u_h^{n+j} \right). \tag{4.78}$$

Unfortunately, although stable, these schemes have very small CFL. For instance, the CFL of the sixth-order scheme is almost six times more restrictive than that of the leapfrog scheme (against twice for the fourth-order) and about ten times for the eighth-order, which implies the use of very small time-steps. This remark combined with the additional storage needed by this approach make it unefficient.

Spectral methods, which are very difficult to implement, can be found in [111].

4.7 Extension to Systems

4.7.1 The Maxwell Equations

The undisputed star method for solving the Maxwell equations is the Yee scheme, introduced in 1966 [120] and still alive today. One of the main reasons for its longevity lies in its simplicity of implementation. The Yee scheme is based on the concept of centered approximation and staggered grids. These two basic ingredients make it second-order, as we shall see later.

So, in order to present this scheme, let us write down an explicit form of the Maxwell equation in 3D with constant scalar coefficients ε_0 , μ_0 :

$$\varepsilon_0 \frac{\partial E_1}{\partial t} + \frac{\partial H_2}{\partial x_3} - \frac{\partial H_3}{\partial x_2} = -J_1, \quad (4.79a)$$

$$\varepsilon_0 \frac{\partial E_2}{\partial t} + \frac{\partial H_3}{\partial x_1} - \frac{\partial H_1}{\partial x_3} = -J_2, \quad (4.79b)$$

$$\varepsilon_0 \frac{\partial E_3}{\partial t} + \frac{\partial H_1}{\partial x_2} - \frac{\partial H_2}{\partial x_1} = -J_3, \quad (4.79c)$$

$$\mu_0 \frac{\partial H_1}{\partial t} - \frac{\partial E_2}{\partial x_3} + \frac{\partial E_3}{\partial x_2} = 0, \quad (4.79d)$$

$$\mu_0 \frac{\partial H_2}{\partial t} - \frac{\partial E_3}{\partial x_1} + \frac{\partial E_1}{\partial x_3} = 0, \quad (4.79e)$$

$$\mu_0 \frac{\partial H_3}{\partial t} - \frac{\partial E_1}{\partial x_2} + \frac{\partial E_2}{\partial x_1} = 0. \quad (4.79f)$$

The principle of the Yee scheme is to write each first-order derivative in space in a centered way versus the corresponding derivative in time of the equation and vice-versa. This technique provides the following scheme:

$$\begin{aligned} & \varepsilon_0 \frac{E_{1p+\frac{1}{2},q,r}^{n+1} - E_{1p+\frac{1}{2},q,r}^n}{\Delta t} + \frac{H_{2p+\frac{1}{2},q,r+\frac{1}{2}}^{n+\frac{1}{2}} - H_{2p+\frac{1}{2},q,r-\frac{1}{2}}^{n+\frac{1}{2}}}{h} \\ & - \frac{H_{3p+\frac{1}{2},q+\frac{1}{2},r}^{n+\frac{1}{2}} - H_{3p+\frac{1}{2},q-\frac{1}{2},r}^{n+\frac{1}{2}}}{h} = -J_{1p+\frac{1}{2},q,r}^{n+\frac{1}{2}}, \end{aligned} \quad (4.80a)$$

$$\begin{aligned} & \varepsilon_0 \frac{E_{2p,q+\frac{1}{2},r}^{n+1} - E_{2p,q+\frac{1}{2},r}^n}{\Delta t} + \frac{H_{3p+\frac{1}{2},q+\frac{1}{2},r}^{n+\frac{1}{2}} - H_{3p-\frac{1}{2},q+\frac{1}{2},r}^{n+\frac{1}{2}}}{h} \\ & - \frac{H_{1p,q+\frac{1}{2},r+\frac{1}{2}}^{n+\frac{1}{2}} - H_{1p,q+\frac{1}{2},r-\frac{1}{2}}^{n+\frac{1}{2}}}{h} = -J_{2p,q+\frac{1}{2},r}^{n+\frac{1}{2}}, \end{aligned} \quad (4.80b)$$

$$\varepsilon_0 \frac{E_{3p,q,r+\frac{1}{2}}^{n+1} - E_{3p,q,r+\frac{1}{2}}^n}{\Delta t} + \frac{H_{1p,q+\frac{1}{2},r+\frac{1}{2}}^{n+\frac{1}{2}} - H_{1p,q-\frac{1}{2},r+\frac{1}{2}}^{n+\frac{1}{2}}}{h} \\ - \frac{H_{2p+\frac{1}{2},q,r+\frac{1}{2}}^{n+\frac{1}{2}} - H_{2p-\frac{1}{2},q,r+\frac{1}{2}}^{n+\frac{1}{2}}}{h} = -J_{3p,q,r+\frac{1}{2}}^{n+\frac{1}{2}}, \quad (4.80c)$$

$$\mu_0 \frac{H_{1p,q+\frac{1}{2},r+\frac{1}{2}}^{n+\frac{1}{2}} - H_{1p,q+\frac{1}{2},r+\frac{1}{2}}^{n-\frac{1}{2}}}{\Delta t} - \frac{E_{2p,q+\frac{1}{2},r+1}^n - E_{2p,q+\frac{1}{2},r}^n}{h} \\ + \frac{E_{3p,q+1,r+\frac{1}{2}}^n - E_{3p,q,r+\frac{1}{2}}^n}{h} = 0, \quad (4.80d)$$

$$\mu_0 \frac{H_{2p+\frac{1}{2},q,r+\frac{1}{2}}^{n+\frac{1}{2}} - H_{2p+\frac{1}{2},q,r+\frac{1}{2}}^{n-\frac{1}{2}}}{\Delta t} - \frac{E_{3p+1,q,r+\frac{1}{2}}^n - E_{3p,q,r+\frac{1}{2}}^n}{h} \\ + \frac{E_{1p+\frac{1}{2},q,r+1}^n - E_{1p+\frac{1}{2},q,r}^n}{h} = 0, \quad (4.80e)$$

$$\mu_0 \frac{H_{3p+\frac{1}{2},q+\frac{1}{2},r}^{n+\frac{1}{2}} - H_{3p+\frac{1}{2},q+\frac{1}{2},r}^{n-\frac{1}{2}}}{\Delta t} - \frac{E_{1p+\frac{1}{2},q+1,r}^n - E_{1p+\frac{1}{2},q,r}^n}{h} \\ + \frac{E_{2p+1,q+\frac{1}{2},r}^n - E_{2p,q+\frac{1}{2},r}^n}{h} = 0. \quad (4.80f)$$

The 2D version of this scheme, written for the TM case for instance, is

$$\varepsilon_0 \frac{E_{1p+\frac{1}{2},q}^{n+1} - E_{1p+\frac{1}{2},q}^n}{\Delta t} - \frac{H_{p+\frac{1}{2},q+\frac{1}{2}}^{n+\frac{1}{2}} - H_{p+\frac{1}{2},q-\frac{1}{2}}^{n+\frac{1}{2}}}{h} = J_{1p+\frac{1}{2},q}^{n+\frac{1}{2}}, \quad (4.81a)$$

$$\varepsilon_0 \frac{E_{2p,q+\frac{1}{2}}^{n+1} - E_{2p,q+\frac{1}{2}}^n}{\Delta t} + \frac{H_{p+\frac{1}{2},q+\frac{1}{2}}^{n+\frac{1}{2}} - H_{p-\frac{1}{2},q+\frac{1}{2}}^{n+\frac{1}{2}}}{h} = J_{2p,q+\frac{1}{2}}^{n+\frac{1}{2}}, \quad (4.81b)$$

$$\mu_0 \frac{H_{p+\frac{1}{2},q+\frac{1}{2}}^{n+\frac{1}{2}} - H_{p+\frac{1}{2},q+\frac{1}{2}}^{n-\frac{1}{2}}}{\Delta t} - \frac{E_{1p+\frac{1}{2},q+1}^n - E_{1p+\frac{1}{2},q}^n}{h} \\ + \frac{E_{2p+1,q+\frac{1}{2}}^n - E_{2p,q+\frac{1}{2}}^n}{h} = 0. \quad (4.81c)$$

As one can see in the above equations, the discrete values of the electric and magnetic fields are not given at the same points. This is the basis of staggered

grid schemes which enable one to center all the equations of the scheme. The locations of the variables in 2D and 3D are given in Fig. 4.2.

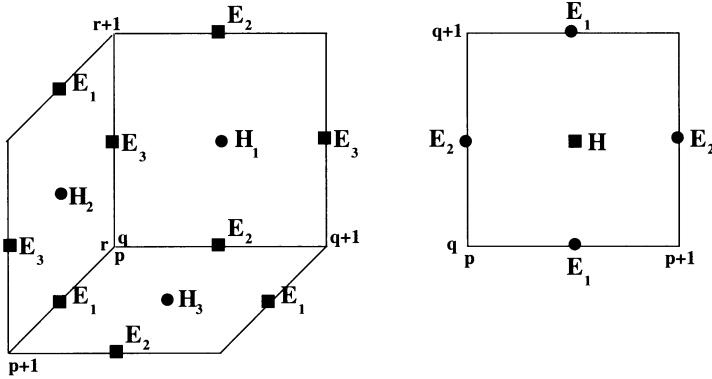


Fig. 4.2. Locations of the fields \mathbf{E} and \mathbf{H} in 3D (*left*) and 2D (*right*)

The extension of such a scheme to higher-orders is quite easy since the equations involve first-order derivatives in one direction. One must simply replace the second-order approximations by higher-order formulae based on Table 4.2.

The fourth-order scheme was studied by Devèze in his thesis [47] and gave quite good results as long as discontinuities in the coefficients were not too large.

Remarks

- When $\mathbf{J} = \mathbf{J}_1 + \sigma\mathbf{E}$, the additional term in \mathbf{E} must also be approximated in a centered way. For a second order approximation in time, we add the terms $(\sigma/2)(E_{1p+\frac{1}{2},q,r}^{n+1} + E_{1p+\frac{1}{2},q,r}^n)$, $(\sigma/2)(E_{2p,q+\frac{1}{2},r}^{n+1} + E_{2p,q+\frac{1}{2},r}^n)$ and $(\sigma/2)(E_{1p,q,r+\frac{1}{2}}^{n+1} + E_{1p,q,r+\frac{1}{2}}^n)$ to the three equations of (4.80a)–(4.80f). Higher-order time differencing requires higher-order centered approximations. The case of the modified equation approach, which we shall discuss below, is more difficult to deal with.
- The treatment of the anisotropic case (with non-diagonal matrices) by using the Yee scheme is not obvious and will be discussed in Chap. 13.

4.7.2 The Elastics System

The approximation equivalent to the Yee scheme for the elastics system was discovered, in the geophysics world, by Madariaga [82] and Virieux [119]. Of

course, as for the Maxwell equations, this scheme was written in the isotropic case described by (1.28) in 2D. Its extension to fourth-order was given by Levander [80] but had a rather short career.

The approximation that we define here will be that of the second-order system in 2D defined in (1.35a) and (1.35b) (in order to simplify the notations we shall replace the variables (v_1, v_2) used in (1.35a) and (1.35b) by (v, w)) in its homogeneous form:

$$\rho \frac{\partial^2 v}{\partial t^2} = (\lambda + 2\mu) \frac{\partial^2 v}{\partial x^2} + \mu \frac{\partial^2 v}{\partial y^2} + (\lambda + \mu) \frac{\partial^2 w}{\partial x \partial y}, \quad (4.82a)$$

$$\rho \frac{\partial^2 w}{\partial t^2} = (\lambda + 2\mu) \frac{\partial^2 w}{\partial y^2} + \mu \frac{\partial^2 w}{\partial x^2} + (\lambda + \mu) \frac{\partial^2 u}{\partial x \partial y}. \quad (4.82b)$$

This system has two new features which justify its study:

- It contains a cross derivative,
- Two different velocities are involved in it.

So, as an example, we shall describe its second-order approximation which was widely studied, in [94] for instance.

Let us first define the approximation of the cross derivative. It is constructed on the basis of the one-dimensional centered approximation of the first derivative as follows:

$$\begin{aligned} \frac{\partial^2 u}{\partial x \partial y} &\simeq \frac{\frac{u_{\ell+1,m+1} - u_{\ell-1,m+1}}{2h} - \frac{u_{\ell+1,m-1} - u_{\ell-1,m-1}}{2h}}{2h} \\ &\simeq \frac{u_{\ell+1,m+1} - u_{\ell-1,m+1} - u_{\ell+1,m-1} + u_{\ell-1,m-1}}{4h^2}. \end{aligned} \quad (4.83)$$

So, by using (4.83) and the second-order approximations of the derivatives in x and y , we obtain:

$$\begin{aligned} \rho \frac{v_{\ell,m}^{n+1} - 2v_{\ell,m}^n + v_{\ell,m}^{n-1}}{\Delta t^2} &= \\ (\lambda + 2\mu) \frac{v_{\ell+1,m}^n - 2v_{\ell,m}^n + v_{\ell-1,m}^n}{h^2} + \mu \frac{v_{\ell,m+1}^n - 2v_{\ell,m}^n + v_{\ell,m-1}^n}{h^2} & \quad (4.84a) \\ + (\lambda + \mu) \frac{w_{\ell+1,m+1}^n - w_{\ell-1,m+1}^n - w_{\ell+1,m-1}^n + w_{\ell-1,m-1}^n}{4h^2}, \end{aligned}$$

$$\begin{aligned} \rho \frac{w_{\ell,m}^{n+1} - 2w_{\ell,m}^n + w_{\ell,m}^{n-1}}{\Delta t^2} = \\ \mu \frac{w_{\ell+1,m}^n - 2w_{\ell,m}^n + w_{\ell-1,m}^n}{h^2} + (\lambda + 2\mu) \frac{w_{\ell,m+1}^n - 2w_{\ell,m}^n + w_{\ell,m-1}^n}{h^2} \quad (4.84b) \\ + (\lambda + \mu) \frac{v_{\ell+1,m+1}^n - v_{\ell-1,m+1}^n - v_{\ell+1,m-1}^n + v_{\ell-1,m-1}^n}{4h^2}. \end{aligned}$$

A fourth-order extension of (1.35a) and (1.35b) was studied by Barbiéra and Cohen [14, 15], but led to very complicated schemes because of the presence of cross derivatives in the system. Moreover, the extension of these fourth-order schemes to heterogeneous media led to approximations which became unstable for large discontinuities and the treatment of free surfaces was very troublesome. Another approach, based on split operators can be found in [16]. Holberg's method [69], which we mentioned for acoustics, was actually applied to elastics.

As we shall see in Chap. 13, variational higher-order versions of the Yee-Madariaga-Virieux schemes can be constructed on the basis of mixed finite elements.

Remark

The Virieux scheme can in fact be reinterpreted as the approximation defined in (4.84a) and (4.84b).

4.7.3 Higher-Order Approximation in Time

Symmetric schemes for systems can be expressed as in (4.49a) and (4.49b) or (4.50a) and (4.50b).

The modified equation approach is however slightly different. We give below its construction at fourth-order for the following formal first order system

$$\frac{\partial \mathbf{u}}{\partial t} = \mathcal{A}\mathbf{v}, \quad (4.85a)$$

$$\frac{\partial \mathbf{v}}{\partial t} = \mathcal{B}\mathbf{u}, \quad (4.85b)$$

whose second-order approximation in time is

$$\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} = \mathcal{A}\mathbf{v}^{n+\frac{1}{2}}, \quad (4.86a)$$

$$\frac{\mathbf{v}^{n+\frac{1}{2}} - \mathbf{v}^{n-\frac{1}{2}}}{\Delta t} = \mathcal{B}\mathbf{u}^n. \quad (4.86b)$$

The Taylor expansion of the leapfrog scheme is

$$\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} = \left(\frac{\partial \mathbf{u}}{\partial t} \right)^n + \frac{\Delta t^2}{24} \left(\frac{\partial^3 \mathbf{u}}{\partial t^3} \right)^n + O(\Delta t^4). \quad (4.87)$$

By combining (4.85a) and (4.85b) and (4.87), we obtain:

$$\frac{\partial^3 \mathbf{u}}{\partial t^3} = \frac{\partial^2}{\partial t^2} \frac{\partial \mathbf{u}}{\partial t} = \frac{\partial^2}{\partial t^2} (\mathcal{A} \mathbf{v}) = \frac{\partial}{\partial t} (\mathcal{A} \frac{\partial \mathbf{v}}{\partial t}), \quad (4.88a)$$

$$= \frac{\partial}{\partial t} (\mathcal{A}(\mathcal{B} \mathbf{u})) = \mathcal{A}(\mathcal{B} \frac{\partial \mathbf{u}}{\partial t}) = \mathcal{A}(\mathcal{B}(\mathcal{A} \mathbf{v})). \quad (4.88b)$$

A similar process provides

$$\frac{\partial^3 \mathbf{v}}{\partial t^3} = \mathcal{B}(\mathcal{A}(\mathcal{B} \mathbf{u})). \quad (4.89)$$

Hence, the modified system is written as

$$\frac{\mathbf{u}_h^{n+1} - \mathbf{u}_h^n}{\Delta t} = \frac{\Delta t^2}{24} \mathcal{A}(\mathcal{B}(\mathcal{A} \mathbf{v}_h^{n+\frac{1}{2}})) + \mathcal{A} \mathbf{v}_h^{n+\frac{1}{2}}, \quad (4.90a)$$

$$\frac{\mathbf{v}_h^{n+\frac{1}{2}} - \mathbf{v}_h^{n-\frac{1}{2}}}{\Delta t} = \frac{\Delta t^2}{24} \mathcal{B}(\mathcal{A}(\mathcal{B} \mathbf{u}_h^n)) + \mathcal{B} \mathbf{u}_h^n. \quad (4.90b)$$

Higher-order versions can be obtained in the same way. Of course, as for the wave equation, higher order derivatives are approximated at lower orders for finite difference approximations.

Remark

As we shall see later, \mathcal{A} and \mathcal{B} can be partial differential or approximate operators.

5. The Dispersion Relation

As we showed in Chap. 3, the solution of the wave equation can be expressed as a sum of plane waves. This remark, which justifies the use of plane waves to analyze continuous models holds in the discrete case [116]. The purpose of this section is to show how to find dispersion relations corresponding to (3.6), (3.15) and (3.24) for the discrete equations defined in Chap. 4. As we shall see in the following chapters, this relation is an important source of information on the stability and accuracy properties of the numerical models.

5.1 Second-Order Schemes for the Wave Equation

We are going to compute the dispersion relations for the lowest order of approximation of the wave equation. This will first enable us to describe the method on an simple scheme and, as we shall see later, will provide the basic formulae for higher-order schemes.

5.1.1 Using Plane Wave Solutions

As in Sect. 4.2, we define a regular mesh of \mathbb{R} with a space-step h and we set $x_\ell = \ell h$. We are looking for a solution of

$$\frac{d^2 u_\ell}{dt^2} - \frac{c^2}{h^2} (u_{\ell+1} - 2u_\ell + u_{\ell-1}) = 0, \quad \forall \ell \in \mathbb{Z}, \quad (5.1)$$

such that

$$u_\ell = e^{i(\omega t - \ell kh)}, \quad k \geq 0. \quad (5.2)$$

By inserting (5.2) into (5.1), we obtain

$$-\omega^2 e^{i(\omega t - \ell kh)} = \frac{c^2}{h^2} \left(e^{i(\omega t - (\ell+1)kh)} - 2e^{i(\omega t - \ell kh)} + e^{i(\omega t - (\ell-1)kh)} \right). \quad (5.3)$$

After simplification by $e^{i(\omega t - \ell kh)}$, this leads to

$$\omega^2 = \frac{c^2}{h^2} (2 - e^{-ikh} - e^{ikh}) = \frac{2c^2}{h^2} (1 - \cos kh), \quad (5.4)$$

which finally shows that

$$\omega^2 = \frac{4c^2}{h^2} \sin^2 \frac{kh}{2} \quad (5.5)$$

is the dispersion relation of (5.1).

This relation, obtained for the scheme semi-discrete in space, can be extended to the fully discrete equation

$$\frac{u_\ell^{n+1} - 2u_\ell^n + u_\ell^{n-1}}{h^2} - c^2 \frac{u_{\ell+1}^n - 2u_\ell^n + u_{\ell-1}^n}{h^2} = 0 \quad \forall (\ell, n) \in \mathbb{Z} \times \mathbb{Z}. \quad (5.6)$$

In this case, the plane wave solution is written as

$$u_\ell^n = e^{i(n\omega\Delta t - \ell kh)}. \quad (5.7)$$

A similar computation gives

$$\frac{1}{\Delta t^2} (e^{-i\omega\Delta t} - 2 + e^{i\omega\Delta t}) - \frac{c^2}{h^2} (e^{-ikh} - 2 + e^{ikh}) = 0, \quad (5.8)$$

which can also be written as

$$\frac{2}{\Delta t^2} (1 - \cos \omega\Delta t) = \frac{2c^2}{h^2} (1 - \cos kh). \quad (5.9)$$

So, we obtain the dispersion relation:

$$\sin^2 \frac{\omega\Delta t}{2} = \frac{c^2\Delta t^2}{h^2} \sin^2 \frac{kh}{2}. \quad (5.10)$$

One can easily check that (5.10) tends to (5.5) when $\Delta t \rightarrow 0$ which itself tends to (3.6) when $h \rightarrow 0$.

5.1.2 Computation by the Discrete Fourier Transform

The same relations can be got by using the *discrete Fourier transform*.

Let us define the 1D discrete Fourier transforms in time and in space.

– In time:

$$\begin{cases} \mathcal{F}_t^h : V_0^1 \rightarrow L^2(\mathbb{R}) \text{ such that} \\ \mathcal{F}_t^h ((u^n)_{n \in \mathbb{Z}}) = \hat{u}(\omega) = \frac{1}{\sqrt{2\pi}} \sum_{n \in \mathbb{Z}} u^n e^{-in\omega\Delta t}. \end{cases} \quad (5.11)$$

– In space:

$$\begin{cases} \mathcal{F}_x^h : V_0^1 \rightarrow L^2(\mathbb{R}) \text{ such that} \\ \mathcal{F}_x^h ((u_\ell)_{\ell \in \mathbb{Z}}) = \hat{u}(\mathbf{k}) = \frac{1}{\sqrt{2\pi}} \sum_{\ell \in \mathbb{Z}} u_\ell e^{-i(\ell kh)}. \end{cases} \quad (5.12)$$

Let D_t denote the discrete operator from V_0^1 into itself defined by

$$(D_t u_h)^n = \frac{u_h^{n+1} - 2u_h^n + u_h^{n-1}}{\Delta t^2}. \quad (5.13)$$

So, the discrete equation can be written as

$$D_t u_h - c^2 \left(\frac{\partial^2 u_h}{\partial x^2} \right)_h^{(2)} = 0. \quad (5.14)$$

By applying the discrete Fourier transforms in time and space to (5.14), we obtain:

$$\widehat{D_t u_h} - c^2 \left(\widehat{\frac{\partial^2 u_h}{\partial x^2}} \right)_h^{(2)} = 0, \quad (5.15)$$

whose explicit form is

$$\begin{aligned} & \sum_{n \in \mathbb{Z}} \sum_{\ell \in \mathbb{Z}} \frac{u_\ell^{n+1} - 2u_\ell^n + u_\ell^{n-1}}{\Delta t^2} e^{-i(n\omega\Delta t + \ell kh)} \\ & - \sum_{n \in \mathbb{Z}} \sum_{\ell \in \mathbb{Z}} \frac{c^2}{h^2} (u_{\ell+1}^n - 2u_\ell^n + u_{\ell-1}^n) e^{-i(n\omega\Delta t + \ell kh)} = 0. \end{aligned} \quad (5.16)$$

By index translation, we obtain

$$\begin{aligned} & \frac{1}{\Delta t^2} \sum_{n \in \mathbb{Z}} \sum_{\ell \in \mathbb{Z}} u_\ell^n (e^{-i((n-1)\omega\Delta t + \ell kh)} - 2e^{-i(n\omega\Delta t + \ell kh)} \\ & \quad + e^{-i((n+1)\omega\Delta t + \ell kh)}) \\ & - \frac{c^2}{h^2} \sum_{n \in \mathbb{Z}} \sum_{\ell \in \mathbb{Z}} u_\ell^n (e^{-i(n\omega\Delta t + (\ell-1)kh)} - 2e^{-i(n\omega\Delta t + \ell kh)} \\ & \quad + e^{-i(n\omega\Delta t + (\ell+1)kh)}) = 0, \end{aligned} \quad (5.17)$$

which shows that, after factorizing,

$$\begin{aligned} & \frac{1}{\Delta t^2} \widehat{u}_h \left(e^{i\omega\Delta t} - 2 + e^{-i\omega\Delta t} \right) \\ & - \frac{c^2}{h^2} \widehat{u}_h \left(e^{ikh} - 2 + e^{-ikh} \right) = 0, \end{aligned} \quad (5.18)$$

where $\widehat{u}_h = \mathcal{F}_t^h (\mathcal{F}_x^h u_h)$.

After simplification by \widehat{u}_h and some computations, we obtain the dispersion relation (5.10).

5.1.3 Symbol of an Operator

The comparison of relations (5.18) and (5.15) shows that one can define the “Fourier transforms” \widehat{D}_t and $(\partial^2 u / \partial x^2)_h^{(2)}$ of the operators D_t and $(\partial^2 u / \partial x^2)_h^{(2)}$. These transforms are equal to $-4/\Delta t^2 \sin^2 \omega \Delta t/2$ and $-4/h^2 \sin^2 kh/2$ respectively. \widehat{D}_t and $\widehat{\Delta}_h^{(2)}$ are called the *symbols* of the corresponding operators. The definition of a symbol implies that, if D_1 and D_2 are two given operators, one can write

$$\widehat{D}_1 \circ \widehat{D}_2 = \widehat{D}_1 \widehat{D}_2. \quad (5.19)$$

Another interesting feature of the symbols comes from the following:

Let

$$\left(\frac{d^q u}{dx^q} \right)_h = \sum_{j=1}^p a_j u_{\ell+j} \quad (5.20)$$

be an approximation of $\frac{d^q u}{dx^q}$. If we look for a plane wave solution $u_0 = e^{i(\omega t - kx)}$ of (5.20), we obtain

$$\sum_{j=1}^p a_j e^{i(\omega t - (\ell+j)hk)} = e^{i(\omega t - \ell hk)} \sum_{j=1}^p a_j e^{ijkh} = u_0 D_h(k), \quad (5.21)$$

where $D_h(k)$ is the symbol of the discrete operator defined by (5.20).

Now, let

$$\left(\frac{\partial^q u}{\partial x_1^q} \right)_h = \sum_{j=1}^p a_j u_{\ell+j,m} \quad (5.22)$$

be an approximation of the partial derivative in \mathbb{R}^2 defined by $\frac{\partial^q u}{\partial x_1^q}$. In the same way, if we set $u = u_0 = e^{i(\omega t - k\ell h - kmh)}$ in (5.22), we obtain

$$\left\{ \begin{aligned} \sum_{j=1}^p a_j e^{i(\omega t - (\ell+j)hk_1 - mhk_2)} &= e^{i(\omega t - \ell hk_1 - mhk_2)} \sum_{j=1}^p a_j e^{ijkh k_1} \\ &= u_0 D_h(k_1). \end{aligned} \right. \quad (5.23)$$

So, we have the same symbol as for (5.20) where k is replaced by k_1 . Of course, this result can be extended to any dimension and for the derivatives in any variable of space.

This result implies that

$$\sum_{r=1}^d \widehat{\left(\frac{\partial^{q_r} u}{\partial x_r^{q_r}} \right)_h} = \sum_{r=1}^d D_h^r(k_r). \quad (5.24)$$

where $D_h^r(k)$ is the symbol of $\frac{d^{q_r} u}{dx^{q_r}}$.

In particular, the symbol of the second-order approximation of Δ is

– In 2D:

$$-\frac{4c^2}{h^2} (\sin^2 \frac{k_1 h}{2} + \sin^2 \frac{k_2 h}{2}), \quad (5.25)$$

– In 3D:

$$-\frac{4c^2}{h^2} (\sin^2 \frac{k_1 h}{2} + \sin^2 \frac{k_2 h}{2} + \sin^2 \frac{k_3 h}{2}). \quad (5.26)$$

Equations (5.25) and (5.26) provide the different dispersion relations in higher-order dimensions.

These properties will be very useful in the following.

Remark

By applying the continuous Fourier transforms to the differential operators, one can obtain their symbols in the same way. The basic symbol is that of a derivative which is i times the Fourier variable. For instance, the Fourier transform of $\partial/\partial t$ is $i\omega$. We give below the symbols of basic operators involved in the wave equations:

- $\widehat{\Delta} = -|\mathbf{k}|^2$,
- $\widehat{\operatorname{div}} = i \sum_{j=1}^d k_j$,
- $\widehat{\nabla} = i(k_1, \dots, k_d)^T$,
- $\widehat{\operatorname{curl}} = i(k_2 - k_3, k_3 - k_1, k_1 - k_2)^T$ (in 3D).

5.2 Higher-Order Approximations in Space

5.2.1 The First Approach

We showed in Chap. 4 that the finite difference approximation obtained by using the first approach can be written as

$$\Delta_h^{(4)} = \frac{4}{3} \Delta_h^{(2)} - \frac{1}{3} \Delta_{2h}^{(2)}. \quad (5.27)$$

This relation enables us to deduce from (5.5) and (5.25) the symbols of Δ in 1D and 2D which will lead to the dispersion relations for this approach.

The 1D Case. In this case, the dispersion relation comes immediately from (5.5)

$$\omega^2 = \frac{4}{3} \frac{4c^2}{h^2} \sin^2 \frac{kh}{2} - \frac{1}{3} \frac{c^2}{h^2} \sin^2 kh. \quad (5.28)$$

Since $\sin^2 kh = 1 - \cos^2 kh = 1 - \left(1 - 2 \sin^2 \frac{kh}{2}\right)^2$, we obtain, after some computations

$$\omega^2 = \frac{4c^2}{h^2} \left(\sin^2 \frac{kh}{2} + \frac{1}{3} \sin^4 \frac{kh}{2} \right). \quad (5.29)$$

The 2D Case. From (5.29) and relation (5.24), we can deduce the 2D dispersion relation for the semi-discrete approximation in space:

$$\omega^2 = \frac{4}{3} \frac{4c^2}{h^2} \left(\sin^2 \frac{k_1 h}{2} + \sin^2 \frac{k_2 h}{2} \right) - \frac{1}{3} \frac{c^2}{h^2} (\sin^2 k_1 h + \sin^2 k_2 h), \quad (5.30)$$

which can be written as

$$\omega^2 = \frac{4c^2}{h^2} \left(\sin^2 \frac{k_1 h}{2} + \sin^2 \frac{k_2 h}{2} + \frac{1}{3} \sin^4 \frac{k_1 h}{2} + \frac{1}{3} \sin^4 \frac{k_2 h}{2} \right). \quad (5.31)$$

Higher-Order Schemes. The symbols of higher-order approximations of Δ come from relations (4.53) and (5.24). From (4.53), we obtain in 1D:

$$\widehat{\left(\frac{\partial^2}{\partial x^2} \right)_h^{(2p)}} = - \sum_{j=1}^p \lambda_j \frac{4c^2}{j^2 h^2} \sin^2 \frac{jkh}{2}. \quad (5.32)$$

Now,

$$\begin{aligned} \sin^2 \frac{(j+1)kh}{2} &= \left(\sin \frac{jkh}{2} \cos \frac{kh}{2} + \sin \frac{kh}{2} \cos \frac{jkh}{2} \right)^2 \\ &= \sin^2 \frac{jkh}{2} \cos^2 \frac{kh}{2} + \sin^2 \frac{kh}{2} \cos^2 \frac{jkh}{2} \\ &\quad + 2 \sin \frac{jkh}{2} \cos \frac{kh}{2} \sin \frac{kh}{2} \cos \frac{jkh}{2}. \end{aligned} \quad (5.33)$$

On the other hand,

$$\begin{aligned}
& \sin \frac{jkh}{2} \cos \frac{kh}{2} \sin \frac{kh}{2} \cos \frac{jkh}{2} \\
&= \frac{1}{4} \left(\cos^2 \frac{(j-1)kh}{2} - \cos^2 \frac{(j+1)kh}{2} \right) \\
&= \frac{1}{4} \left(\sin^2 \frac{(j+1)kh}{2} - \sin^2 \frac{(j-1)kh}{2} \right),
\end{aligned} \tag{5.34}$$

so that, by setting $w_j = \sin^2 jkh/2$, we obtain the equation

$$w_{j+1} = 2w_j(1 - 2w_1) - w_{j-1} + 2w_1, \quad w_0 = 0, \quad w_1 = \sin^2 \frac{kh}{2}, \tag{5.35}$$

whose solution is

$$\begin{aligned}
w_j &= \frac{1}{2} - \frac{1}{4} \left(\frac{-1}{2w_1^2 - 1 + \sqrt{w_1^2(w_1^2 - 1)}} \right)^j \\
&\quad - \frac{1}{4} \left(\frac{1}{1 - 2w_1^2 + \sqrt{w_1^2(w_1^2 - 1)}} \right)^j.
\end{aligned} \tag{5.36}$$

Actually, by writing w_j in terms of complex exponentials, one can see that $w_j = \sin^2 jkh/2$, which proves that it belongs to \mathbb{R} .¹

So, by inserting (5.36) into (5.32) (with the help of Maple [98]), we obtain the expression of (5.32) in terms of powers of $\sin^2 kh/2$.

The computation of the symbols of successive higher-order schemes shows that we have the following remarkable property:

$$\widehat{\left(\frac{\partial^2}{\partial x^2} \right)_h^{(2p+2)}} - \widehat{\left(\frac{\partial^2}{\partial x^2} \right)_h^{(2p)}} = -a_{p+1} \sin^{2p+2} \frac{kh}{2}. \tag{5.37}$$

We give in Table 5.1 the values of a_p for $1 \leq p \leq 7$.

Table 5.1 shows that the higher-order approximations of $-\Delta$ by the first approach are positive, which was not obvious because of the negative coefficients involved in them. A general study of these higher-order symbols can be found in [6].

Of course, the symbols of higher-dimensional schemes can be immediately deduced from Table 5.1.

¹ One can also deduce (5.36) directly from the exponential form of $\sin^2 jkh/2$.

Table 5.1. The coefficients a_j for approximations from 2nd-order to 14th-order

a_1	a_2	a_3	a_4	a_5	a_6	a_7
4	$\frac{4}{3}$	$\frac{32}{45}$	$\frac{16}{35}$	$\frac{512}{1575}$	$\frac{512}{2079}$	$\frac{4096}{21021}$

5.2.2 The Second Approach

The 1D Case. From (4.60) and (4.62), we derive

$$\left(\frac{\partial^2}{\partial x^2} \right)_h^{(4')} = \tilde{D}_{1,h}^{(4)} \circ D_{1,h}^{(4)}, \quad (5.38)$$

where

$$D_{1,h}^{(4)} = \frac{9}{8} D_{1,h}^{(2)} - \frac{1}{8} D_{1,2h}^{(2)} \text{ and } \tilde{D}_{1,h}^{(4)} = \frac{9}{8} \tilde{D}_{1,h}^{(2)} - \frac{1}{8} \tilde{D}_{1,2h}^{(2)}. \quad (5.39)$$

These relations imply that

$$\widehat{\left(\frac{\partial^2}{\partial x^2} \right)_h^{(4')}} = \left(\frac{9}{8} \widehat{\tilde{D}_{1,h}^{(2)}} - \frac{1}{8} \widehat{\tilde{D}_{1,2h}^{(2)}} \right) \left(\frac{9}{8} \widehat{D_{1,h}^{(2)}} - \frac{1}{8} \widehat{D_{1,2h}^{(2)}} \right). \quad (5.40)$$

So, from the symbols of $D_{1,h}^{(2)}$ and $\tilde{D}_{1,h}^{(2)}$ which are both obviously

$$\widehat{D_{1,h}^{(2)}} = \widehat{\tilde{D}_{1,h}^{(2)}} = \frac{2i}{h} \sin \frac{kh}{2}, \quad (5.41)$$

we can deduce the symbol $D_h(kh)$ of $\left(\frac{\partial^2}{\partial x^2} \right)_h^{(4')}$:

$$\widehat{\left(\frac{\partial^2}{\partial x^2} \right)_h^{(4')}} = \left(\frac{9}{8} \frac{2i}{h} \sin \frac{kh}{2} - \frac{1}{8} \frac{2i}{3h} \sin \frac{3kh}{2} \right)^2, \quad (5.42)$$

which provides after some trigonometric computations

$$\widehat{\left(\frac{\partial^2}{\partial x^2} \right)_h^{(4')}} = -\frac{1}{h^2} \left(4 \sin^2 \frac{kh}{2} + \frac{4}{3} \sin^4 \frac{kh}{2} + \frac{1}{9} \sin^6 \frac{kh}{2} \right). \quad (5.43)$$

From (5.43), we derive immediately the dispersion relation for this approach:

$$\omega^2 = \frac{c^2}{h^2} \left(4 \sin^2 \frac{kh}{2} + \frac{4}{3} \sin^4 \frac{kh}{2} + \frac{1}{9} \sin^6 \frac{kh}{2} \right). \quad (5.44)$$

The 2D Case. By using (5.24), we immediately obtain the 2D dispersion relation:

$$\begin{aligned}\omega^2 = & \frac{c^2}{h^2} \left(4 \left(\sin^2 \frac{k_1 h}{2} + \sin^2 \frac{k_2 h}{2} \right) \right. \\ & + \frac{4}{3} \left(\sin^4 \frac{k_1 h}{2} + \sin^4 \frac{k_2 h}{2} \right) \\ & \left. + \frac{1}{9} \left(\sin^6 \frac{k_1 h}{2} + \sin^6 \frac{k_2 h}{2} \right) \right). \quad (5.45)\end{aligned}$$

Higher-Order Schemes. Here, the symbols of higher-order approximations of $\partial^2/\partial x^2$ come from the relation (4.60) combined with (4.62), and (4.63a) and (4.63b). Their general form is

$$\begin{aligned}\widehat{\left(\frac{\partial^2}{\partial x^2} \right)_h^{(2r')}} &= \left(\sum_{j=1}^r 2i \frac{\lambda'_j}{(2j-1)h} \sin(2j-1) \frac{kh}{2} \right)^2 \\ &= -\frac{4}{h^2} \sum_{p=1}^r \sum_{q=1}^r \frac{\lambda'_p \lambda'_q}{(2p-1)(2q-1)} \sin(2p-1) \frac{kh}{2} \sin(2q-1) \frac{kh}{2}. \quad (5.46)\end{aligned}$$

Now, we have

$$\begin{aligned}& \sin \left[(2p-1) \frac{kh}{2} \right] \sin \left[(2q-1) \frac{kh}{2} \right] \\ &= \frac{1}{2} \left(\cos \left[2(q-p) \frac{kh}{2} \right] - \cos \left[2(p+q-1) \frac{kh}{2} \right] \right) \\ &= \sin^2(p+q-1) \frac{kh}{2} - \sin^2(p-q) \frac{kh}{2}. \quad (5.47)\end{aligned}$$

Hence, we can write

$$\begin{aligned}\widehat{\left(\frac{\partial^2}{\partial x^2} \right)_h^{(2r')}} &= -\frac{4}{h^2} \sum_{p=1}^r \sum_{q=1}^r \frac{\lambda'_p \lambda'_q}{(2p-1)(2q-1)} \\ &\times \left(\sin^2 \left[(p+q-1) \frac{kh}{2} \right] - \sin^2 \left[(p-q) \frac{kh}{2} \right] \right). \quad (5.48)\end{aligned}$$

By inserting (5.36) into (5.48), we obtain the symbols of the schemes in terms of $\sin^2 kh/2$ only. The computation of the first orders shows that we have, for $r > 1$:

$$\widehat{\left(\frac{\partial^2}{\partial x^2} \right)_h^{(2r')}} = \sum_{p=1}^r a_p \sin^{2p} \frac{kh}{2} + \sum_{p=r+1}^{2r-1} b_{r,p} \sin^{2p} \frac{kh}{2}, \quad (5.49)$$

Table 5.2. The coefficients $b_{r,p}$ for approximations from 4th-order to 14th-order

$p =$	$r + 1$	$r + 2$	$r + 3$	$r + 4$	$r + 5$	$r + 6$
$r = 2$	$\frac{1}{9}$					
$r = 3$	$\frac{1}{10}$	$\frac{9}{400}$				
$r = 4$	$\frac{689}{8400}$	$\frac{3}{112}$	$\frac{25}{3136}$			
$r = 5$	$\frac{407}{6048}$	$\frac{493}{18816}$	$\frac{25}{2304}$	$\frac{1225}{331776}$		
$r = 6$	$\frac{11597}{206976}$	$\frac{769}{31680}$	$\frac{42635}{3649536}$	$\frac{245}{45056}$	$\frac{3969}{1982464}$	
$r = 7$	$\frac{78103}{1647360}$	$\frac{5241127}{237219840}$	$\frac{6815}{585728}$	$\frac{240443}{38658048}$	$\frac{1323}{425984}$	$\frac{53361}{44302336}$

where the coefficients a_p are those given in Table 5.1 and $b_{r,p}$ are coefficients which depend on the order of the scheme. In Table 5.2, we give the values (of course positive) of $b_{r,p}$ for $2 \leq r \leq 7$.

5.3 Approximation in Time

5.3.1 Second-Order Approximation in Time

In order to obtain the dispersion relation of the second-order approximation in time, we must only replace, as for the second-order approximations, ω by $4/\Delta t^2 \sin^2 \omega \Delta t / 2$ in all dispersion relations of the semi-discrete approximations in space. For instance, the dispersion relation of the 2D equation fully discretized by the first approach to fourth-order in space is

$$\begin{aligned} & \sin^2 \frac{\omega \Delta t}{2} \\ &= \frac{c^2 \Delta t^2}{h^2} \left(\sin^2 \frac{k_1 h}{2} + \sin^2 \frac{k_2 h}{2} + \frac{1}{3} \sin^4 \frac{k_1 h}{2} + \frac{1}{3} \sin^4 \frac{k_2 h}{2} \right). \end{aligned} \tag{5.50}$$

5.3.2 The Modified Equation Approach

For the modified equation approach, we must add the symbol of the corrective term $c^4 \Delta t^2 / 12 (\Delta^2)_h^{(2)}$ to the dispersion relation of the fully discretized equation. By using relation (4.43), the symbol of $(\Delta^2)_h^{(2)}$ can be written as

$$\widehat{(\Delta^2)_h^{(2)}} = \left(\widehat{\Delta_h^{(2)}} \right)^2. \quad (5.51)$$

So, we obtain

– In 1D:

$$\frac{c^4 \Delta t^2}{12} \widehat{(\Delta^2)_h^{(2)}} = \frac{c^4 \Delta t^4}{3h^4} \sin^4 \frac{kh}{2}, \quad (5.52)$$

– In 2D:

$$\begin{aligned} & \frac{c^4 \Delta t^2}{12} \widehat{(\Delta^2)_h^{(2)}} \\ &= \frac{c^4 \Delta t^4}{3h^4} \left(\sin^4 \frac{k_1 h}{2} + 2 \sin^2 \frac{k_1 h}{2} \sin^2 \frac{k_2 h}{2} + \sin^4 \frac{k_2 h}{2} \right). \end{aligned} \quad (5.53)$$

For instance, the dispersion relation of the modified equation combined with the scheme based on the first approach is, in 2D

$$\begin{aligned} & \sin^2 \frac{\omega \Delta t}{2} \\ &= \frac{c^2 \Delta t^2}{h^2} \left(\sin^2 \frac{k_1 h}{2} + \sin^2 \frac{k_2 h}{2} + \frac{1}{3} \sin^4 \frac{k_1 h}{2} + \frac{1}{3} \sin^4 \frac{k_2 h}{2} \right) \\ & \quad - \frac{c^4 \Delta t^4}{3h^4} \left(\sin^4 \frac{k_1 h}{2} + 2 \sin^2 \frac{k_1 h}{2} \sin^2 \frac{k_2 h}{2} + \sin^4 \frac{k_2 h}{2} \right). \end{aligned} \quad (5.54)$$

The case of higher-order schemes is, for this approach, more complicated. In fact, the higher-order approximations of the biharmonic operators are not the squares of approximations of Δ . Therefore, their symbols must be computed explicitly. This requires some more computations in 1D but is much more difficult in 2D since the discrete operators cannot be decomposed into elementary 1D discrete operators. So, one must compute their symbols directly in 2D.

We are going to show this in the example given in (4.76). The fourth-order approximation of Δ^2 involved in this scheme is

$$\begin{aligned}
& (\Delta^2)_h^{(4)} u_h \\
&= \frac{1}{24h^4} [700u_{\ell,m} - 284(u_{\ell+1,m} + u_{\ell-1,m} + u_{\ell,m+1} + u_{\ell,m-1}) \\
&\quad + 50(u_{\ell+2,m} + u_{\ell-2,m} + u_{\ell,m+2} + u_{\ell,m-2}) \\
&\quad - 4(u_{\ell+3,m} + u_{\ell-3,m} + u_{\ell,m+3} + u_{\ell,m-3}) \\
&\quad + 64(u_{\ell+1,m+1} + u_{\ell-1,m+1} + u_{\ell+1,m-1} + u_{\ell-1,m-1}) \\
&\quad - (u_{\ell+2,m+2} + u_{\ell-2,m+2} + u_{\ell+2,m-2} + u_{\ell-2,m-2})] .
\end{aligned} \tag{5.55}$$

In order to obtain the dispersion relation of (5.55), we must insert the two-dimensional plane wave solution

$$u_h = e^{i(\omega t - k_1 x_1 - k_2 x_2)} \tag{5.56}$$

into (5.55). We obtain

$$\begin{aligned}
& \widehat{(\Delta^2)_h^{(4)}} \\
&= \frac{1}{6h^4} [175 - 142(\cos k_1 h + \cos k_2 h) \\
&\quad + 25(\cos 2k_1 h + \cos 2k_2 h) - 2(\cos 3k_1 h + \cos 3k_2 h) \\
&\quad + 64 \cos k_1 h \cos k_2 h - \cos 2k_1 h \cos 2k_2 h] ,
\end{aligned} \tag{5.57}$$

which finally shows that

$$\begin{aligned}
& \widehat{(\Delta^2)_h^{(4)}} \\
&= \frac{16}{3h^4} \left[3 \left(\sin^4 \frac{k_1 h}{2} + \sin^4 \frac{k_2 h}{2} \right) + 2 \left(\sin^6 \frac{k_1 h}{2} + \sin^6 \frac{k_2 h}{2} \right) \right. \\
&\quad \left. + 6 \sin^2 \frac{k_1 h}{2} \sin^2 \frac{k_2 h}{2} - 2 \sin^4 \frac{k_1 h}{2} \sin^4 \frac{k_2 h}{2} \right. \\
&\quad \left. + 2 \sin^2 \frac{k_1 h}{2} \sin^2 \frac{k_2 h}{2} \left(\sin^2 \frac{k_1 h}{2} + \sin^2 \frac{k_2 h}{2} \right) \right] .
\end{aligned} \tag{5.58}$$

5.3.3 Symmetric Schemes

The dispersion relation for symmetric schemes defined in (4.48), for instance, is more difficult to compute.

Let us take the discrete Fourier transform in space of (4.47):

$$\frac{1}{\Delta t^2} \sum_{j=-p}^p a_{|j|} u_h^{n+j} = c^2 \Delta_h \left(\sum_{j=-q}^q b_{|j|} u_h^{n+j} \right). \quad (5.59)$$

Now, let us recall the general form of a symmetric scheme:

$$u_h^n = e^{in\omega \Delta t - \mathbf{k} \cdot \mathbf{x}}. \quad (5.60)$$

By inserting (5.60) into (5.59), we obtain the relation

$$\sum_{j=-p}^p a_{|j|} e^{i\omega j \Delta t} - c^2 \Delta t^2 D_h(\mathbf{k}) \sum_{j=-q}^q b_{|j|} e^{i\omega j \Delta t} = 0, \quad (5.61)$$

where $D_h(\mathbf{k})$ is the symbol of Δ_h .

By coupling the symmetric terms of (5.61), we obtain:

$$\sum_{j=0}^p a_j \cos j\omega \Delta t - c^2 \Delta t^2 D_h(\mathbf{k}) \sum_{j=0}^q b_j \cos j\omega \Delta t = 0. \quad (5.62)$$

For instance, we obtain for the fourth-order scheme given in (4.71):

$$\begin{aligned} & \frac{1}{\Delta t^2} \left(-\frac{\theta+2}{6} e^{2i\omega \Delta t} + \frac{1+2\theta}{3} e^{i\omega \Delta t} - \theta \right. \\ & \quad \left. + \frac{1+2\theta}{3} e^{-i\omega \Delta t} - \frac{\theta+2}{6} e^{-2i\omega \Delta t} \right) \\ &= \left(\frac{5+2\theta}{12} e^{i\omega \Delta t} + \frac{1-2\theta}{6} + \frac{5+2\theta}{12} e^{-i\omega \Delta t} \right) c^2 D_h(\mathbf{k}). \end{aligned} \quad (5.63)$$

After some computations, we can rewrite this relation as

$$\begin{aligned} & \frac{1}{\Delta t^2} \left(-8(2+\theta) \sin^4 \frac{\omega \Delta t}{2} + 12 \sin^2 \frac{\omega \Delta t}{2} \right) \\ &= \left(3 - (2\theta+5) \sin^2 \frac{\omega \Delta t}{2} \right) c^2 D_h(\mathbf{k}). \end{aligned} \quad (5.64)$$

As we shall see later, (5.64) is more convenient for the study of the accuracy.

As we showed, most of the dispersion relations could be computed on the basis of 1D symbols. This kind of computation can be easily extended to higher-order schemes and 3D by using relations (4.53), (4.62) and (4.68)–(4.73). Higher-order in time comes from (4.75) and (4.78).

5.4 The Case of Systems

5.4.1 The Maxwell Equations

Obtaining a dispersion relation of discretized systems is a little more complicated because of the presence of different equations. We are going to illustrate this kind of computation on the Yee scheme for the Maxwell system in 3D defined in (4.80a)–(4.80f) where we set $\mathbf{J} = 0$.

So, let us look for a plane wave solution defined in (3.11a)–(3.11d) of these equations. For instance, we obtain for (4.80a)

$$\begin{aligned} \varepsilon_0 E_{01}/\Delta t & (e^{i((n+1)\omega\Delta t+(p+\frac{1}{2})k_1 h+qk_2 h+rk_3 h)} \\ & - e^{i(n\omega\Delta t+(p+\frac{1}{2})k_1 h+qk_2 h+rk_3 h)}) \\ & + H_{02}/h (e^{i((n+\frac{1}{2})\omega\Delta t+(p+\frac{1}{2})k_1 h+qk_2 h+(r+\frac{1}{2})k_3 h)} \\ & - e^{i((n+\frac{1}{2})\omega\Delta t+(p+\frac{1}{2})k_1 h+qk_2 h+(r-\frac{1}{2})k_3 h)}) \\ & - H_{03}/h (e^{i((n+\frac{1}{2})\omega\Delta t+(p+\frac{1}{2})k_1 h+(q+\frac{1}{2})k_2 h+rk_3 h)} \\ & - e^{i((n+\frac{1}{2})\omega\Delta t+(p+\frac{1}{2})k_1 h+(q-\frac{1}{2})k_2 h+rk_3 h)}) = 0. \end{aligned} \quad (5.65)$$

By dividing (5.65) by $e^{i((n+\frac{1}{2})\omega\Delta t+(p+\frac{1}{2})k_1 h+qk_2 h+rk_3 h)}$, and collecting the exponentials, we obtain

$$\frac{\varepsilon_0 E_{01}}{\Delta t} \sin \frac{\omega \Delta t}{2} = \frac{H_{03}}{h} \sin \frac{k_2 h}{2} - \frac{H_{02}}{h} \sin \frac{k_3 h}{2}. \quad (5.66)$$

After iterating this process to (4.80b)–(4.80f), we finally obtain:

$$\frac{\varepsilon_0 E_{01}}{\Delta t} \sin \frac{\omega \Delta t}{2} = \frac{H_{03}}{h} \sin \frac{k_2 h}{2} - \frac{H_{02}}{h} \sin \frac{k_3 h}{2}, \quad (5.67a)$$

$$\frac{\varepsilon_0 E_{02}}{\Delta t} \sin \frac{\omega \Delta t}{2} = \frac{H_{01}}{h} \sin \frac{k_3 h}{2} - \frac{H_{03}}{h} \sin \frac{k_1 h}{2}, \quad (5.67b)$$

$$\frac{\varepsilon_0 E_{03}}{\Delta t} \sin \frac{\omega \Delta t}{2} = \frac{H_{02}}{h} \sin \frac{k_1 h}{2} - \frac{H_{01}}{h} \sin \frac{k_2 h}{2}, \quad (5.67c)$$

$$\frac{\mu_0 H_{01}}{\Delta t} \sin \frac{\omega \Delta t}{2} = \frac{E_{02}}{h} \sin \frac{k_3 h}{2} - \frac{E_{03}}{h} \sin \frac{k_2 h}{2}, \quad (5.67d)$$

$$\frac{\mu_0 H_{02}}{\Delta t} \sin \frac{\omega \Delta t}{2} = \frac{E_{03}}{h} \sin \frac{k_1 h}{2} - \frac{E_{01}}{h} \sin \frac{k_3 h}{2}, \quad (5.67e)$$

$$\frac{\mu_0 E_{03}}{\Delta t} \sin \frac{\omega \Delta t}{2} = \frac{E_{01}}{h} \sin \frac{k_2 h}{2} - \frac{E_{02}}{h} \sin \frac{k_1 h}{2}. \quad (5.67f)$$

This system can be written as

$$\sin \frac{\omega \Delta t}{2} \mathbf{E}_0 = \underline{\underline{\underline{C}}}_1 \mathbf{H}_0, \quad (5.68a)$$

$$\sin \frac{\omega \Delta t}{2} \mathbf{H}_0 = \underline{\underline{\underline{C}}}_2 \mathbf{E}_0, \quad (5.68b)$$

where $\underline{\underline{\underline{C}}}_1$ and $\underline{\underline{\underline{C}}}_2$ are the 3×3 matrices

$$\underline{\underline{\underline{C}}}_1 = \frac{\Delta t}{\varepsilon_0 h} \underline{\underline{C}}, \quad \underline{\underline{\underline{C}}}_2 = -\frac{\Delta t}{\mu_0 h} \underline{\underline{C}}, \quad (5.69)$$

$$\underline{\underline{C}} = \begin{pmatrix} 0 & -\sin \frac{k_3 h}{2} & \sin \frac{k_2 h}{2} \\ \sin \frac{k_3 h}{2} & 0 & -\sin \frac{k_1 h}{2} \\ -\sin \frac{k_2 h}{2} & \sin \frac{k_1 h}{2} & 0 \end{pmatrix}. \quad (5.70)$$

At this stage, the solution can be computed in two ways:

- either by inserting (5.68a) into (5.68b) (or vice versa), which leads to the three-dimensional eigenvalues problem

$$\sin^2 \frac{\omega \Delta t}{2} \mathbf{H}_0 = \underline{\underline{\underline{C}}}_2 \underline{\underline{\underline{C}}}_1 \mathbf{H}_0, \quad (5.71)$$

where $\sin^2 \frac{\omega \Delta t}{2}$ is the eigenvalue and \mathbf{H}_0 the eigenvector,

- or directly solve the sixth-dimensional eigenvalue problem

$$\sin \frac{\omega \Delta t}{2} \mathbf{V}_0 = \underline{\underline{\widetilde{C}}} \mathbf{V}_0, \quad (5.72)$$

where \mathbf{V}_0 is the vector $(E_{01}, E_{02}, E_{03}, H_{01}, H_{02}, H_{03})^T$ and $\underline{\underline{\widetilde{C}}}$ is the 6×6 matrix:

$$\underline{\underline{\widetilde{C}}} = \begin{pmatrix} 0 & \underline{\underline{\underline{C}}}_1 \\ \underline{\underline{\underline{C}}}_2 & 0 \end{pmatrix}. \quad (5.73)$$

Both methods provide a stationary mode for each field and the dispersion relation

$$\sin^2 \frac{\omega \Delta t}{2} = \frac{\Delta t^2}{\varepsilon_0 \mu_0 h^2} \left(\sin^2 \frac{k_1 h}{2} + \sin^2 \frac{k_2 h}{2} \right), \quad (5.74)$$

which is the same dispersion relation as that obtained for the second-order approximation of the wave equation (5.25), since $c^2 \varepsilon_0 \mu_0 = 1$.

Higher-order approximations in space will be treated in the same way. Higher-order schemes such as the scheme given in (4.49a) and (4.49b) can be

treated as the symmetric schemes for the wave equation. On the other hand, the modified equation approach defined in (4.90a) and (4.90b) will need a more careful treatment and will be discussed later.

Remark

Actually, the same process is used for the continuous Maxwell equations in the isotropic case. Equivalent results (which actually correspond to the limit of the discrete process when $h \rightarrow 0$ and $\Delta t \rightarrow 0$) are then obtained.

5.4.2 The Elastics System

The case of the elastics system is more complicated since two waves with two different velocities are involved in it. If we search for a plane wave solution of the form

$$v = v_0 e^{i(\omega t + k_1 x + k_2 y)}, \quad (5.75a)$$

$$w = w_0 e^{i(\omega t + k_1 x + k_2 y)}, \quad (5.75b)$$

of the discrete system (4.84a) and (4.84b), we obtain:

$$\begin{aligned} \rho \sin^2 \frac{\omega \Delta t}{2} v_0 &= \frac{\Delta t^2}{h^2} ((\lambda + 2\mu) \sin^2 \frac{k_1 h}{2} v_0 + \mu \sin^2 \frac{k_2 h}{2} v_0 \\ &\quad + (\lambda + \mu) \sin \frac{k_1 h}{2} \sin \frac{k_2 h}{2} \cos \frac{k_1 h}{2} \cos \frac{k_2 h}{2} w_0), \end{aligned} \quad (5.76a)$$

$$\begin{aligned} \rho \sin^2 \frac{\omega \Delta t}{2} w_0 &= \frac{\Delta t^2}{h^2} ((\lambda + \mu) \sin \frac{k_1 h}{2} \sin \frac{k_2 h}{2} \cos \frac{k_1 h}{2} \cos \frac{k_2 h}{2} v_0 \\ &\quad + \mu \sin^2 \frac{k_1 h}{2} w_0 + (\lambda + 2\mu) \sin^2 \frac{k_2 h}{2} w_0). \end{aligned} \quad (5.76b)$$

So, we obtain two dispersion relations which are the eigenvalues of the problem

$$A\mathbf{X} = \rho \sin^2 \frac{\omega \Delta t}{2} \mathbf{X}, \quad (5.77)$$

where $\mathbf{X} = (v_0, w_0)^T$ and

$$A = \gamma \begin{pmatrix} (\lambda + 2\mu)s_1^2 + \mu s_2^2 & (\lambda + \mu)s_1 s_2 c_1 c_2 \\ (\lambda + \mu)s_1 s_2 c_1 c_2 & \mu s_1^2 + (\lambda + 2\mu)s_2^2, \end{pmatrix}, \quad (5.78)$$

with $\gamma = \Delta t^2/h^2$, $s_p = \sin(k_p h/2)$, $c_p = \cos(k_p h/2)$, $p = 1, 2$.

Equation (5.77) provides the two following dispersion relations:

$$\begin{aligned} \rho \sin^2 \frac{\omega \Delta t}{2} &= \frac{\Delta t^2}{h^2} ((\lambda + 2\mu)(\sin^2 \frac{k_1 h}{2} + \sin^2 \frac{k_2 h}{2}) \\ &\quad - (\lambda + \mu) \sin^2 \frac{k_1 h}{2} \sin^2 \frac{k_2 h}{2}), \end{aligned} \tag{5.79a}$$

$$\begin{aligned} \rho \sin^2 \frac{\omega \Delta t}{2} &= \frac{\Delta t^2}{h^2} (\mu(\sin^2 \frac{k_1 h}{2} + \sin^2 \frac{k_2 h}{2}) \\ &\quad + (\lambda + \mu) \sin^2 \frac{k_1 h}{2} \sin^2 \frac{k_2 h}{2}). \end{aligned} \tag{5.79b}$$

One can easily check that, when h and Δt tend both to 0, (5.79a) tends to $\rho\omega^2 = (\lambda + 2\mu)|\mathbf{k}|^2$ and (5.79b) tends to $\rho\omega^2 = \mu|\mathbf{k}|^2$ which are the two dispersion relations obtained in the continuous case.

6. Stability of the Schemes

6.1 General Presentation

Obviously, the purpose of this book is not to provide a general theory of stability for numerical approximations of ordinary differential equations. This has already been done by specialists (see for example [58] or [62]). What we are going to do here is give an intuitive approach to the notion of stability of a scheme for the wave equations. This approach is actually based on plane wave solutions which can be regarded as elementary solutions of the wave equations, as we saw in Sect. 3.1.

Let us consider the plane wave solution of a wave equation

$$u = e^{i(\omega t + \mathbf{k} \cdot \mathbf{x})}. \quad (6.1)$$

Since $\pm\sqrt{\omega/|\mathbf{k}|}$ are the two velocities of propagation of the wave, we require that $\omega \in \mathbb{R}^{+*}$. Now, if we fix \mathbf{k} in the dispersion relation of a scheme, ω (which can, in this case, also be written as ω_h) turns into a function of \mathbf{k} , h and Δt which could, for some values of h and Δt , be complex valued.

Let us examine this phenomenon in a simple example. If we set $c\Delta t/h = \alpha$ in the dispersion relation of the second-order in space and time approximation of the wave equation, we obtain

$$\sin^2 \frac{\omega \Delta t}{2} = \alpha^2 \sin^2 \frac{kh}{2}. \quad (6.2)$$

Obviously, if $\alpha = 2$ for instance, $\sin^2 \omega \Delta t / 2$ is strictly greater than 1 as soon as kh is strictly greater than $\pi/6$. This implies that ω is complex.

Now, if $\omega = \omega_R + i\omega_I$, the velocity loses its physical sense and, on the other hand, our plane wave (6.1) becomes

$$u = e^{-\omega_I t} e^{i(\omega_R t + \mathbf{k} \cdot \mathbf{x})}, \quad (6.3)$$

so that for $\omega_I > 0$, we shall obtain evanescent waves¹ and, $\omega_I < 0$ produces exponentially increasing waves. Of course, both occurrences² are not physical and the second one leads to the phenomenon of blow up of the solution. This phenomenon is called numerical instability.

So, the purpose of the following in this chapter is to determine for which values of h and Δt , ω remains real. As we shall see, the answer to this simple question will not always be easy to find. The stability condition obtained by such an analysis is also called the *von Neumann stability condition*.

6.2 Positivity of an Operator

As we know, the symbol of $-\Delta$ is $|\mathbf{k}|^2$ which is always positive. This result actually coincides with the functional notion of positivity of $-\Delta$ in $H^1(\mathbb{R}^d)$ for instance, which can be written as

$$\int_{\mathbb{R}^d} |\nabla u|^2 d\mathbf{x} \geq 0 \quad \forall u \in H^1(\mathbb{R}^d). \quad (6.4)$$

This property of Δ is fundamental for the existence of the solution of the wave equation. The notion of positivity is also necessary for the stability of the numerical solution. For instance, we saw that the symbol of $-(\partial^2/\partial x^2)_h^{(2)}$ is $4c^2/h^2 \sin^2 kh/2$ which is also always positive and, actually, relation (6.2) shows that if this symbol were negative for some values of kh , $\sin^2 \omega \Delta t/2$ would also be negative for these values and ω would not be real in this case.

Now, we show that the functional definition of the positivity of a continuous or discrete operator is equivalent to the positivity of its symbol. This property will significantly simplify the study of positivity.

As we saw before, both differential and discrete operators operate in Hilbert spaces on which one can define Fourier transforms. Since the scalar products can be written in terms of the norm, by applying Plancherel or Parseval theorems on the conservation of the norms, we can write

$$(Du, u) = (\widehat{D}u, \widehat{u}), \quad (6.5)$$

where $(,)$ is the scalar product in the Hilbert space.

¹ This kind of phenomenon is called “numerical dissipation” and must be avoided for the wave equations which satisfy a principle of energy conservation.

² Which are generally both effective in the case of the wave equations for centered approximations.

Positivity of D can be then written as

$$(Du, u) > 0. \quad (6.6)$$

By using the fact that, by definition, $\widehat{D}u = \widehat{D}\widehat{u}$ and by combining (6.5) and (6.6), we obtain the inequality

$$(\widehat{D}\widehat{u}, \widehat{u}) > 0. \quad (6.7)$$

Now, by assuming that $\widehat{D}|\widehat{u}|^2$ belongs to $L^1(\mathbb{R}^d)$ (which is generally true), (6.7) can be written as

$$\int_{\mathbb{R}^d} \widehat{D}|\widehat{u}|^2 d\mathbf{k} > 0, \quad \forall \widehat{u} \in L^2(\mathbb{R}^d). \quad (6.8)$$

So, if \widehat{D} is not positive almost everywhere, there is an open set Ω of \mathbb{R}^d such that $\forall \mathbf{k} \in \Omega, \widehat{D}(\mathbf{k}) < 0$. Since (6.8) holds for all \widehat{u} , one can choose \widehat{u} such that its support is in Ω . By using (6.8), this would imply that

$$\int_{\mathbb{R}^d} \widehat{D}(\mathbf{k})|\widehat{u}|^2 d\mathbf{k} < 0, \quad (6.9)$$

which is impossible.

This proves that $D > 0 \implies \widehat{D} > 0$.

Conversely, if $\widehat{D} > 0$, then (6.8) (and also (6.7)) hold. So, $\widehat{D}u = \widehat{D}\widehat{u}$ and (6.5) imply the positivity of D .

The positivity of the approximation of a positive differential operator is the first step in the proof of stability.

6.3 Second-Order Approximation in Time

We are going to study the stability of the second-order approximation in time of the wave equation combined with different approximations in space.

We saw that all the dispersion relations given above can be expressed in terms of $\sin^2 kh/2$. So, for our study, it is more convenient to set

$$\alpha = \frac{c\Delta t}{h}, X = \sin^2 \frac{kh}{2} \text{ in 1D}, X_j = \sin^2 \frac{k_j h}{2}, j = 1, 2 \text{ in 2D}. \quad (6.10)$$

On the other hand, we denote by $1/h^2 D_h$ the symbol of $-\Delta_h$ expressed by using (6.10).

With this notation, a general form of a second-order approximation in time and a $2p$ th-order approximation in space is

$$4 \sin^2 \frac{\omega \Delta t}{2} = \alpha^2 D_h. \quad (6.11)$$

Since D_h is positive for all the approximations that we defined in the previous section and $\omega \in \mathbb{R}$ if and only if $0 \leq \sin^2 \omega \Delta t / 2 \leq 1$, CFL is then written as

$$\alpha \leq \alpha_M = \frac{2}{\sup_{\mathbf{X} \in [0,1]^d} \sqrt{D_h(\mathbf{X})}}, \quad (6.12)$$

where $\mathbf{X} = X$ in 1D and $\mathbf{X} = (X_1, \dots, X_d)$ in dimension d .

With this notation, the stability condition is

$$\frac{c \Delta t}{h} \leq \alpha_M. \quad (6.13)$$

6.3.1 Second-Order Approximation in Space

For a second-order scheme in space, one obviously has

$$D_h(\mathbf{X}) = 4 \sum_{j=1}^d X_j \Rightarrow \alpha_M = \frac{1}{\sqrt{d}}, \quad (6.14)$$

in dimension d .

6.3.2 A Basic Property

In order to easily compute the stability conditions of higher-order approximations in space combined with a leapfrog scheme in time, we now give a general property of these approximations in

Proposition 5. *Let*

$$D_{1,h}(X) = \sum_{p=1}^r c_p X^{2p}, \quad c_p \geq 0 \quad \forall p = 1..r \quad (6.15)$$

be the symbol of an approximation of $-\partial^2 / \partial x^2$ and

$$D_{2,h}(\mathbf{X}) = D_{1,h}(X_1) + D_{1,h}(X_2) \quad (6.16)$$

and

$$D_{3,h}(\mathbf{X}) = D_{1,h}(X_1) + D_{1,h}(X_2) + D_{1,h}(X_3) \quad (6.17)$$

the symbols of the same approximation in 2D and 3D.

The stability condition of the discrete wave equation (4.1) based on these approximations and a leapfrog scheme in time is

– In 1D:

$$\alpha \leq \alpha_{1,M} = 2 \left(\sum_{p=1}^r c_p \right)^{-\frac{1}{2}}, \quad (6.18)$$

– In 2D:

$$\alpha \leq \alpha_{2,M} = \frac{\sqrt{2}}{2} \alpha_{1,M}. \quad (6.19)$$

– In 3D:

$$\alpha \leq \alpha_{3,M} = \frac{\sqrt{3}}{3} \alpha_{1,M}. \quad (6.20)$$

Proof. Since $c_p \geq 0 \forall p = 1..r$ and $X \geq 0$, $D_{1,h}(X)$ is increasing on $[0, 1]$ and reaches its maximum at $X = 1$. Thus (6.18) comes immediately from (6.12). On the other hand, let $\mathbf{X}_M = (X_{1M}, X_{2M})$ be the point at which $D_{2,h}(\mathbf{X})$ reaches its maximum when $\mathbf{X} \in [0, 1]^2$ and let us set $X_M = \max(X_{1M}, X_{2M})$. Since $D_{1,h}(X)$ is increasing on $[0, 1]$, we have

$$D_{2,h}(\mathbf{X}_M) = D_{1,h}(X_{1M}) + D_{1,h}(X_{2M}) \leq D_{1,h}(X_M) + D_{1,h}(X_M), \quad (6.21)$$

which implies that $X_{1M} = X_{2M} = X_M$.

So, the maximum of $D_{2,h}(\mathbf{X})$ is reached on the diagonal $X_1 = X_2$ of $[0, 1]^2$. On this diagonal, we have

$$D_{2,h}(\mathbf{X}_M) = 2D_{1,h}(X_1), \quad (6.22)$$

which reaches its maximum at $X_1 = 1$ and (6.19) comes from (6.18) and (6.12). A similar proof leads to (6.20). \diamond

6.3.3 Application to Higher-Order Approximation in Space

Tables 5.1 and 5.2 and the construction of higher-order approximations given in Sect. 4.5 show that the first and second approaches are within the framework of proposition 5. So, we can easily deduce from Tables 5.1 and 5.2 the stability conditions for higher-order approximations in space and a leapfrog scheme in time for both approaches. The exact and approximative values of these conditions are given in Tables 6.1 and 6.2 in 1D. The 2D and 3D cases are deduced from Proposition 5.

Remark

The stability conditions for the second approach are all rational because we have, in this case:

$$-\left(\frac{\partial^2}{\partial x^2}\right)_h^{(4)} = \widehat{D}_{h,1}^{(4)} \circ D_{h,1}^{(4)} \implies -\left(\widehat{\frac{\partial^2}{\partial x^2}}\right)_h^{(4)} = \left(\widehat{D}_{h,1}^{(4)}\right)^2,$$

since $\widehat{D}_{h,1}^{(4)} = \widehat{D}_{h,1}^{(4)}$.

Table 6.1. The stability conditions in 1D for approximations from 4th-order to 14th-order by using the first approach

Order	4	6	8
$\alpha_M =$	$\frac{\sqrt{3}}{2}$	$\frac{3}{34}\sqrt{85}$	$\frac{3}{32}\sqrt{70}$
$\alpha_M \simeq$	0.866 025	0.813 489	0.784 368
Order	10	12	14
$\alpha_M =$	$\frac{5}{16}\sqrt{6}$	$\frac{15\sqrt{82\,929}}{5744}$	$\frac{105\sqrt{14\,401\,101}}{537\,104}$
$\alpha_M \simeq$	0.765 465	0.752 021	0.741 871

Table 6.2. The stability conditions in 1D for approximations from 4th-order to 14th-order by using the second approach

Order	4	6	8	10	12	14
$\alpha_M =$	$\frac{6}{7}$	$\frac{120}{149}$	$\frac{1680}{2161}$	$\frac{40\,320}{53\,089}$	$\frac{887\,040}{1\,187\,803}$	$\frac{46\,126\,080}{62\,566\,171}$
$\alpha_M \simeq$	0.857 142	0.805 369	0.777 417	0.759 479	0.746 790	0.737 236

6.4 The Modified Equation Approach

For the modified equation approach, we shall only study fourth-order approximations in 1D and 2D. Higher dimension and higher-order approximations are technically too complicated.

6.4.1 Preliminary Results

In this section, we denote $(1/h^4)D'_h$ the symbol of $(\Delta_h^{(2)})^2$ expressed by using (6.10). This symbol is actually equal to

$$\begin{aligned} -D'_h &= 16X^2 \quad \text{in 1D}, \\ -D'_h &= 16(X_1 + X_2)^2 \quad \text{in 2D}. \end{aligned}$$

In the case of the modified equation approach, the stability condition can be written as

$$\left\{ \begin{array}{l} \sup_{\mathbf{X} \in [0,1]^d} \frac{\alpha^2}{4} \tilde{D}_h(\mathbf{X}) \leq 1, \\ \text{where } \tilde{D}_h = D_h(\mathbf{X}) - \frac{\alpha^2}{12} D'_h(\mathbf{X}). \end{array} \right. \quad (6.23)$$

The first step of the study of stability in this case is the determination of conditions of positivity of \tilde{D}_h which is the symbol of the modified operator in space. For this, we first notice that for the first and second approaches, all the terms of \tilde{D}_h are positive except for $1/3(1 - \alpha^2)X^2$ in 1D and $1/3(X_1^2 + X_2^2 - \alpha^2(X_1 + X_2)^2)$ in 2D. In 1D the positivity of this term is obtained as soon as $\alpha \leq 1$. In 2D, we have

$$2X_1X_2 \leq X_1^2 + X_2^2 \Rightarrow (X_1 + X_2)^2 \leq 2(X_1^2 + X_2^2),$$

which implies that

$$(X_1^2 + X_2^2 - \alpha^2(X_1 + X_2)^2) \geq (1 - 2\alpha^2)(X_1^2 + X_2^2), \quad (6.24)$$

so that the positivity of \tilde{D}_h is obtained for $\alpha \leq \sqrt{2}/2$.

All this study can be summarized in the following lemma:

Lemma 1. *A sufficient condition for the positivity of \tilde{D}_h is*

$$-\alpha \leq 1 \text{ in 1D},$$

$$-\alpha \leq \frac{\sqrt{2}}{2} \text{ in 2D}.$$

6.4.2 Fourth-Order Approximation in Space: First Approach

The 1D Case. For the first approach, we have, in 1D,

$$\tilde{D}_h = 4 \left(X + \frac{1}{3}X^2 - \frac{\alpha^2}{3}X^2 \right) \quad (6.25)$$

and the stability condition (6.23) becomes

$$\sup_{X \in [0,1]} \alpha^2 \left(X + \frac{1}{3}X^2 - \frac{\alpha^2}{3}X^2 \right) \leq 1. \quad (6.26)$$

For $X = 1$, we obtain from (6.26)

$$-\alpha^4 + 4\alpha^2 - 3 \leq 0. \quad (6.27)$$

This inequality holds for $\alpha^2 \leq 1$ or $\alpha^2 \geq 3$. For $\alpha^2 \leq 1$, \tilde{D}_h is always positive and increasing on $[0, 1]$. So,

$$\tilde{D}_h(X) \leq \tilde{D}_h(1). \quad (6.28)$$

This inequality shows that (6.23) is satisfied for $\alpha^2 \leq 1$. This condition combined with the positivity of \tilde{D}_h ensures the stability of the scheme for $\alpha^2 \leq 1$.

Now, we have

$$\tilde{D}_h(1) = \frac{4 - \alpha^2}{3}, \quad (6.29)$$

which is always negative for $\alpha^2 > 4$. This shows that \tilde{D}_h is not positive for these values of α and the scheme is not stable. We must now check if $3 \leq \alpha^2 \leq 4$ is an acceptable interval.

The maximum of \tilde{D}_h is reached for

$$X_0 = \frac{3}{2(\alpha^2 - 1)}. \quad (6.30)$$

For this value, we have

$$\frac{\alpha^2}{4} \tilde{D}_h(X_0) = \frac{3\alpha^2}{4(\alpha^2 - 1)} > 1, \quad \forall \alpha^2 \in [3, 4[. \quad (6.31)$$

Equation (6.31) shows that the relation (6.23) does not hold for $3 \leq \alpha^2 < 4$. Formally, the scheme could be stable for $\alpha = 2$ but, since this value is isolated, it is not realistic for a numerical model. So, we can write:

$$\alpha_M = 1. \quad (6.32)$$

A general study of the 1D case for the first approach, carried out in [6], shows that this stability condition holds for any order.

The 2D Case. In this case,

$$\tilde{D}_h = 4 \left(X_1 + X_2 + \frac{1}{3}(X_1^2 + X_2^2) - \frac{\alpha^2}{3}(X_1 + X_2)^2 \right) \quad (6.33)$$

and (6.23) can be written as

$$\sup_{\mathbf{X} \in [0,1]^2} \alpha^2 \left(X_1 + X_2 + \frac{1}{3}(X_1^2 + X_2^2) - \frac{\alpha^2}{3}(X_1 + X_2)^2 \right) \leq 1. \quad (6.34)$$

By setting $X = 1$ in (6.34), we obtain

$$-4\alpha^4 + 8\alpha^2 - 3 \leq 0. \quad (6.35)$$

The solutions of this inequality are

$$\alpha^2 \leq \frac{1}{2} \text{ or } \alpha^2 \geq \frac{3}{2}. \quad (6.36)$$

On the other hand, the extrema of (6.33) are given by the system

$$\frac{2}{3} (X_1 - \alpha^2(X_1 + X_2)) + 1 = 0, \quad (6.37a)$$

$$\frac{2}{3} (X_2 - \alpha^2(X_1 + X_2)) + 1 = 0, \quad (6.37b)$$

whose unique solution is

$$X_1^0 = \frac{3}{2(2\alpha^2 - 1)} = X_2^0. \quad (6.38)$$

Since

$$(X_1^0, X_2^0) \in [0, 1]^2 \Leftrightarrow \alpha^2 \geq 5/4, \quad (6.39)$$

the maximum of \tilde{D}_h is reached on the boundary of $[0, 1]^2$ for $\alpha \leq 1/2$.

For obvious symmetry reasons, the boundary of $[0, 1]^2$ can be reduced to the segments S_1 and S_2 defined by

$$S_1 = \{(X_1, X_2) \in [0, 1]^2 \text{ such that } X_1 = 0\}, \quad (6.40a)$$

$$S_2 = \{(X_1, X_2) \in [0, 1]^2 \text{ such that } X_2 = 1\}, \quad (6.40b)$$

on which we have

$$\tilde{D}_h(0, X_2) = 4 \left(X_2 + \frac{1}{3}X_2^2 - \frac{\alpha^2}{3}X_2^2 \right), \quad (6.41a)$$

$$\tilde{D}_h(X_1, 1) = 4 \left(X_1 + \frac{1}{3}X_1^2 + \frac{4}{3} - \frac{\alpha^2}{3}(X_1^2 + 1) \right). \quad (6.41b)$$

A simple computation shows that these two functions increase on $[0, 1]$ when $\alpha^2 \leq 1/2$ and so, the maximum of \tilde{D}_h is reached at $(X_1^0, X_2^0) = (1, 1)$. This fact combined with Lemma 1 implies that (6.34) is satisfied for $\alpha^2 \leq 1/2$ and the scheme is stable.

Although this study provides a sufficient condition for stability for the scheme, we will show that it is in fact necessary.

Equation (6.36) shows that the scheme could be stable only if $\alpha^2 \geq 3/2$. For these values, we know that the extremum belongs to $[0, 1]^2$.

We have:

$$\alpha^2 \tilde{D}_h(X_1^0, X_2^0) = \frac{3\alpha^2}{2(2\alpha^2 - 1)}, \quad (6.42)$$

which is strictly greater than 1 when $\alpha \in [3/2, 2[$. So, in this interval, (6.34) is not satisfied.

On the other hand,

$$\tilde{D}_h(1, 1) = \frac{4}{3}(2 - \alpha^2) \quad (6.43)$$

is negative as soon as $\alpha^2 > 2$. This implies that the approximation of $-\Delta$ is not positive and then, the scheme cannot be stable.

Finally, we showed that the stability condition is verified for $\alpha^2 \in [0, 1/2] \cup \{2\}$. Since the value $\alpha^2 = 2$ is not acceptable because it is isolated, we can conclude:

$$\alpha_M = \frac{\sqrt{2}}{2}. \quad (6.44)$$

6.4.3 Fourth-Order Approximation in Space: Second Approach

We are now going to study the second approach. Since the computations are more complicated in this case, we shall limit our study to sufficient conditions. Actually, a further tedious study can show that these conditions are also necessary.

The 1D Case. In this case, we have

$$\tilde{D}_h = 4 \left(X + \frac{1}{3}X^2 - \frac{\alpha^2}{3}X^2 + \frac{1}{36}X^3 \right) \quad (6.45)$$

and the stability condition (6.23) becomes

$$\sup_{X \in [0,1]} \alpha^2 \left(X + \frac{1}{3} X^2 - \frac{\alpha^2}{3} X^2 + \frac{1}{36} X^3 \right) \leq 1. \quad (6.46)$$

For $X = 1$, we obtain from (6.46)

$$-\frac{1}{3} \alpha^4 + \frac{49}{36} \alpha^2 - 1 \leq 0. \quad (6.47)$$

This inequality holds for $\alpha^2 \leq (49 - \sqrt{673})/24$ or $\alpha^2 \geq (49 + \sqrt{673})/24$.

On the other hand, the derivative of \tilde{D}_h is equal to zero at the points

$$X_1^0 = 4(\alpha^2 - 1) - 2\sqrt{1 - 8\alpha^2 + 4\alpha^4}, \quad (6.48a)$$

$$X_2^0 = 4(\alpha^2 - 1) + 2\sqrt{1 - 8\alpha^2 + 4\alpha^4}. \quad (6.48b)$$

The study of these functions of α^2 shows that they are negative for $\alpha^2 \leq (49 - \sqrt{673})/24$.

So, \tilde{D}_h is monotone and positive (thanks to Lemma 1) for $X \in [0, 1]$ and $\tilde{D}_h(0) = 0$. These properties imply that \tilde{D}_h is increasing for $X \in [0, 1]$ and reaches its maximum at $X = 1$. Since $\tilde{D}_h(1) \leq 1$ when $\alpha^2 \leq (49 - \sqrt{673})/24$, (6.46) is satisfied for these values of α^2 and so,

$$\alpha \leq \alpha_M = \frac{1}{2} \sqrt{\frac{1}{6}(49 - \sqrt{673})} \simeq 0.9801734 \quad (6.49)$$

is a sufficient condition of stability of the scheme.

The 2D Case. Here, we have:

$$\begin{aligned} \tilde{D}_h = & 4 \left(X_1 + X_2 + \frac{1}{3}(X_1^2 + X_2^2) - \frac{\alpha^2}{3}(X_1 + X_2)^2 \right. \\ & \left. + \frac{1}{36}(X_1^3 + X_2^3) \right) \end{aligned} \quad (6.50)$$

and (6.23) can be written as

$$\sup_{\mathbf{X} \in [0,1]^2} \alpha^2 \left(X_1 + X_2 + \frac{1}{3}(X_1^2 + X_2^2) - \frac{\alpha^2}{3}(X_1 + X_2)^2 \right. \\ \left. + \frac{1}{36}(X_1^3 + X_2^3) \right) \leq 1. \quad (6.51)$$

By setting $X = 1$ in (6.51), we obtain the inequality

$$-24\alpha^4 + 49\alpha^2 - 18 \leq 0, \quad (6.52)$$

whose solutions are

$$\alpha^2 \leq \frac{49 - \sqrt{673}}{48} \text{ or } \alpha^2 \geq \frac{49 + \sqrt{673}}{48}. \quad (6.53)$$

On the other hand, the possible extrema of (6.50) are given by the equations

$$\frac{1}{12}X_1^2 - \frac{2}{3}\alpha^2(X_1 + X_2) + \frac{1}{3}X_1 + 1 = 0, \quad (6.54a)$$

$$\frac{1}{12}X_2^2 - \frac{2}{3}\alpha^2(X_1 + X_2) + \frac{1}{3}X_2 + 1 = 0, \quad (6.54b)$$

whose solutions are

$$\begin{cases} X_1 = 2(2 \pm \sqrt{1 - 16\alpha^2}), \\ X_2 = 2(-2 \pm \sqrt{1 - 16\alpha^2}), \end{cases} \quad (6.55a)$$

$$\begin{cases} X_1 = 8\alpha^2 - 4 \pm 2\sqrt{16\alpha^4 - 16\alpha^2 + 1}, \\ X_2 = 8\alpha^2 - 4 \pm 2\sqrt{16\alpha^4 - 16\alpha^2 + 1}. \end{cases} \quad (6.55b)$$

Obviously, the solutions given in (6.55a) are never real for $\alpha^2 \leq (49 - \sqrt{673})/48$ and a study of the functions given in (6.55b) shows that the second class of solutions is always negative for these values of α^2 .

So, the maximum of \tilde{D}_h is reached on the segments S_1 or S_2 . On these segments, we have

$$\tilde{D}_h(0, X_2) = 4 \left(X_2 + \frac{1}{3}X_2^2 - \frac{\alpha^2}{3}X_2^2 + \frac{1}{36}X_2^3 \right), \quad (6.56a)$$

$$\tilde{D}_h(X_1, 1) = 4 \left(X_1 + \frac{1}{3}X_1^2 - \frac{\alpha^2}{3}(X_1 + 1)^2 + \frac{1}{36}X_1^3 + \frac{49}{36} \right). \quad (6.56b)$$

The first function is the same as (6.45) for $X = X_2$ and reaches its maximum at $X_2 = 1$ and the study of the second function shows that it increases from $\tilde{D}_h(0, 1)$ to $\tilde{D}_h(1, 1)$. Then, the maximum of \tilde{D}_h is reached at the point $(1, 1)$ and so, (6.51) is satisfied for $\alpha^2 \leq (49 - \sqrt{673})/48$. This result combined with Lemma 1 leads to the following sufficient condition for stability:

$$\alpha \leq \alpha_M = \frac{1}{4} \sqrt{\frac{1}{3}(49 - \sqrt{673})} \simeq 0.693\,087\,2. \quad (6.57)$$

Remarks

- As for the second-order discretization in time, we obtain the 2D stability conditions by dividing the 1D ones by $\sqrt{2}$. However, no general theorem can be given at this stage about this phenomenon.

2. The stability conditions for the second approach in 1D and 2D are slightly less than those obtained for the first approach. As we shall see later, this small difference has a great influence on features of the two approaches.

6.5 Symmetric Schemes

6.5.1 First Method

We now study the stability of the scheme given in (4.48). By setting $\lambda = \sin^2 \omega \Delta t / 2$ and $\alpha = c \Delta t / h$ in (5.64), we obtain the equation:

$$-8(2 + \theta)\lambda^2 + 12\lambda = (3 - (2\theta + 5)\lambda) D, \quad (6.58)$$

where $D = \alpha^2 \tilde{D}_h(\mathbf{X})$ and \tilde{D}_h is the symbol of Δ_h multiplied by h^2 .

This equation, second-order in λ , must have its roots (which are actually squares of sine functions) positive and less than or equal to 1. We are going to carry out the computation for $\theta = 0$. For this value, the roots of (6.58) are:

$$\lambda_1 = \frac{12 + 5D}{32} + \frac{1}{32} \sqrt{25D^2 - 72D + 144}, \quad (6.59a)$$

$$\lambda_2 = \frac{12 + 5D}{32} - \frac{1}{32} \sqrt{25D^2 - 72D + 144}. \quad (6.59b)$$

The discriminant of the term under the square root is equal to -2304 , which implies that this term is positive for any value of D . On the other hand, λ_1 is obviously always positive. Let us now look for the values of D so that $\lambda_1 \leq 1$, which can be written as

$$\sqrt{25D^2 - 72D + 144} \leq 20 - 5D. \quad (6.60)$$

Since the right-hand side is negative for $D > 4$, solutions of inequality (6.60) must be such that $0 \leq D \leq 4$ (since D must be positive). On this interval, $20 - 5D$ is positive and inequality (6.60) is equivalent to the inequality

$$25D^2 - 72D + 144 \leq (20 - 5D)^2, \quad (6.61)$$

whose solution is

$$D \leq 2. \quad (6.62)$$

Now, the values for which λ_2 is positive are given by

$$12 + 5D \geq \sqrt{25D^2 - 72D + 144}. \quad (6.63)$$

Since D is positive, the left-hand side is always positive and inequality (6.63) can be rewritten as

$$(12 + 5D)^2 \geq 25D^2 - 72D + 144. \quad (6.64)$$

This relation is true as soon as $D \geq 0$.

The inequality $\lambda_2 \leq 1$ can be written as

$$5D - 20 \leq \sqrt{25D^2 - 72D + 144}, \quad (6.65)$$

which is always true for $0 \leq D \leq 4$. For $D \geq 4$, the left-hand side is positive and inequality (6.65) is equivalent to the inequality

$$(5D - 20)^2 \leq 25D^2 - 72D + 144, \quad (6.66)$$

which is true for $D \geq 0$.

So, the intersection of the solutions of the different inequalities is:

$$0 \leq D \leq 2. \quad (6.67)$$

Now, since $D = \alpha^2 \tilde{D}_h(\mathbf{X})$, the stability condition of (4.48) can be written as

$$\alpha \leq \alpha_M = \frac{\sqrt{2}}{\sup_{\mathbf{X} \in [0,1]^d} \sqrt{D_h(\mathbf{X})}}. \quad (6.68)$$

This condition is equal to the condition obtained for a leapfrog scheme in time divided by $\sqrt{2}$.

For $\theta > 0$, one can show by the same method (just more complicated because of the presence of θ) that the stability condition can be written as

$$\alpha \leq \alpha_M = \frac{\sqrt{2 \frac{2\theta + 1}{\theta + 1}}}{\sup_{\mathbf{X} \in [0,1]^d} \sqrt{D_h(\mathbf{X})}}. \quad (6.69)$$

which tends to the leapfrog stability condition when $\theta \rightarrow \infty$. However, as we shall see later, the accuracy of the method decreases when θ increases and, therefore, one cannot use values of this parameter which are too large.

The stability condition of the higher-order symmetric schemes can be derived from relation (5.62) by setting $\cos j\omega\Delta t = 1 - 2\sin^2 j\omega\Delta t/2$ and using (5.36). Then, by setting $\lambda = \sin^2 \omega\Delta t/2$, we obtain a p th-order polynomial equation for a $2p$ th-order scheme.

The main drawback of this method is the computation of the polynomial in λ which is not so easy. For this reason, we are going to give a second method which is actually classically used for ODE.

6.5.2 Second Method

By setting $z = e^{i\omega j \Delta t}$ in (5.61) and by multiplying by z^p , we obtain the symmetric polynomial in z :

$$\sum_{j=-p}^p a_{|j|} z^{p+j} - D \sum_{j=-q}^q b_{|j|} z^{p+j} = 0, \quad (6.70)$$

where D is defined in the same way as in (6.58).

Since the only admissible values of ω are such that $\omega \in \mathbb{R}$, the roots of the polynomial equation defined in (6.70) must satisfy the condition

$$|z| = 1. \quad (6.71)$$

So, the stability condition is obtained for the maximal value of D for which all the roots of (6.70) satisfy (6.71).

For example, we obtain for the scheme defined in (4.48) when $\theta = 0$:

$$4z^4 + (5D - 4)z^3 + 2Dz^2 + (5D - 4)z + 4 = 0. \quad (6.72)$$

Of course, it is difficult and even impossible to obtain the exact solutions of such an equation as well as higher-order equations in λ for the first method. For this reason, such equations must be solved by numerical methods by testing different values of D .

This example shows that although the polynomials involved in this method can be obtained much more easily, these polynomials are of $2p$ th-order for $2p$ th-order of approximation. this

Remark

Although the symmetric scheme is less stable than the modified equation approach, one iteration in time of this scheme is however less expensive since, unlike the modified equation, it only uses the stencil of the fourth-order approximation in space without any additional points.

6.6 The Case of Systems

The Maxwell equations present no additional difficulty versus the wave equation since their dispersion relation can be expressed in the same way as the one for the wave equation (as shown in (5.74)).

This is not the case for the elastics system for which we gave in (5.79a) and (5.79b) two different dispersion relations corresponding to the two velocities of the elastic waves. By using the change of variables defined in (6.10), we obtain, from these dispersion relations, the two following double inequalities:

$$0 \leq \sup_{\mathbf{X} \in [0,1]^2} \left[\frac{\Delta t^2}{\rho h^2} ((\lambda + 2\mu)(X_1 + X_2) - (\lambda + \mu)X_1 X_2) \right] \leq 1, \quad (6.73a)$$

$$0 \leq \sup_{\mathbf{X} \in [0,1]^2} \left[\frac{\Delta t^2}{\rho h^2} (\mu(X_1 + X_2) + (\lambda + \mu)X_1 X_2) \right] \leq 1. \quad (6.73b)$$

Both functions involved in (6.73a) and (6.73b) are of the form:

$$f(X_1, X_2) = a(X_1 + X_2) + bX_1 X_2, \quad (6.74)$$

which reaches its extrema at $(X_1, X_2) = (-a/b, -a/b)$.

For (6.73b), $(-a/b, -a/b)$ is negative and thus not acceptable and, for (6.73a), it is equal to $((\lambda + 2\mu)/(\lambda + \mu), (\lambda + 2\mu)/(\lambda + \mu))$ and the two components of this vector are greater than 1.

So, in both cases, the maximum of f is reached on the boundary of $[0, 1]^2$. In the second case, f is always positive and reaches its maximum at $(X_1, X_2) = (1, 1)$ and, in the first case, we obtain, on S_1 and S_2 defined by (6.40a) and (6.40b):

$$0 \leq \sup_{X_2 \in [0,1]} \left[\frac{\Delta t^2}{\rho h^2} (\lambda + 2\mu) X_2 \right] \leq 1, \quad (6.75a)$$

$$0 \leq \sup_{X_1 \in [0,1]} \left[\frac{\Delta t^2}{\rho h^2} (\mu X_1 + (\lambda + \mu)) \right] \leq 1. \quad (6.75b)$$

Here also, the maximum is obviously reached at $(X_1, X_2) = (1, 1)$. For this point, (6.73a) and (6.73b) provide:

$$\frac{\Delta t^2}{\rho h^2} (\lambda + 3\mu) \leq 1, \quad (6.76a)$$

$$\frac{\Delta t^2}{\rho h^2} (\lambda + 2\mu) \leq 1. \quad (6.76b)$$

Since the most restrictive relation is (6.76a), it is the stability condition of the scheme, which, by taking into account (1.40) and (1.41), can be rewritten as

$$\frac{\Delta t}{h} \leq \frac{1}{\sqrt{V_P^2 + V_S^2}}. \quad (6.77)$$

The computation of the stability condition of higher-order methods is more complicated because it involves higher-order polynomials. Some examples of

such computations for fourth-order methods in space and time can be found in [14].

Although involving higher-order polynomials whose roots are not always easy to find, the plane wave analysis provides a straightforward method to compute the stability condition of a numerical model. Energy techniques, developed in Chap. 9, which are more appropriate for heterogeneous media, are, however, more difficult.

6.7 A Numerical Illustration

We give below a numerical illustration of an instability phenomenon in 1D. We propagate an initial data on a segment of length 12 and we set $u(0, t) = u(12, t) = 0$. The scheme used is a fourth-order scheme in space (first approach) and in time (modified equation approach) which is exact for its maximal stability condition, as we shall see in the next section. The exact solution (which corresponds to $\alpha = 1$) is given as a dotted line and the solution obtained for $\alpha = 1.001$ is the continuous line.

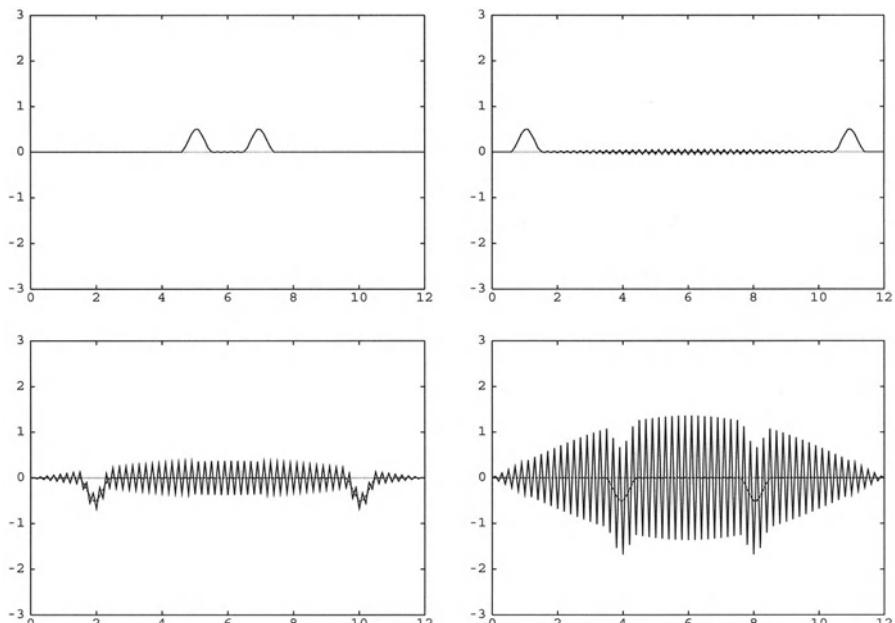


Fig. 6.1. Evolution of an instability phenomenon over time. The solutions are drawn for $t = 1\text{ s}$ (above right), $t = 5\text{ s}$ (above left), $t = 8\text{ s}$ (below right), $t = 10\text{ s}$ (below left)

7. Numerical Dispersion and Anisotropy

7.1 Phase and Group Velocities

As we saw in a previous section, a plane wave solution $e^{i(\omega t - \mathbf{k} \cdot \mathbf{x})}$ of the continuous wave equation provides the following dispersion relation:

$$\omega^2 = c^2 |\mathbf{k}|^2. \quad (7.1)$$

If we assume that $c > 0$, this can be written as

$$c = \frac{\omega}{|\mathbf{k}|}. \quad (7.2)$$

The same approach applied to an approximation of the wave equation provides a dispersion relation between the discrete pulsation ω_h and the wave vector \mathbf{k} . In the same wave, we can define the discrete velocity c_h of the wave propagated by this approximated wave equation:

$$c_h = \frac{\omega_h}{|\mathbf{k}|}. \quad (7.3)$$

Then, we are able to introduce the *numerical dispersion coefficient*:

$$q_h = \frac{c_h}{c} = \frac{\omega_h}{\omega}. \quad (7.4)$$

This non-dimensional quantity measures the error in the velocity of the numerical wave which is a basic error for wave equations. These velocities are called *phase velocities* of the wave. They correspond to a *monochromatic* wave, i.e. a wave of a given frequency. A better measure is given by the *group velocity* which takes into account a whole “packet” of waves around a given frequency.

In order to obtain this group velocity, we first define an initial value of the wave equation as the superposition of monochromatic waves given by the following relation:

$$u_0(\mathbf{x}) = \int_{|\mathbf{k}-\mathbf{k}_0|<\varepsilon} a(\mathbf{k}) e^{-i\mathbf{k} \cdot \mathbf{x}} d\mathbf{k}, \quad (7.5)$$

where ε is a small given parameter and $a(\mathbf{k})$ is the amplitude of the plane wave.

Then, we know that each monochromatic wave propagates with a pulsation $\omega(\mathbf{k})$, so that we can write:

$$u(\mathbf{x}, t) = \int_{|\mathbf{k} - \mathbf{k}_0| < \varepsilon} a(\mathbf{k}) e^{i(\omega(\mathbf{k})t - \mathbf{k} \cdot \mathbf{x})} d\mathbf{k}. \quad (7.6)$$

Now, let us write the Taylor expansion of ω_h around \mathbf{k}_0 :

$$\omega(\mathbf{k}) = \omega(\mathbf{k}_0) + \mathbf{grad}\omega(\mathbf{k}_0) \cdot (\mathbf{k} - \mathbf{k}_0) + O(|\mathbf{k} - \mathbf{k}_0|^2). \quad (7.7)$$

By inserting (7.7) into (7.6) we obtain:

$$\begin{aligned} u(\mathbf{x}, t) = & e^{i(\omega(\mathbf{k}_0) - \mathbf{grad}\omega(\mathbf{k}_0) \cdot \mathbf{k}_0)t} \\ & \times \int_{|\mathbf{k} - \mathbf{k}_0| < \varepsilon} a(\mathbf{k}) e^{-i[(\mathbf{x} - \mathbf{grad}\omega(\mathbf{k}_0)t)\mathbf{k} - O(|\mathbf{k} - \mathbf{k}_0|^2 t)]} d\mathbf{k}, \end{aligned} \quad (7.8)$$

which can be rewritten as

$$\begin{aligned} u(\mathbf{x}, t) = & e^{i(\omega(\mathbf{k}_0) - \mathbf{grad}\omega(\mathbf{k}_0) \cdot \mathbf{k}_0)t} u_0(\mathbf{x} - \mathbf{grad}\omega(\mathbf{k}_0)t) \\ & + tO(|\mathbf{k} - \mathbf{k}_0|^2). \end{aligned} \quad (7.9)$$

From (7.9), we can derive the modulus of u :

$$|u(\mathbf{x}, t)| = |u_0(\mathbf{x} - \mathbf{grad}\omega(\mathbf{k}_0)t)| + tO(|\mathbf{k} - \mathbf{k}_0|^2). \quad (7.10)$$

Equation (7.9) shows that, to a first approximation, the modulus of u propagates with a velocity equal to

$$c = |\mathbf{grad}\omega(\mathbf{k}_0)|. \quad (7.11)$$

This velocity is the group velocity of the wave. The definition of the discrete group velocity is derived in the same way and the numerical dispersion is given by the non-dimensional coefficient:

$$q_h = \frac{|\mathbf{grad}\omega_h|}{|\mathbf{grad}\omega|}. \quad (7.12)$$

The error given by the group velocity is generally greater than that given by the phase velocity but is more significant. As we shall see later, the difference between the two errors grows with the order of the method.

7.2 The Concept of Numerical Dispersion

Besides the error committed in the velocity, q_h measures the numerical dispersion of the scheme which produces parasitic waves around the solution.

In fact, these parasitic waves arise from the following process: As we know, the continuous velocity c in a homogeneous medium is a constant, whereas the discrete velocity c_h is a function of \mathbf{k} and, therefore, of the frequency of the wave. So, when the wave is polychromatic, the dependency on the frequency implies that each monochromatic component of the wave moves with a different velocity. When c_h depends significantly on the frequency (which occurs when, for instance, the space-step is too large), some of these components leave the physical wave and produce a sequence of parasitic waves which have no physical meaning. Of course, such a phenomenon can seriously damage the results obtained by a numerical model. Some schemes can develop such parasitic waves even for small space-steps. They are called *dispersive* schemes. A dispersive scheme can even be a higher-order scheme. Its dispersive character derives from the constant in front of the leading power of h in the Taylor expansion of q_h .

We give a numerical illustration of this phenomenon in Fig. 7.1.

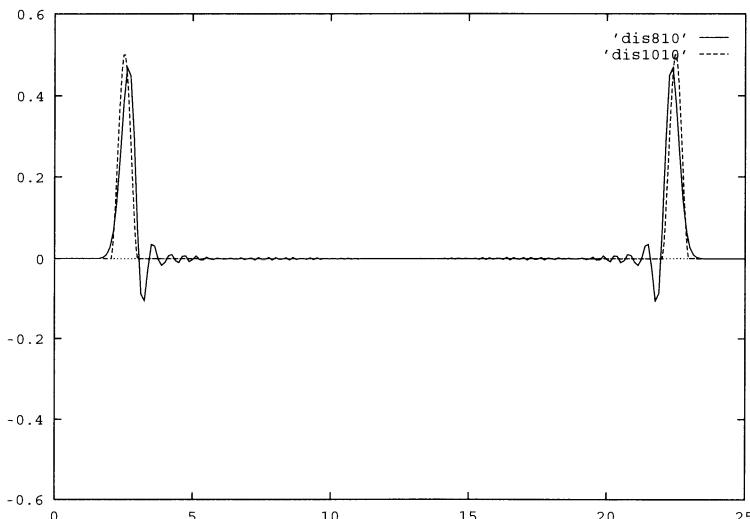


Fig. 7.1. Comparison of the solution obtained with a dispersive scheme (*continuous line*) and the exact solution (*dotted line*)

7.3 Order of the Numerical Dispersion

A natural way to use q_h is to study the leading term of its Taylor expansion in h . This term is generally of the same order as the scheme but, for some

approximations (such as finite element methods), superconvergence phenomena can appear. Anyhow, such a study is good for checking the method.

We now give these Taylor expansions for some semi-discrete and fully discrete schemes in 1D for the wave equation. In the following, q_h is given for the phase velocity.

7.3.1 Schemes Semi-Discrete in Space

We give below the Taylor expansions for the second-order and fourth-order approximations in space of the wave equation.

Second-Order.

$$\begin{aligned} q_h &= \frac{2}{kh} \sin \frac{kh}{2} \\ &= 1 - \frac{k^2 h^2}{24} + \frac{k^4 h^4}{1920} - \frac{k^6 h^6}{322560} + O(h^8). \end{aligned} \tag{7.13}$$

Fourth-Order: First Approach.

$$\begin{aligned} q_h &= \frac{2}{kh} \sqrt{\sin^2 \frac{kh}{2} + \frac{1}{3} \sin^4 \frac{kh}{2}} \\ &= 1 - \frac{k^4 h^4}{180} + \frac{k^6 h^6}{2016} + O(h^8). \end{aligned} \tag{7.14}$$

Fourth-Order: Second Approach.

$$\begin{aligned} q_h &= \frac{2}{kh} \sqrt{\sin^2 \frac{kh}{2} + \frac{1}{3} \sin^4 \frac{kh}{2} + \frac{1}{36} \sin^6 \frac{kh}{2}} \\ &= 1 - \frac{3}{640} k^4 h^4 + \frac{k^6 h^6}{3584} + O(h^8). \end{aligned} \tag{7.15}$$

One can see that the coefficient of the fourth-order term is slightly smaller for the second approach than for the first one.

7.3.2 Fully Discrete Schemes: Second-Order in Time

Here are the Taylor expansions in h and Δt for the previous schemes with a leapfrog approximation in time.

Second-Order.

$$\begin{aligned}
q_h &= \frac{2}{ck\Delta t} \arcsin \left(\frac{c\Delta t}{h} \sin \frac{kh}{2} \right) \\
&\simeq 1 - \frac{k^2 h^2}{24} + \frac{c^2 k^2 \Delta t^2}{24} + \frac{k^4 h^4}{1920} \\
&\quad - \frac{c^2 k^4 h^2 \Delta t^2}{192} + \frac{3}{640} c^4 k^4 \Delta t^4 + \dots,
\end{aligned} \tag{7.16}$$

which provides, after replacing Δt by $\frac{\alpha h}{c}$,

$$q_h = 1 - \frac{k^2 h^2}{24} (1 - \alpha^2) + \frac{k^4 h^4}{1920} (1 - 10\alpha^2 + 9\alpha^4) + O(h^6). \tag{7.17}$$

Fourth-Order: First Approach.

$$\begin{aligned}
q_h &= \frac{2}{ck\Delta t} \arcsin \left(\frac{c\Delta t}{h} \sqrt{\sin^2 \frac{kh}{2} + \frac{1}{3} \sin^4 \frac{kh}{2}} \right) \\
&\simeq 1 - \frac{k^4 h^4}{180} + \frac{c^2 k^2 \Delta t^2}{24} + \frac{3}{640} c^4 k^4 \Delta t^4 + \dots
\end{aligned} \tag{7.18}$$

So, by taking into account the relation $\Delta t = \frac{\alpha h}{c}$:

$$q_h = 1 + \frac{\alpha^2 k^2 h^2}{24} - k^4 h^4 \left(\frac{1}{180} - \frac{3}{640} \alpha^4 \right) + O(h^6). \tag{7.19}$$

Fourth-Order: Second Approach.

$$\begin{aligned}
q_h &= \frac{2}{ck\Delta t} \arcsin \left(\frac{c\Delta t}{h} \sqrt{\sin^2 \frac{kh}{2} + \frac{1}{3} \sin^4 \frac{kh}{2} + \frac{1}{36} \sin^6 \frac{kh}{2}} \right) \\
&\simeq 1 - \frac{3}{640} k^4 h^4 + \frac{c^2 k^2 \Delta t^2}{24} + \frac{3}{360} c^4 k^4 \Delta t^4 + \dots
\end{aligned} \tag{7.20}$$

which also gives, after setting $\Delta t = \alpha h/c$,

$$q_h = 1 + \frac{\alpha^2 k^2 h^2}{24} - \frac{3}{640} k^4 h^4 (1 - \alpha^4) + O(h^6). \tag{7.21}$$

Equations (7.19) and (7.21) show the second-order global accuracy of these schemes. However, by taking small values (generally $\alpha = \alpha_M/2$ is enough), one can reduce the effect of the second-order term.

7.3.3 Fully Discrete Schemes: The Modified Equation Approach

Of course, the use of a fourth-order approximation in time provides a global fourth-order accuracy.

Fourth-Order: First Approach.

$$\begin{aligned} q_h &= \frac{2}{ck\Delta t} \arcsin \left(\frac{c\Delta t}{h} \sqrt{\sin^2 \frac{kh}{2} + \frac{1}{3} \sin^4 \frac{kh}{2} \left(1 - \frac{c^2 \Delta t^2}{h^2} \right)} \right) \\ &\simeq 1 - \frac{k^4 h^4}{180} + \frac{c^2 k^4 h^2 \Delta t^2}{144} - \frac{c^4 k^4 \Delta t^4}{720} + \dots \end{aligned} \quad (7.22)$$

So, by replacing Δt by $\frac{\alpha h}{c}$, we obtain:

$$q_h = 1 - \frac{k^4 h^4}{720} (4 - 5\alpha^2 + \alpha^4) + O(h^6). \quad (7.23)$$

Fourth-Order: Second Approach.

$$\begin{aligned} q_h &= \frac{2}{ck\Delta t} \arcsin \left(\frac{c\Delta t}{h} \sqrt{\sin^2 \frac{kh}{2} + \frac{1}{3} \sin^4 \frac{kh}{2} \left(1 - \frac{c^2 \Delta t^2}{h^2} \right) + \frac{1}{36} \sin^6 \frac{kh}{2}} \right) \\ &\simeq 1 - \frac{3}{640} k^4 h^4 + \frac{c^2 k^4 h^2 \Delta t^2}{144} - \frac{c^4 k^4 \Delta t^4}{720} + \dots \end{aligned} \quad (7.24)$$

which provides, finally:

$$q_h = 1 - \frac{1}{5760} (27 - 40\alpha^2 + 8\alpha^4) k^4 h^4 + O(h^6). \quad (7.25)$$

In the two cases above, one can see that a second-order approximation of the corrective term provides a fourth-order global accuracy.

7.3.4 Fully Discrete Schemes: The Fourth-Order Symmetric Scheme

In this case, we have two values of ω_h derived (for $\theta = 0$) from (6.59a) and (6.59b). One of them corresponds to the physical wave and the other is the pulsation of a numerical parasitic wave. The corresponding numerical dispersion coefficients are given by

$$q_{h1} = \frac{2}{ck\Delta t} \arcsin(\sqrt{\lambda_1}), \quad (7.26a)$$

$$q_{h2} = \frac{2}{ck\Delta t} \arcsin(\sqrt{\lambda_2}). \quad (7.26b)$$

Now, we give the Taylor expansions of these coefficients for the two fourth-order schemes obtained by the first and second approaches.

Fourth-Order: First Approach.

$$\begin{aligned}
q_{h1} &\simeq \frac{2\pi}{3ck\Delta t} \\
&+ \left(\frac{1}{12} - \frac{1}{1080} k^4 h^4 + \frac{1}{12096} k^6 h^6 + O(h^8) \right) \sqrt{3} ck \Delta t \\
&+ \left(\frac{1}{32} - \frac{1}{1440} k^4 h^4 + \frac{1}{16128} k^6 h^6 + O(h^8) \right) \sqrt{3} c^3 k^3 \Delta t^3 \\
&+ \dots, \\
q_{h2} &\simeq 1 - \frac{k^4 h^4}{180} - \frac{7}{1440} c^4 k^4 \Delta t^4 \\
&+ \frac{k^6 h^6}{2016} - \frac{41}{12096} c^6 k^6 \Delta t^6 + \dots
\end{aligned} \tag{7.28}$$

Fourth-Order: Second Approach.

$$\begin{aligned}
q_{h1} &\simeq \frac{2\pi}{3ck\Delta t} \\
&+ \left(\frac{1}{12} - \frac{1}{1280} k^4 h^4 + \frac{1}{21504} k^6 h^6 + O(h^8) \right) \sqrt{3} ck \Delta t \\
&+ \left(\frac{1}{32} - \frac{3}{5120} k^4 h^4 + \frac{1}{28672} k^6 h^6 + O(h^8) \right) \sqrt{3} c^3 k^3 \Delta t^3 \\
&+ \dots, \\
q_{h2} &\simeq 1 - \frac{3}{640} k^4 h^4 - \frac{7}{1440} c^4 k^4 \Delta t^4 \\
&+ \frac{k^6 h^6}{3584} - \frac{41}{12096} c^6 k^6 \Delta t^6 + \dots
\end{aligned} \tag{7.29}$$

Obviously, q_{h1} corresponds to a parasitic wave whose velocity tends to infinity when Δt tends to zero. In principle, such parasitic waves have an amplitude decreasing with Δt . On the other hand, (7.28) and (7.30) show the fourth-order character of this approximation in time.

The Taylor expansion of the second root of (6.58) when $\theta \neq 0$ provides, for example, for the first approach:

$$\begin{aligned}
q_{h2} &\simeq 1 - \frac{k^4 h^4}{180} - \left(\frac{7}{1440} + \frac{1}{144} \theta \right) c^4 k^4 \Delta t^4 \\
&+ \frac{k^6 h^6}{2016} - \left(\frac{41}{12096} + \frac{7}{1728} \theta + \frac{1}{864} \theta^2 \right) c^6 k^6 \Delta t^6 + \dots
\end{aligned} \tag{7.31}$$

which shows that the truncation error grows linearly with θ and, therefore, the numerical dispersion of the scheme. This is why we cannot use values of θ which are too large, although the stability of the scheme increases with this parameter. This remark holds for the second approach too.

7.3.5 Error Committed on the Group Velocities

We shall close this section by giving below the Taylor expansions of the numerical dispersion for the different approaches, semi-discrete in space in 1D.

Second-Order.

$$\begin{aligned} q_h &= \frac{\sin \frac{kh}{2} \cos \frac{kh}{2}}{\sqrt{1 - \cos^2 \frac{kh}{2}}} \\ &\simeq 1 - \frac{k^2 h^2}{8} + \frac{k^4 h^4}{384} - \frac{k^6 h^6}{46080} + O(h^8). \end{aligned} \quad (7.32)$$

Fourth-Order: First Approach.

$$\begin{aligned} q_h &= \frac{\sqrt{3}}{3} \frac{\sin \frac{kh}{2} \cos \frac{kh}{2} \left(-5 + 2 \cos^2 \frac{kh}{2} \right)}{\sqrt{1 - 5 \cos^2 \frac{kh}{2} + \cos^4 \frac{kh}{2}}} \\ &\simeq 1 - \frac{k^4 h^4}{36} + \frac{k^6 h^6}{288} + O(h^8). \end{aligned} \quad (7.33)$$

Fourth-Order: Second Approach.

$$\begin{aligned} q_h &= \frac{1}{2} \frac{\sin \frac{kh}{2} \cos \frac{kh}{2} \left(\cos^2 \frac{kh}{2} - 3 \right)}{\sqrt{1 - \cos^2 \frac{kh}{2}}} \\ &\simeq 1 - \frac{3}{128} k^4 h^4 + \frac{k^6 h^6}{512} + O(h^8). \end{aligned} \quad (7.34)$$

If we compare the Taylor expansions of the numerical dispersion coefficients for the phase and group velocities, we can check that the second-order term is multiplied by 3 for the group velocity, the fourth-order term by 5 and the sixth-order scheme by 7.

This phenomenon is easily understandable in 1D since, if $q_h = 1 + Ck^{2n}h^{2n} + O(h^{2n+2})$ for the phase velocity, we can write: $\omega_h = k + Ck^{2n+1}h^{2n} + O(h^{2n+2})$. Then $\omega'_h = 1 + (2n+1)Ck^{2n}h^{2n} + O(h^{2n+2})$ and the numerical dispersion of the group velocity is such that $q_h = 1 + (2n+1)Ck^{2n}h^{2n} + O(h^{2n+2})$.

Remark

The Taylor expansion of the symbols of the discrete operators are actually the Fourier transforms of the Taylor expansions of the operators applied to a function u divided by the Fourier transform of u . For example, we have, for the first approach in 1D:

$$\begin{aligned}\Delta_h u(x_\ell) &= \frac{1}{h^2} \left(-\frac{1}{12}u(x_{\ell-2}) + \frac{4}{3}u(x_{\ell-1}) - \frac{5}{2}u(x_\ell) \right. \\ &\quad \left. + \frac{4}{3}u(x_{\ell+1}) - \frac{1}{12}u(x_{\ell+2}) \right) = \frac{\partial^2 u}{\partial x^2}(x_\ell) \\ &\quad - \frac{1}{90} \frac{\partial^6 u}{\partial x^6}(x_\ell) h^4 + O(h^6),\end{aligned}\tag{7.35}$$

whose Fourier transform divided by \hat{u} is

$$\widehat{\Delta}_h = \frac{4}{h^2} \left(\sin^2 \frac{kh}{2} + \frac{1}{3} \sin^4 \frac{kh}{2} \right) = k^2 - \frac{k^6 h^4}{90} + O(h^6).\tag{7.36}$$

7.4 Change of Variables

As we saw in the previous sections, the relations between the numerical pulsation ω_h and the symbol of the difference operator in space is, in the case of a semi-discrete approximation in space, of the form

$$\omega_h = \frac{1}{h} \sqrt{\tilde{D}_h(\mathbf{k}h)},\tag{7.37}$$

where $\tilde{D}_h = h^2 D_h$ and D_h is the symbol of the discrete operator. From (7.37), we obtain the general form of the numerical dispersion of a finite difference scheme:

$$q_h = \frac{1}{c|\mathbf{k}|h} \sqrt{\tilde{D}_h(\mathbf{k}h)}.\tag{7.38}$$

In the fully discrete case, we have:

$$\omega_h = \frac{2}{\Delta t} f(\alpha^2 \tilde{D}_h(\mathbf{k}h, \alpha)),\tag{7.39}$$

where $\alpha = c\Delta t/h$ and, therefore:

$$q_h = \frac{2}{c|\mathbf{k}|\Delta t} f(\alpha^2 \tilde{D}_h(\mathbf{k}h, \alpha)). \quad (7.40)$$

In order to study the error committed on the velocity on the wave versus significant physical and numerical quantities involved in the propagation of a wave by a numerical scheme, we set the following changes of variables:

– In 1D:

$$kh = 2\pi K. \quad (7.41)$$

– In 2D:

$$\begin{cases} k_1 h = 2\pi K \cos \varphi, \\ k_2 h = 2\pi K \sin \varphi. \end{cases} \quad (7.42)$$

– In 3D:

$$\begin{cases} k_1 h = 2\pi K \sin \varphi \cos \theta, \\ k_2 h = 2\pi K \sin \varphi \sin \theta, \\ k_3 h = 2\pi K \cos \varphi. \end{cases} \quad (7.43)$$

Obviously, φ and θ determine the direction of propagation of the wave. The physical meaning of K is less evident.

As we said before, a plane wave with a pulsation and a velocity equal to ω and c can be characterized by

-its frequency: $f = \frac{\omega}{2\pi}$,

-its wavelength: $\lambda = \frac{c}{f} = \frac{2\pi c}{\omega}$.

By using the dispersion relation of the continuous wave equation $\omega^2 = c^2 |\mathbf{k}|^2$, we can write, for $\omega > 0$:

$$\lambda = \frac{2\pi}{|\mathbf{k}|}. \quad (7.44)$$

Let N_λ be the number of points of a mesh per wavelength in a given direction of space. We have:

$$N_\lambda = \frac{\lambda}{h} = \frac{2\pi}{|\mathbf{k}|h} = \frac{1}{K}. \quad (7.45)$$

On the basis of this change of variables, we obtain for (7.38):

– In 1D:

$$q_h = \frac{1}{\pi K} \sqrt{\tilde{D}_h(K)}. \quad (7.46)$$

– In 2D:

$$q_h = \frac{1}{\pi K} \sqrt{\tilde{D}_h(K, \varphi)}. \quad (7.47)$$

– In 3D:

$$q_h = \frac{1}{\pi K} \sqrt{\tilde{D}_h(K, \varphi, \theta)}. \quad (7.48)$$

For (7.40), one must first notice that

$$c|\mathbf{k}|\Delta t = \alpha|\mathbf{k}|h = 2\pi\alpha \quad (7.49)$$

and, therefore:

– In 1D:

$$q_h = \frac{1}{\pi\alpha K} f(\alpha^2 \tilde{D}_h(K, \alpha)). \quad (7.50)$$

– In 2D:

$$q_h = \frac{1}{\pi\alpha K} f(\alpha^2 \tilde{D}_h(K, \varphi, \alpha)). \quad (7.51)$$

– In 3D:

$$q_h = \frac{1}{\pi\alpha K} f(\alpha^2 \tilde{D}_h(K, \varphi, \theta, \alpha)). \quad (7.52)$$

The above forms of q_h enable us to draw two kinds of curves which characterize a numerical scheme:

- for given α , φ and θ , we obtain the *dispersion curves* of the scheme, which indicate the error committed on the velocity versus the inverse of the number of points per wavelength,
- for given α and K , we obtain the *isotropy curves* (in 2D or also anisotropy surfaces in 3D) of the scheme which visualize the error committed on the velocity versus the direction of propagation.

7.5 Some Useful Properties of the Schemes

7.5.1 Relation between 1D and Higher Dimensions

The symmetric character of the schemes used to approximate the second order derivatives implies that the symbols of the discrete operators can be expressed as a linear combination of cosines of ℓkh and, therefore, as a linear combination of even powers of sines of $\ell kh/2$.

So, let

$$D_{h,1}(kh) = \sum_{\ell=1}^{N_D} \lambda_\ell \sin^{2\ell} \frac{kh}{2} \quad (7.53)$$

be the symbol of the discrete operator approximating a second-order derivative in space.

The symbol for $\Delta = \sum_{j=1}^d \partial^2 / \partial x_j^2$ is:

$$D_{h,d}(\mathbf{k}h) = \sum_{j=1}^d \sum_{\ell=1}^{N_D} \lambda_\ell \sin^{2\ell} \frac{k_j h}{2}. \quad (7.54)$$

Now, when the waves propagate colinearly to the vector whose all components are equal to 1, we can set $k_j = k, \forall j = 1..d$ and we obtain:

$$D_{h,d}(\mathbf{k}h) = d \sum_{\ell=1}^{N_D} \lambda_\ell \sin^{2\ell} \frac{kh}{2}. \quad (7.55)$$

Relation (7.38) shows that the numerical dispersion of the scheme semi-discrete in space in that direction can be written as

$$q_h = \frac{1}{c|\mathbf{k}|h} \sqrt{d \sum_{\ell=1}^{N_D} \lambda_\ell \sin^{2\ell} \frac{kh}{2}}. \quad (7.56)$$

Since, in this case, $|\mathbf{k}| = \sqrt{dk}$, we have:

$$q_h = \frac{1}{ckh} \sqrt{\sum_{\ell=1}^{N_D} \lambda_\ell \sin^{2\ell} \frac{kh}{2}}, \quad (7.57)$$

which provides the numerical dispersion of the discrete operator in 1D.

In particular, in 2D, the numerical dispersion along the diagonal is the same as that of the one-dimensional case.

Remark

The symbol of the operator of the fourth-order modified equation approach (or the second-order scheme in space and time) in dD is of the form:

$$q_h = \frac{1}{\alpha|\mathbf{k}|h} \arcsin \left(\alpha \sqrt{\sum_{j=1}^d \sum_{\ell=1}^{N_D} \lambda_\ell \sin^{2\ell} \frac{k_j h}{2} + \mu \frac{\alpha^2}{3} \left(\sum_{j=1}^d \sin^2 \frac{k_j h}{2} \right)^2} \right). \quad (7.58)$$

By setting $\alpha' = \alpha\sqrt{d}$, we obtain, for $k_j = k, \forall j = 1..d$:

$$q_h = \frac{1}{\alpha'kh} \sqrt{\sum_{\ell=1}^{N_D} \lambda_\ell \sin^{2\ell} \frac{kh}{2} + \mu \frac{\alpha'^2}{3} \sin^4 \frac{kh}{2}}, \quad (7.59)$$

which is the numerical dispersion of the scheme in 1D.

7.5.2 Two Remarkable Schemes

After applying the change of variables, the numerical dispersion of the second-order approximation in space and time can be written as

$$q_h = \frac{1}{\pi\alpha K} \arcsin(\alpha \sin(\pi K)). \quad (7.60)$$

For $\alpha = \alpha_M = 1$, this relation becomes

$$q_h = \frac{1}{\pi K} \arcsin(\sin(\pi K)) = 1. \quad (7.61)$$

In other words, we have, in this case: $c_h = c$ and no error is committed on the velocity.

Similarly, the numerical dispersion of the fourth-order scheme in 1D obtained by the first approach in space and the modified equation approach in time can be written as

$$q_h = \frac{1}{\pi\alpha K} \arcsin \left(\alpha \sqrt{\sin^2(\pi K) + \frac{1}{3}(1 - \alpha^2) \sin^4(\pi K)} \right) \quad (7.62)$$

and we obtain, for $\alpha = \alpha_M = 1$:

$$q_h = \frac{1}{\pi K} \arcsin(\sin(\pi K)) = 1. \quad (7.63)$$

Remarks

1. The positivity of the sine function under the square root derives from the fact that we are only interested in $0 \leq K \leq 1/2$.
2. This result is actually more general. For the second-order scheme, we have:

$$\frac{u_j^{n+1} - 2u_j^n + u_j^{n-1}}{\Delta t^2} - c^2 \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} = 0. \quad (7.64)$$

For $c\Delta t = h$, this scheme provides:

$$u_j^{n+1} + u_j^{n-1} = u_{j+1}^n + u_{j-1}^n. \quad (7.65)$$

This characteristics scheme gives an exact solution.

A similar computation leads to the same characteristics scheme for the fourth-order modified equation approach with the first approach in space. As we said before, the slight difference between the dispersions relations of the first and second approaches is enough to destroy this remarkable property in the second approach.

7.6 Dispersion Curves

In Figs. 7.2–7.10, we show dispersion curves given by (7.51) for various schemes. The reciprocal of the number of points per wavelength K is given along the abscissae and the numerical dispersion q_h along the ordinates.

For the schemes semi-discrete in space, we shall give the dispersion curves corresponding to the phase and group velocities for $\varphi = 0, \pi/12, \pi/6, \pi/4$.

For the fully discrete schemes, we shall give the dispersion curves corresponding to the phase velocity for α varying from 0 to α_M in steps of 0.1, for $\varphi = 0$ and $\varphi = \pi/4$.

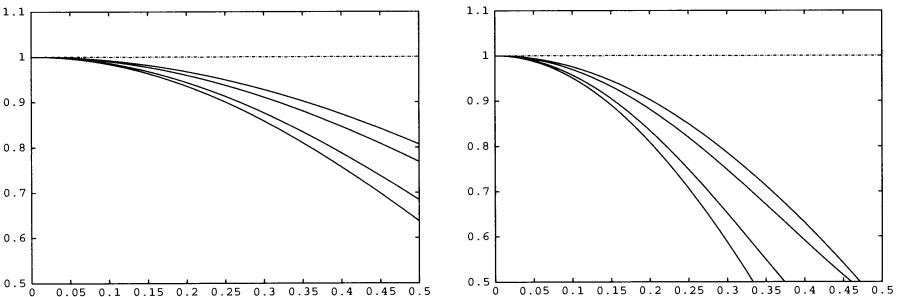


Fig. 7.2. Dispersion curves for the second-order scheme semi-discrete in space for the phase (left) and the group (right) velocities. The curves are given from $\varphi = 0$ (lower curves) to $\varphi = \pi/4$ (upper curves)

7.6.1 Second and Fourth-Order Schemes, Semi-Discrete in Space

In Figs. 7.2–7.4, we give the dispersion curves of second-order and fourth-order approximations in space for the phase and group velocities. We can first notice that, as expected, the group velocities provide a larger but more realistic numerical dispersion than the phase velocities. On the other hand, the gain of accuracy realized by fourth-order schemes over second-order schemes is obvious. Finally, Figs. 7.3 and 7.4 show that the additional term which appears in the dispersion relation of the second approach in space does not provide a significant change versus the first approach.

7.6.2 Schemes, Second-Order in Time and Fourth-Order in Space

In Figs. 7.5–7.7, we give the dispersion curves for the second-order approximations in time. All the curves increase with α but only the curves provided

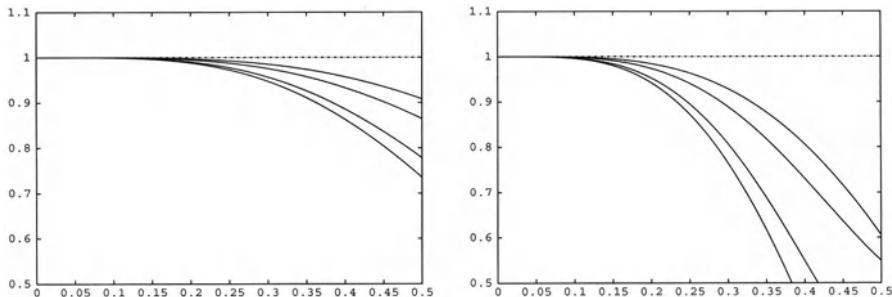


Fig. 7.3. Dispersion curves for the fourth-order scheme semi-discrete in space (first approach) for the phase (*left*) and the group (*right*) velocities. The curves are given from $\varphi = 0$ (*lower curves*) to $\varphi = \pi/4$ (*upper curves*)

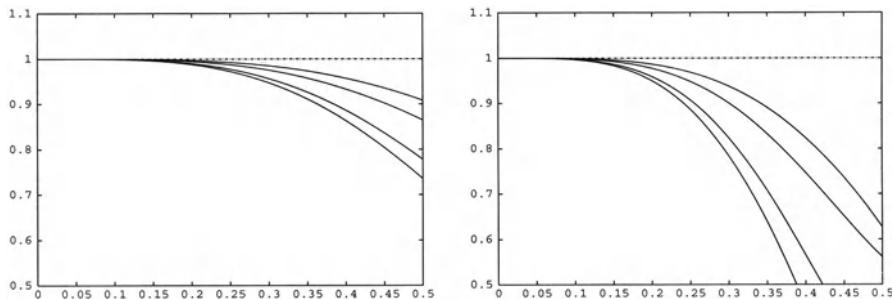


Fig. 7.4. Dispersion curves for the fourth-order scheme semi-discrete in space (second approach) for the phase (*left*) and the group (*right*) velocities. The curves are given from $\varphi = 0$ (*lower curves*) to $\varphi = \pi/4$ (*upper curves*)

by the second-order in space tend to 1 with this parameter. The curves corresponding to fourth-order approximations become greater than 1 when α tends to α_M and, therefore, are less accurate for large values of α .

7.6.3 Schemes, Fourth-Order in Time and Space

In Figs. 7.8–7.10, we give the dispersion curves for the fourth-order approximations in time. The first approach with the modified equation approach is obviously the best one. Its accuracy increases with α and we obtain no error along the diagonal for the maximum value of α . The second approach has the same features except for the perfect accuracy for α_M along the diagonal. However, its accuracy is very good. The symmetric schemes does not have the same features. Their accuracy decreases slightly when α_M increases and remains around the accuracy of the scheme semi-discrete in space.

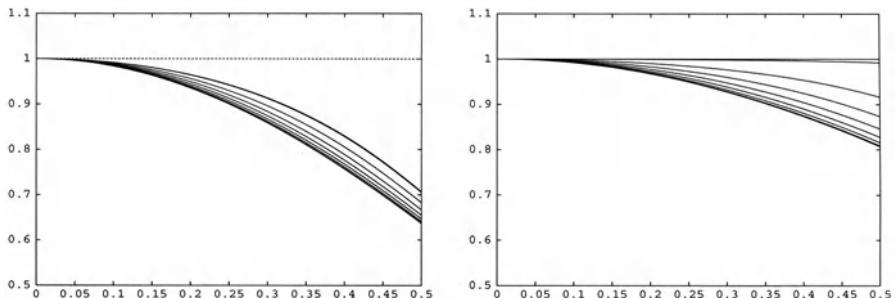


Fig. 7.5. Dispersion curves for the scheme, second-order in time and space for $\varphi = 0$ (left) to $\varphi = \pi/4$ (right). The curves are drawn for α varying from 0 to $\alpha_M \approx 0.707\,106$. For $\alpha = \alpha_M$ and $\varphi = \pi/4$, the curve is constant and equal to 1

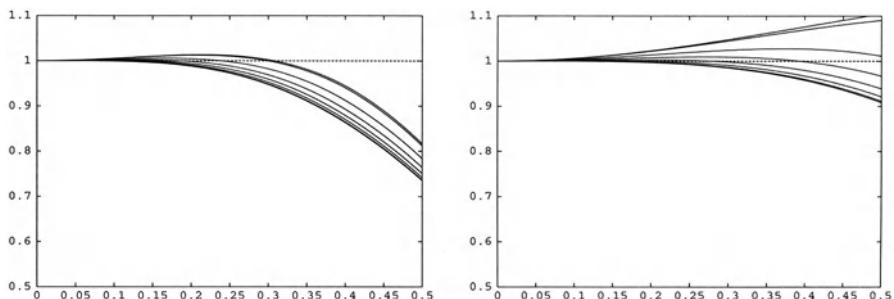


Fig. 7.6. Dispersion curves for the scheme, second-order in time and fourth-order in space (first approach) for $\varphi = 0$ (left) to $\varphi = \pi/4$ (right). The curves are drawn for α varying from 0 to $\alpha_M \approx 0.612\,372$ in steps of 0.1

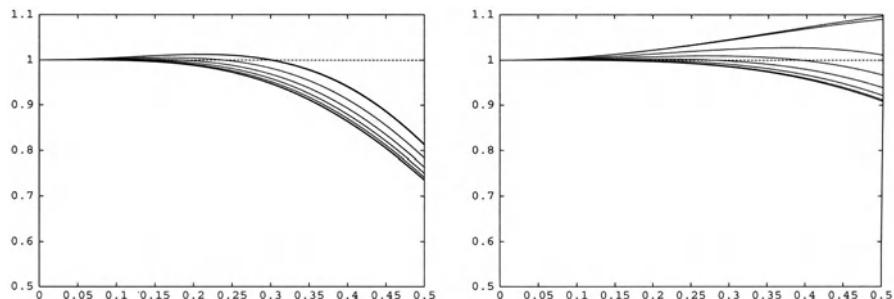


Fig. 7.7. Dispersion curves for the scheme, second-order in time and fourth-order in space (second approach) for $\varphi = 0$ (left) to $\varphi = \pi/4$ (right). The curves are drawn for α varying from 0 to $\alpha_M \approx 0.606\,091\,5$ in steps of 0.1

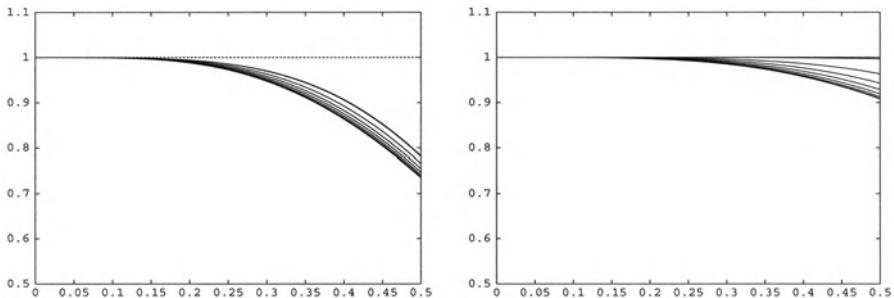


Fig. 7.8. Dispersion curves for the scheme, fourth-order in time and space (first approach) for $\varphi = 0$ (left) to $\varphi = \pi/4$ (right). The curves are drawn for α varying from 0 to $\alpha_M \simeq 0.707\,106$ in steps of 0.1. For $\alpha = \alpha_M$ and $\varphi = \pi/4$, the curve is constant and equal to 1

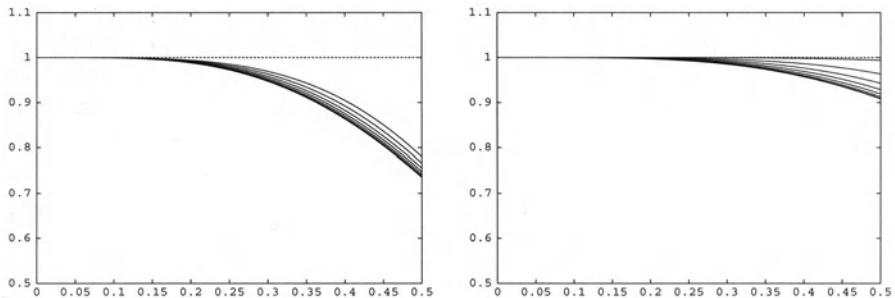


Fig. 7.9. Dispersion curves for the scheme, fourth-order in time and space (second approach) for $\varphi = 0$ (left) to $\varphi = \pi/4$ (right). The curves are drawn for α varying from 0 to $\alpha_M \simeq 0.707\,106$ in steps of 0.1. In this case, for $\alpha = \alpha_M$ and $\varphi = \pi/4$, the curve is close to 1 but not uniformly equal to 1

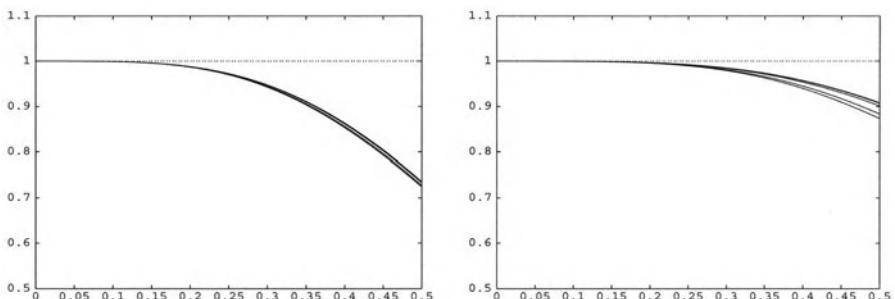


Fig. 7.10. Dispersion curves for the scheme, fourth-order in time and space (symmetric scheme with $\theta = 0$, first approach in space) for $\varphi = 0$ (left) to $\varphi = \pi/4$ (right). The curves are drawn for α varying from 0 to $\alpha_M \simeq 0.433\,012\,7$ in steps of 0.1

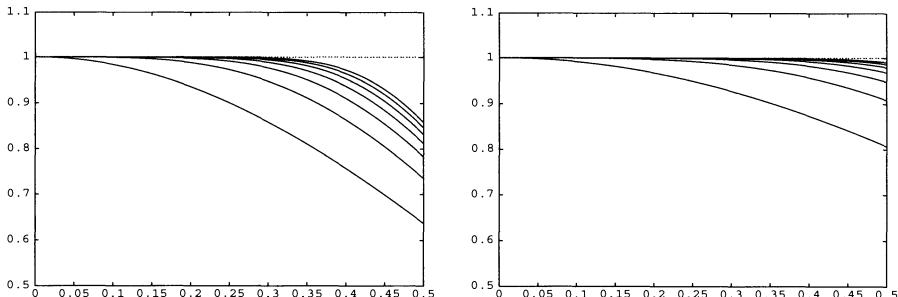


Fig. 7.11. Comparison of dispersion curves for finite difference approximations in space whose orders go from 2 to 14 for $\varphi = 0$ (left) to $\varphi = \pi/4$ (right). Of course, the curves tend to 1 with the order of the approximation

7.6.4 Comparison with Higher-Order Schemes in Space

Now, we are going to compare the semi-discrete approximations in space for orders from 2 to 14 when $\varphi = 0$ and $\varphi = \pi/4$. Figure 7.11 shows that, obviously, for all the orders, the error committed along the diagonal is smaller than that along the axis. It is however interesting to notice that the gain of accuracy decreases with the order of the scheme. This shows that one must find a good balance between the accuracy and the order of the scheme whose cost grows with the order. It seems that this balance lies around the 8th or 10th-order schemes.

7.7 Isotropy Curves

Of course, for the isotropic phenomenon given by the wave equation, the value of the velocity of the wave does not depend on the direction of propagation. However, as we saw in the previous section, the accuracy of its approximation can depend on this direction, since the approximated velocity depends on \mathbf{k} . A good way to visualize the anisotropy induced by the approximation is to draw the numerical dispersion q_h in polar coordinates versus the angle φ . In Fig. 7.12 we give these curves for the second-order and fourth-order (first approach) approximations in space. The curves are drawn from $N_\lambda = 1/K = 2$ to $N_\lambda = 1/K = \infty$. The curves in both figures are drawn for $N_\lambda = 2, 3, 4, 5, 8, 10, \infty (\simeq 10\,000)$. Of course, the isotropy becomes better when the number of points per wavelength increases and the last curve is a circle of radius equal to 1.

7.8 The Elastics System

As we saw in Sect. 5.4, the elastics system has two dispersion relations given by (5.79a) and (5.79b) which provide two values, ω_h^P and ω_h^S , of ω_h defined

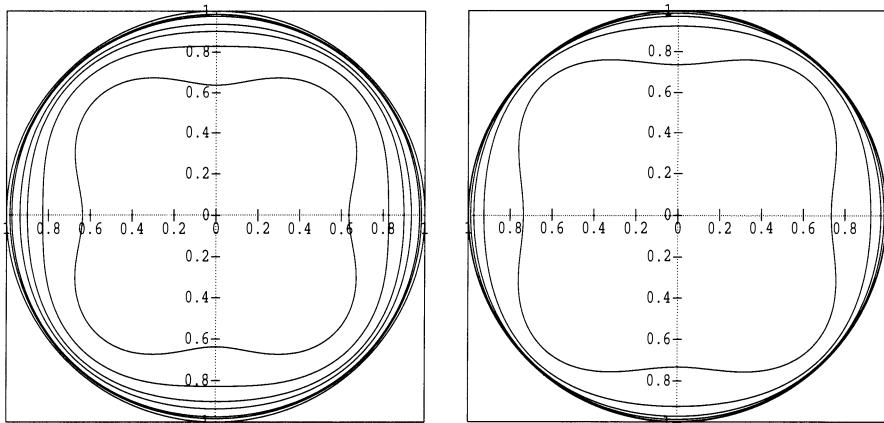


Fig. 7.12. Isotropy curves for a semi-discretization in space of second-order (*left*) and fourth-order (first approach) (*right*). The curves are given for $N_\lambda = 2$ (*interior curve*) to $N_\lambda = \infty$ (*exterior circle*)

by:

$$\omega_h^P = \frac{2}{\Delta t} \arcsin \left(\left(\frac{\Delta t^2}{\rho h^2} ((\lambda + 2\mu)(\sin^2 \frac{k_1 h}{2} + \sin^2 \frac{k_2 h}{2}) - (\lambda + \mu) \sin^2 \frac{k_1 h}{2} \sin^2 \frac{k_2 h}{2}) \right)^{1/2} \right), \quad (7.66a)$$

$$\omega_h^S = \frac{2}{\Delta t} \arcsin \left(\left(\frac{\Delta t^2}{\rho h^2} (\mu(\sin^2 \frac{k_1 h}{2} + \sin^2 \frac{k_2 h}{2}) + (\lambda + \mu) \sin^2 \frac{k_1 h}{2} \sin^2 \frac{k_2 h}{2}) \right)^{1/2} \right). \quad (7.66b)$$

By dividing ω_h^P and ω_h^S by ω_P and ω_S defined in (3.25) and (3.26), we obtain the two numerical dispersions coefficients corresponding to the *P*-wave and the *S*-wave:

$$q_h^P = \frac{\omega_h^P}{|\mathbf{k}|V_P}, \quad (7.67a)$$

$$q_h^S = \frac{\omega_h^S}{|\mathbf{k}|V_S}. \quad (7.67b)$$

In this case, we are going to express q_h^P and q_h^S versus the variable K and the following non-dimensional quantities:

$$\alpha' = \frac{\Delta t}{h} V_S, \quad (7.68)$$

$$\nu = \frac{\lambda}{2(\lambda + \mu)}, \quad (7.69)$$

where ν is the Poisson coefficient of the equation, which varies from 0 to 0.5 and measures the contrast between the two velocities V_P and V_S . Actually, this coefficient is connected to V_P and V_S in the following way:

$$\nu = \frac{\lambda}{2(\lambda + \mu)} = \frac{1}{2} - \frac{\mu}{2(\lambda + \mu)} = \frac{1}{2} - \frac{1}{2\frac{\lambda + \mu}{\mu}} = \frac{1}{2} - \frac{1}{2\frac{\lambda + 2\mu}{\mu} - 2}.$$

So, since $(\lambda + 2\mu)/\mu = V_P^2/V_S^2$, we finally obtain:

$$\nu = \frac{1}{2} \left(1 - \frac{1}{\frac{V_P^2}{V_S^2} - 1} \right) \quad (7.70)$$

and, conversely:

$$\frac{V_P^2}{V_S^2} = \frac{2(1 - \nu)}{1 - 2\nu} = \frac{\nu_1}{\nu_2}. \quad (7.71)$$

Now, by noticing that

$$\frac{\lambda + \mu}{\mu} = \frac{\lambda + 2\mu}{\mu} - 1 = \frac{V_P^2}{V_S^2} - 1 = \frac{1}{\nu_2}$$

and, taking into account relations (7.68)–(7.71), we obtain the following expressions of q_h^P and q_h^S :

$$q_h^P = \frac{\sqrt{\nu_2}}{\alpha' \sqrt{\nu_1} \pi K} \arcsin \left(\alpha' \sqrt{\frac{\nu_1}{\nu_2}} \left((X_1 + X_2) - \frac{1}{\nu_1} X_1 X_2 \right)^{1/2} \right), \quad (7.72a)$$

$$q_h^S = \frac{1}{\alpha' \pi K} \arcsin \left(\alpha' \left(X_1 + X_2 + \frac{1}{\nu_2} X_1 X_2 \right)^{1/2} \right), \quad (7.72b)$$

where $X_1 = \sin^2(\pi K \cos \varphi)$ and $X_2 = \sin^2(\pi K \sin \varphi)$.

With the above notation, the stability condition can be written as

$$\alpha' \leq \alpha'_M = \sqrt{\frac{1 - 2\nu}{3 - 4\nu}}. \quad (7.73)$$

We now give, in Figs. 7.13 and 7.14, some dispersion curves derived from q_h^P and q_h^S for different values of ν and φ .

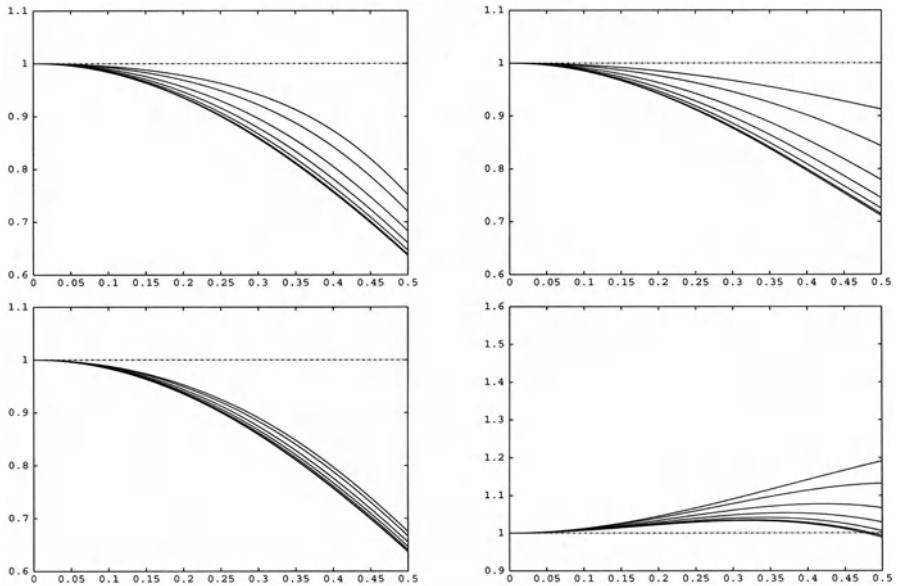


Fig. 7.13. Dispersion curves in V_P (above) and V_S (below) for the second-order approximation of the elastics system for $\varphi = 0$ (left) and $\varphi = \pi/4$ (right). Here $\nu = 0.1$ and the different curves increase with α' from 0 to $\alpha'_M \simeq 0.554\,700\,1$ in steps of 0.1

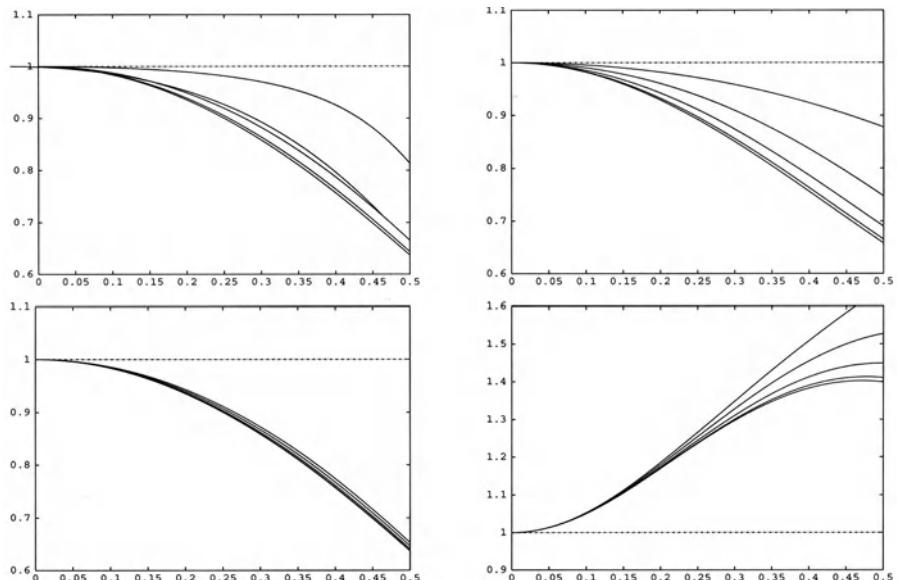


Fig. 7.14. Dispersion curves in V_P (above) and V_S (below) for the second-order approximation of the elastics system for $\varphi = 0$ (left) and $\varphi = \pi/4$ (right). Here $\nu = 0.4$ and the different curves increase with α' from 0 to $\alpha'_M \simeq 0.377\,964\,4$ in steps of 0.1

8. Construction of the Schemes in Heterogeneous Media

8.1 A General Framework

As we saw in Chap. 4, the second-order approximation as well as the fourth-order approximation in space obtained by the second approach led to the following decomposition of the discrete Laplace operator:

$$\Delta_{h,p}^{(q)} = \tilde{D}_{p,h}^{(q)} \circ D_{p,h}^{(q)} = -(D_{p,h}^{(q)})^* \circ D_{p,h}^{(q)}, \quad (8.1)$$

where p is the dimension of space and q is the order of the approximation.

Let us extend this approach to the heterogeneous, second-order operator $\text{div} \circ (\gamma(\mathbf{x})\mathbf{grad})$.

A natural approximation of this operator is given by:

$$\begin{aligned} \tilde{\Delta}_{h,p}^{(q)} = [\text{div} \circ (\gamma(\mathbf{x})\mathbf{grad})]_h &= \tilde{D}_{p,h}^{(q)} \circ (\gamma(\mathbf{x})D_{p,h}^{(q)}) \\ &= -(D_{p,h}^{(q)})^* \circ (\gamma(\mathbf{x})D_{p,h}^{(q)}). \end{aligned} \quad (8.2)$$

This factorization will be the basis of the construction of the discrete operators. Once the general framework of the approximation defined, we give the actual discrete operators.

8.2 Second-Order Approximations

As in the homogeneous case, the simplest and most basic operator is the 1D second-order approximation. Since we have:

$$(\Delta_{h,1}^{(2)} u)_\ell = \frac{u_{\ell+1} - 2u_\ell + u_{\ell-1}}{h^2} = \frac{\frac{u_{\ell+1} - u_\ell}{h} - \frac{u_\ell - u_{\ell-1}}{h}}{h}, \quad (8.3)$$

an obvious expression of $\tilde{\Delta}_{h,1}^{(2)}$ is given by

$$(\tilde{\Delta}_{h,1}^{(2)} u)_\ell = \frac{1}{h} \left(\gamma_{\ell+\frac{1}{2}} \frac{u_{\ell+1} - u_\ell}{h} - \gamma_{\ell-\frac{1}{2}} \frac{u_\ell - u_{\ell-1}}{h} \right), \quad (8.4)$$

where $\gamma_{\ell+1/2}$ is the value of γ at the point $x_{\ell+1/2} = (\ell + 1/2)h$.

The extension of (8.4) to the 2D case can easily be derived in the following form:

$$\begin{aligned} & (\tilde{\Delta}_{h,2}^{(2)} u)_{\ell,m} = \\ & \frac{1}{h} \left(\gamma_{\ell+\frac{1}{2},m} \frac{u_{\ell+1,m} - u_{\ell,m}}{h} - \gamma_{\ell-\frac{1}{2},m} \frac{u_{\ell,m} - u_{\ell-1,m}}{h} \right. \\ & \left. + \gamma_{\ell,m+\frac{1}{2}} \frac{u_{\ell,m+1} - u_{\ell,m}}{h} - \gamma_{\ell,m-\frac{1}{2}} \frac{u_{\ell,m} - u_{\ell,m-1}}{h} \right), \end{aligned} \quad (8.5)$$

where $\gamma_{\ell+1/2,m} = \gamma((\ell + 1/2)h, mh)$ and $\gamma_{\ell,m+1/2} = \gamma(\ell h, (m + 1/2)h)$.

In the same way, we obtain

$$\begin{aligned} & (\tilde{\Delta}_{h,3}^{(2)} u)_{\ell,m} = \\ & \frac{1}{h} \left(\gamma_{\ell+\frac{1}{2},m,n} \frac{u_{\ell+1,m,n} - u_{\ell,m,n}}{h} - \gamma_{\ell-\frac{1}{2},m,n} \frac{u_{\ell,m,n} - u_{\ell-1,m,n}}{h} \right. \\ & \left. + \gamma_{\ell,m+\frac{1}{2},n} \frac{u_{\ell,m+1,n} - u_{\ell,m,n}}{h} - \gamma_{\ell,m-\frac{1}{2},n} \frac{u_{\ell,m,n} - u_{\ell,m-1,n}}{h} \right. \\ & \left. + \gamma_{\ell,m,n+\frac{1}{2}} \frac{u_{\ell,m,n+1} - u_{\ell,m,n}}{h} - \gamma_{\ell,m,n-\frac{1}{2}} \frac{u_{\ell,m,n} - u_{\ell,m,n-1}}{h} \right), \end{aligned} \quad (8.6)$$

where $\gamma_{\ell+1/2,m,n} = \gamma((\ell + 1/2)h, mh, nh)$, $\gamma_{\ell,m+1/2,n} = \gamma(\ell h, (m + 1/2)h, nh)$ and $\gamma_{\ell,m,n+1/2} = \gamma(\ell h, mh, (n + 1/2)h)$.

8.3 Higher-Order Approximations: First Approach

As in the homogeneous case, we have, for this approach:

$$\tilde{\Delta}_{h,p}^{(2q)} = \sum_{j=1}^q \tilde{\Delta}_{jh,p}^{(2)}. \quad (8.7)$$

So, in order to obtain these schemes, we must first define $\tilde{\Delta}_{jh,1}^{(2)}$. Since

$$(\Delta_{jh,1}^{(2)} u)_\ell = \frac{u_{\ell+j} - 2u_\ell + u_{\ell-j}}{(jh)^2} = \frac{\frac{u_{\ell+j} - u_\ell}{jh} - \frac{u_\ell - u_{\ell-j}}{jh}}{jh}, \quad (8.8)$$

we obtain, here also:

$$(\tilde{\Delta}_{jh,1}^{(2)} u)_\ell = \frac{1}{jh} \left(\gamma_{\ell+\frac{j}{2}} \frac{u_{\ell+j} - u_\ell}{jh} - \gamma_{\ell-\frac{j}{2}} \frac{u_\ell - u_{\ell-j}}{jh} \right), \quad (8.9)$$

where $\gamma_{\ell+\frac{j}{2}}$ is defined as above.

The higher-dimensional discrete operators can easily be derived from (8.9).

In particular, the fourth-order approximation can be written as

$$\tilde{\Delta}_{h,p}^{(4)} = \frac{4}{3} \tilde{\Delta}_{h,p}^{(2)} - \frac{1}{3} \tilde{\Delta}_{2h,p}^{(2)}. \quad (8.10)$$

By applying (8.9) for $j = 2$, we obtain:

– In 1D:

$$(\tilde{\Delta}_{h,1}^{(4)} u)_\ell = \frac{4}{3h} \left(\gamma_{\ell+\frac{1}{2}} \frac{u_{\ell+1} - u_\ell}{h} - \gamma_{\ell-\frac{1}{2}} \frac{u_\ell - u_{\ell-1}}{h} \right) - \frac{1}{6h} \left(\gamma_{\ell+1} \frac{u_{\ell+2} - u_\ell}{2h} - \gamma_{\ell-1} \frac{u_\ell - u_{\ell-2}}{2h} \right). \quad (8.11)$$

– In 2D:

$$\begin{aligned} (\tilde{\Delta}_{h,2}^{(4)} u)_{\ell,m} &= \\ &\frac{4}{3h} \left(\gamma_{\ell+\frac{1}{2},m} \frac{u_{\ell+1,m} - u_{\ell,m}}{h} - \gamma_{\ell-\frac{1}{2},m} \frac{u_{\ell,m} - u_{\ell-1,m}}{h} \right. \\ &\left. + \gamma_{\ell,m+\frac{1}{2}} \frac{u_{\ell,m+1} - u_{\ell,m}}{h} - \gamma_{\ell,m-\frac{1}{2}} \frac{u_{\ell,m} - u_{\ell,m-1}}{h} \right) \\ &- \frac{1}{6h} \left(\gamma_{\ell+1,m} \frac{u_{\ell+2,m} - u_{\ell,m}}{2h} - \gamma_{\ell-1,m} \frac{u_{\ell,m} - u_{\ell-2,m}}{2h} \right. \\ &\left. + \gamma_{\ell,m+1} \frac{u_{\ell,m+2} - u_{\ell,m}}{2h} - \gamma_{\ell,m-1} \frac{u_{\ell,m} - u_{\ell,m-2}}{2h} \right). \end{aligned} \quad (8.12)$$

The 3D case is obtained in the same way.

8.4 Higher-Order Approximations: Second Approach

By using (8.2) and (4.61b), we obtain, for this approach, in 1D:

$$\begin{aligned} (\tilde{\Delta}_{h,1}^{(4)} u)_\ell = & \\ \frac{1}{h} \sum_{i=1}^k \alpha'_i (\gamma_{\ell+i-\frac{1}{2}} (D_{1,h}^{(2k)} u)_{\ell+i-\frac{1}{2}} - \gamma_{\ell-i+\frac{1}{2}} (D_{1,h}^{(2k)} u)_{\ell-i+\frac{1}{2}}). & \end{aligned} \quad (8.13)$$

For higher dimensions, we shall define the “discrete” partial derivatives $D_{x,h}^{(2k)}$, $D_{y,h}^{(2k)}$ and $D_{z,h}^{(2k)}$ which correspond to the operator $D_{1,h}^{(2k)}$ applied to each direction of space. With this notations, we obtain:

$$D_{2,h}^{(2k)} = (D_{x,h}^{(2k)}, D_{y,h}^{(2k)}) \quad (8.14)$$

and

$$D_{3,h}^{(2k)} = (D_{x,h}^{(2k)}, D_{y,h}^{(2k)}, D_{z,h}^{(2k)}), \quad (8.15)$$

which provides:

– In 2D:

$$\begin{aligned} (\tilde{\Delta}_{h,2}^{(2k)} u)_{\ell,m} = & \\ \frac{1}{h} \sum_{i=1}^k \alpha'_i (\gamma_{\ell+i-\frac{1}{2},m} (D_{x,h}^{(2k)} u)_{\ell+i-\frac{1}{2},m} - \gamma_{\ell-i+\frac{1}{2},m} (D_{x,h}^{(2k)} u)_{\ell-i+\frac{1}{2},m}) + & \\ \frac{1}{h} \sum_{i=1}^k \alpha'_i (\gamma_{\ell,m+i-\frac{1}{2}} (D_{y,h}^{(2k)} u)_{\ell,m+i-\frac{1}{2}} - \gamma_{\ell,m-i+\frac{1}{2}} (D_{y,h}^{(2k)} u)_{\ell,m-i+\frac{1}{2}}). & \end{aligned} \quad (8.16)$$

– In 3D:

$$\begin{aligned} (\tilde{\Delta}_{h,3}^{(2k)} u)_{\ell,m,n} = & \\ \frac{1}{h} \sum_{i=1}^k \alpha'_i (\gamma_{\ell+i-\frac{1}{2},m,n} (D_{x,h}^{(2k)} u)_{\ell+i-\frac{1}{2},m,n} - \gamma_{\ell-i+\frac{1}{2},m,n} (D_{x,h}^{(2k)} u)_{\ell-i+\frac{1}{2},m,n}) + & \\ \frac{1}{h} \sum_{i=1}^k \alpha'_i (\gamma_{\ell,m+i-\frac{1}{2},n} (D_{y,h}^{(2k)} u)_{\ell,m+i-\frac{1}{2},n} - \gamma_{\ell,m-i+\frac{1}{2},n} (D_{y,h}^{(2k)} u)_{\ell,m-i+\frac{1}{2},n}) + & \\ \frac{1}{h} \sum_{i=1}^k \alpha'_i (\gamma_{\ell,m,n+i-\frac{1}{2}} (D_{z,h}^{(2k)} u)_{\ell,m,n+i-\frac{1}{2}} - \gamma_{\ell,m,n-i+\frac{1}{2}} (D_{z,h}^{(2k)} u)_{\ell,m,n-i+\frac{1}{2}}). & \end{aligned} \quad (8.17)$$

The fourth-order version of the above approach can be written as follows:

– In 1D:

$$\begin{aligned}
 (\tilde{\Delta}_{h,1}^{(4)} u)_\ell = & \\
 \frac{1}{24h} \left(\gamma_{\ell-\frac{3}{2}} \frac{u_{\ell-3} - 27u_{\ell-2} + 27u_{\ell-1} - u_\ell}{24h} \right. & \\
 - 27\gamma_{\ell-\frac{1}{2}} \frac{u_{\ell-2} - 27u_{\ell-1} + 27u_\ell - u_{\ell+1}}{24h} & \\
 + 27\gamma_{\ell+\frac{1}{2}} \frac{u_{\ell-1} - 27u_\ell + 27u_{\ell+1} - u_{\ell+2}}{24h} & \\
 \left. - \gamma_{\ell+\frac{3}{2}} \frac{u_\ell - 27u_{\ell+1} + 27u_{\ell+2} - u_{\ell+3}}{24h} \right). &
 \end{aligned} \tag{8.18}$$

– In 2D:

$$\begin{aligned}
 (\tilde{\Delta}_{h,2}^{(4)} u)_{\ell,m} = & \\
 \frac{1}{24h} \left(\gamma_{\ell-\frac{3}{2},m} \frac{u_{\ell-3,m} - 27u_{\ell-2,m} + 27u_{\ell-1,m} - u_{\ell,m}}{24h} \right. & \\
 - 27\gamma_{\ell-\frac{1}{2},m} \frac{u_{\ell-2,m} - 27u_{\ell-1,m} + 27u_{\ell,m} - u_{\ell+1,m}}{24h} & \\
 + 27\gamma_{\ell+\frac{1}{2},m} \frac{u_{\ell-1,m} - 27u_{\ell,m} + 27u_{\ell+1,m} - u_{\ell+2,m}}{24h} & \\
 - \gamma_{\ell+\frac{3}{2},m} \frac{u_{\ell,m} - 27u_{\ell+1,m} + 27u_{\ell+2,m} - u_{\ell+3,m}}{24h} & \\
 + \gamma_{\ell,m-\frac{3}{2}} \frac{u_{\ell,m-3} - 27u_{\ell,m-2} + 27u_{\ell,m-1} - u_{\ell,m}}{24h} & \\
 - 27\gamma_{\ell,m-\frac{1}{2}} \frac{u_{\ell,m-2} - 27u_{\ell,m-1} + 27u_{\ell,m} - u_{\ell,m+1}}{24h} & \\
 + 27\gamma_{\ell,m+\frac{1}{2}} \frac{u_{\ell,m-1} - 27u_{\ell,m} + 27u_{\ell,m+1} - u_{\ell,m+2}}{24h} & \\
 \left. - \gamma_{\ell,m+\frac{3}{2}} \frac{u_{\ell,m} - 27u_{\ell,m+1} + 27u_{\ell,m+2} - u_{\ell,m+3}}{24h} \right). &
 \end{aligned} \tag{8.19}$$

Remarks

1. The point values of γ are often defined as mean values of this function centered at the point of definition. For instance, one can have:

$$\gamma_{\ell+\frac{1}{2}} = \frac{1}{h} \int_{\ell h}^{(\ell+1)h} \gamma(x) dx, \quad (8.20)$$

$$\gamma_{\ell+\frac{1}{2},m} = \frac{1}{2h^2} \int_{(m-1)h}^{(m+1)h} \int_{\ell h}^{(\ell+1)h} \gamma(x, y) dx dy, \quad (8.21)$$

$$\gamma_{\ell+\frac{1}{2},m,n} = \frac{1}{4h^3} \int_{(n-1)h}^{(n+1)h} \int_{(m-1)h}^{(m+1)h} \int_{\ell h}^{(\ell+1)h} \gamma(x, y, z) dx dy dz. \quad (8.22)$$

2. In the same way as (8.14) and (8.15), one can write:

$$\tilde{D}_{2,h}^{(2k)} = \tilde{D}_{x,h}^{(2k)} + \tilde{D}_{y,h}^{(2k)} \quad (8.23)$$

and

$$\tilde{D}_{3,h}^{(2k)} = \tilde{D}_{x,h}^{(2k)} + \tilde{D}_{y,h}^{(2k)} + \tilde{D}_{z,h}^{(2k)}. \quad (8.24)$$

3. Approximations of operators with cross derivatives such as $\partial/\partial x(\gamma\partial/\partial y)$ can be obtained in the same way, on the basis of the decomposition given in (4.83).

8.5 The Case of Arakawa's Scheme

The extension to non-homogeneous cases of Arakawa's scheme defined in (4.16) can be derived by noticing that this scheme can be rewritten as

$$\begin{aligned} & \left(\check{\Delta}_h^{(2)} u \right)_{\ell,m} \\ &= \frac{2}{3} \frac{1}{h^2} (u_{\ell+1,m} + u_{\ell-1,m} - 4u_{\ell,m} + u_{\ell,m+1} + u_{\ell,m-1}) \\ &+ \frac{1}{3} \frac{1}{2h^2} (u_{\ell+1,m+1} + u_{\ell-1,m+1} - 4u_{\ell,m} + u_{\ell+1,m-1} + u_{\ell-1,m-1}). \end{aligned} \quad (8.25)$$

In particular, the second term of this expression can be reinterpreted as a second-order approximation of Δ rotated by $\pi/4$, so that we can write:

$$\begin{aligned}
& \frac{1}{2h^2}(u_{\ell+1,m+1} + u_{\ell-1,m+1} - 4u_{\ell,m} + u_{\ell+1,m-1} + u_{\ell-1,m-1}) = \\
& \frac{\frac{u_{\ell+1,m+1} - u_{\ell,m}}{h\sqrt{2}} - \frac{u_{\ell,m} - u_{\ell-1,m-1}}{h\sqrt{2}}}{h\sqrt{2}} \\
& + \frac{\frac{u_{\ell-1,m+1} - u_{\ell,m}}{h\sqrt{2}} - \frac{u_{\ell,m} - u_{\ell+1,m-1}}{h\sqrt{2}}}{h\sqrt{2}}. \tag{8.26}
\end{aligned}$$

So, the heterogeneous version of Arakawa's scheme can be written as

$$\begin{aligned}
& \left(\tilde{\Delta}_{h,2}^{(2)} u \right)_{\ell,m} \\
& = \frac{2}{3h} \left(\gamma_{\ell-\frac{1}{2},m} \frac{u_{\ell,m} - u_{\ell-1,m}}{h} - \gamma_{\ell+\frac{1}{2},m} \frac{u_{\ell+1,m} - u_{\ell,m}}{h} \right. \\
& \quad \left. + \gamma_{\ell,m-\frac{1}{2}} \frac{u_{\ell,m} - u_{\ell,m-1}}{h} - \gamma_{\ell,m+\frac{1}{2}} \frac{u_{\ell,m+1} - u_{\ell,m}}{h} \right) \\
& \quad + \frac{1}{3h\sqrt{2}} \left(\gamma_{\ell+\frac{1}{2},m+\frac{1}{2}} \frac{u_{\ell+1,m+1} - u_{\ell,m}}{h\sqrt{2}} \right. \\
& \quad \left. - \gamma_{\ell-\frac{1}{2},m+\frac{1}{2}} \frac{u_{\ell,m} - u_{\ell-1,m-1}}{h\sqrt{2}} + \gamma_{\ell-\frac{1}{2},m+\frac{1}{2}} \frac{u_{\ell-1,m+1} - u_{\ell,m}}{h\sqrt{2}} \right. \\
& \quad \left. - \gamma_{\ell+\frac{1}{2},m-\frac{1}{2}} \frac{u_{\ell,m} - u_{\ell+1,m-1}}{h\sqrt{2}} \right). \tag{8.27}
\end{aligned}$$

Remark

Such an approach can be generalized to higher-order approximations with diagonal terms by rewriting the diagonal terms as a combination of second-order approximations rotated by $\pi/4$. Fourth-order approximations of this kind can be found in [28].

8.6 Approximation in Time

8.6.1 Second-Order Approximation in Time

The leapfrog approximation in time of the heterogeneous wave equation:

$$\eta(\mathbf{x}) \frac{\partial^2 u}{\partial t^2}(\mathbf{x}, t) - \operatorname{div}(\gamma(\mathbf{x}) \mathbf{grad} u(\mathbf{x}, t)) = 0, \tag{8.28}$$

can be written as follows:

$$\eta_r \frac{u_r^{n+1} - 2u_r^n + u_r^{n-1}}{\Delta t^2} - (\tilde{\Delta}_{h,p}^{(2q)} u_h^n)_r = 0, \quad (8.29)$$

where $r = \ell$ in 1D, $r = (\ell, m)$ in 2D, $r = (\ell, m, n)$ in 3D and η_r is the value of η at the point \mathbf{x}_r or a mean value on a cube centered at \mathbf{x}_r .

8.6.2 The Heterogeneous Modified Equation

Higher-order approximations obtained by symmetric schemes can be written in the same way as the leapfrog approximation. This is not the case for the modified equation approach which involves the square of the heterogeneous operator in space. In order to express this approach, we first recall the Taylor expansion of the leapfrog approximation at a point t_n :

$$\frac{u^{n+1} - 2u^n + u^{n-1}}{\Delta t^2} = \left(\frac{\partial^2 u}{\partial t^2} \right)^n + \frac{\Delta t^2}{12} \left(\frac{\partial^4 u}{\partial t^4} \right)^n + O(\Delta t^4). \quad (8.30)$$

By using (8.28), we obtain:

$$\begin{aligned} \frac{\partial^4 u}{\partial t^4} &= \frac{\partial^2}{\partial t^2} \left(\frac{\partial^2 u}{\partial t^2} \right) = \frac{\partial^2}{\partial t^2} \left(\frac{1}{\eta} \operatorname{div}(\gamma \mathbf{grad} u) \right) \\ &= \frac{1}{\eta} \operatorname{div} \left(\gamma \mathbf{grad} \left(\frac{\partial^2 u}{\partial t^2} \right) \right) \\ &= \frac{1}{\eta} \operatorname{div} \left(\gamma \mathbf{grad} \left(\frac{1}{\eta} \operatorname{div}(\gamma \mathbf{grad} u) \right) \right). \end{aligned}$$

Since we only need a second-order approximation in space of the correction term, its natural approximation is:

$$\left(\frac{1}{\eta} \operatorname{div} \left(\gamma \mathbf{grad} \left(\frac{1}{\eta} \operatorname{div}(\gamma \mathbf{grad}) \right) \right) \right) = \frac{1}{\eta} \tilde{\Delta}_{h,p}^{(2q)} \circ \left(\frac{1}{\eta} \tilde{\Delta}_{h,p}^{(2q)} \right). \quad (8.31)$$

So, we can write the heterogeneous modified equation:

$$\eta \frac{u_h^{n+1} - 2u_h^n + u_h^{n-1}}{\Delta t^2} + \tilde{\Delta}_{h,p}^{(2q)} u_h^n - \frac{\Delta t^2}{12} \left(\tilde{\Delta}_{h,p}^{(2q)} \left(\frac{1}{\eta} \tilde{\Delta}_{h,p}^{(2q)} u_h^n \right) \right) = 0. \quad (8.32)$$

8.6.3 Expression of the Correction Term

We must now give an explicit expression of $\tilde{\Delta}_{h,p}^{(2q)} \circ \left(\frac{1}{\eta} \tilde{\Delta}_{h,p}^{(2q)} \right)$. In 1D, we set

$$\begin{aligned} v_\ell &= \frac{1}{h} \left(\gamma_{\ell+\frac{1}{2}} \left(\frac{f_{\ell+1} - f_\ell}{h} \right) - \gamma_{\ell-\frac{1}{2}} \left(\frac{f_\ell - f_{\ell-1}}{h} \right) \right), \\ f_\ell &= \frac{1}{\eta_\ell h} \left(\gamma_{\ell+\frac{1}{2}} \left(\frac{u_{\ell+1} - u_\ell}{h} \right) - \gamma_{\ell-\frac{1}{2}} \left(\frac{u_\ell - u_{\ell-1}}{h} \right) \right), \end{aligned} \quad (8.33)$$

such that:

$$\begin{aligned} \left(\tilde{\Delta}_{h,1}^{(2q)} \left(\frac{1}{\eta} \tilde{\Delta}_{h,1}^{(2q)} u_h \right) \right)_\ell &= v_\ell = \\ -\frac{1}{h^3} \frac{\gamma_{\ell+\frac{1}{2}}}{\eta_{\ell+1}} \left(\gamma_{\ell+\frac{3}{2}} \left(\frac{u_{\ell+2} - u_{\ell+1}}{h} \right) - \gamma_{\ell+\frac{1}{2}} \left(\frac{u_{\ell+1} - u_\ell}{h} \right) \right) \\ +\frac{1}{\eta_\ell h^3} \left(\gamma_{\ell+\frac{1}{2}} + \gamma_{\ell-\frac{1}{2}} \right) \left(\gamma_{\ell+\frac{1}{2}} \left(\frac{u_{\ell+1} - u_\ell}{h} \right) - \gamma_{\ell-\frac{1}{2}} \left(\frac{u_\ell - u_{\ell-1}}{h} \right) \right) \\ -\frac{1}{h^3} \frac{\gamma_{\ell-\frac{1}{2}}}{\eta_{\ell-1}} \left(\gamma_{\ell-\frac{1}{2}} \left(\frac{u_\ell - u_{\ell-1}}{h} \right) - \gamma_{\ell-\frac{3}{2}} \left(\frac{u_{\ell-1} - u_{\ell-2}}{h} \right) \right). \end{aligned} \quad (8.34)$$

In 2D, we introduce the decomposition of the discrete operator $\tilde{\Delta}_{h,2}^{(2q)}$ in the same way as in (8.14):

$$\tilde{\Delta}_{h,2}^{(2q)} = \tilde{\Delta}_{h,x}^{(2q)} + \tilde{\Delta}_{h,y}^{(2q)}. \quad (8.35)$$

This decomposition enables us to write:

$$\begin{aligned} \tilde{\Delta}_{h,2}^{(2)} \circ \left(\frac{1}{\eta} \tilde{\Delta}_{h,2}^{(2)} \right) &= \tilde{\Delta}_{h,x}^{(2)} \circ \left(\frac{1}{\eta} \tilde{\Delta}_{h,x}^{(2)} \right) + \tilde{\Delta}_{h,x}^{(2)} \circ \left(\frac{1}{\eta} \tilde{\Delta}_{h,y}^{(2)} \right) \\ &\quad + \tilde{\Delta}_{h,y}^{(2)} \circ \left(\frac{1}{\eta} \tilde{\Delta}_{h,x}^{(2)} \right) + \tilde{\Delta}_{h,y}^{(2)} \circ \left(\frac{1}{\eta} \tilde{\Delta}_{h,y}^{(2)} \right). \end{aligned} \quad (8.36)$$

By using (8.14) and (8.23), we can easily obtain the expressions of the discrete second-order partial derivatives in x and y . The cross derivatives are a little more complicated.

As in (8.33), let us set:

$$v_{\ell,m} =$$

$$\frac{1}{h} \left(\gamma_{\ell+\frac{1}{2},m} \left(\frac{f_{\ell+1,m} - f_{\ell,m}}{h} \right) - \gamma_{\ell-\frac{1}{2},m} \left(\frac{f_{\ell,m} - f_{\ell-1,m}}{h} \right) \right), \quad (8.37)$$

$$f_{\ell,m} =$$

$$\frac{1}{\eta_{\ell,m} h} \left(\gamma_{\ell,m+\frac{1}{2}} \left(\frac{u_{\ell,m+1} - u_{\ell,m}}{h} \right) - \gamma_{\ell,m-\frac{1}{2}} \left(\frac{u_{\ell,m} - u_{\ell,m-1}}{h} \right) \right).$$

From (8.37), we can deduce:

$$\begin{aligned} & \left(\tilde{\Delta}_{h,x}^{(2)} \left(\frac{1}{\eta} \tilde{\Delta}_{h,y}^{(2)} u_h \right) \right)_{\ell,m} = -\frac{1}{h^3} \frac{\gamma_{\ell+\frac{1}{2},m}}{\eta_{\ell+1,m}} \\ & \left(\gamma_{\ell+1,m+\frac{1}{2}} \frac{u_{\ell+1,m+1} - u_{\ell+1,m}}{h} - \gamma_{\ell+1,m-\frac{1}{2}} \frac{u_{\ell+1,m} - u_{\ell+1,m-1}}{h} \right) \\ & + 2 \frac{\gamma_{\ell,m}}{\eta_{\ell,m} h^3} \left(\gamma_{\ell,m+\frac{1}{2}} \frac{u_{\ell,m+1} - u_{\ell,m}}{h} - \gamma_{\ell,m-\frac{1}{2}} \frac{u_{\ell,m} - u_{\ell,m-1}}{h} \right) \quad (8.38) \\ & - \frac{1}{h^3} \frac{\gamma_{\ell-\frac{1}{2},m}}{\eta_{\ell-1,m}} \\ & \left(\gamma_{\ell-1,m+\frac{1}{2}} \frac{u_{\ell-1,m+1} - u_{\ell-1,m}}{h} - \gamma_{\ell-1,m-\frac{1}{2}} \frac{u_{\ell-1,m} - u_{\ell-1,m-1}}{h} \right) \end{aligned}$$

and, in a similar way:

$$\begin{aligned} & \left(\tilde{\Delta}_{h,y}^{(2)} \left(\frac{1}{\eta} \tilde{\Delta}_{h,x}^{(2)} u_h \right) \right)_{\ell,m} = -\frac{1}{h^3} \frac{\gamma_{\ell,m+\frac{1}{2}}}{\eta_{\ell,m+1}} \\ & \left(\gamma_{\ell+\frac{3}{2},m+1} \frac{u_{\ell+1,m+1} - u_{\ell,m+1}}{h} - \gamma_{\ell-\frac{1}{2},m+1} \frac{u_{\ell,m+1} - u_{\ell-1,m+1}}{h} \right) \quad (8.39) \\ & + \frac{1}{\eta_{\ell,m} h^3} \left(\gamma_{\ell,m+\frac{1}{2}} + \gamma_{\ell,m-\frac{1}{2}} \right) \\ & \left(\gamma_{\ell+\frac{1}{2},m} \frac{u_{\ell+1,m} - u_{\ell,m}}{h} - \gamma_{\ell-\frac{1}{2},m} \frac{u_{\ell,m} - u_{\ell-1,m}}{h} \right) \end{aligned}$$

$$-\frac{1}{h^3} \frac{\gamma_{\ell,m-\frac{1}{2}}}{\eta_{\ell,m-1}} \\ \left(\gamma_{\ell-\frac{1}{2},m-1} \frac{u_{\ell+1,m-1} - u_{\ell,m-1}}{h} - \gamma_{\ell-\frac{1}{2},m-1} \frac{u_{\ell,m-1} - u_{\ell-1,m-1}}{h} \right).$$

8.7 Approximation of the Boundary Conditions

Although natural for second-order schemes, the approximation of boundary conditions raises rather difficult problems in higher-order approximations, in particular in the case of systems.

In order to simplify our presentation, we shall consider, in the following, the boundary of a 2D problem located at the point x_0 and parallel to the y axis. The 3D case does not introduce additional difficulties.

8.7.1 Second-Order Schemes

For second-order schemes, one can easily derive both Dirichlet and Neumann conditions.

The Dirichlet Condition. The Dirichlet condition for the wave equation defined in (1.43) can obviously be written as

$$u_{0,m}^n = g(x_0, y_m, t_n). \quad (8.40)$$

So, the value of the discrete Laplace operator at a point (x_1, y_m) is for the classical second-order scheme:

$$(\tilde{\Delta}_{h,2}^{(2)} u)_{1,m} = \\ \frac{1}{h} \left(\gamma_{1+\frac{1}{2},m} \frac{u_{2,m} - u_{1,m}}{h} - \gamma_{1-\frac{1}{2},m} \frac{u_{1,m} - g(x_0, y_m, t_n)}{h} \right. \\ \left. + \gamma_{1,m+\frac{1}{2}} \frac{u_{1,m+1} - u_{1,m}}{h} - \gamma_{1,m-\frac{1}{2}} \frac{u_{1,m} - u_{1,m-1}}{h} \right), \quad (8.41)$$

which provides a natural definition of $(\tilde{\Delta}_{h,2}^{(2)} u)_{1,m}$.

The Neumann Condition. The Neumann condition defined in (1.44) can be expressed in a straightforward way, by adding exterior values of u :

$$\frac{u_{1,m}^n - u_{-1,m}^n}{2h} = g(x_0, y_m, t_n). \quad (8.42)$$

On the other hand, we can write the discrete wave equation at a point (x_0, y_m) :

$$\begin{aligned} & \eta_{0,m} \frac{u_{0,m}^{n+1} - 2u_{0,m}^n + u_{0,m}^{n-1}}{\Delta t^2} - \\ & \frac{1}{h} \left(\gamma_{\frac{1}{2},m} \frac{u_{1,m}^n - u_{0,m}^n}{h} - \gamma_{-\frac{1}{2},m} \frac{u_{0,m}^n - u_{-1,m}^n}{h} \right. \\ & \left. + \gamma_{0,m+\frac{1}{2}} \frac{u_{0,m+1}^n - u_{0,m}^n}{h} - \gamma_{0,m-\frac{1}{2}} \frac{u_{0,m}^n - u_{0,m-1}^n}{h} \right) = 0. \end{aligned} \quad (8.43)$$

So, by combining (8.42) and (8.43), we can eliminate the exterior value $u_{-1,m}^n$ and include the Neumann condition into the scheme as follows:

$$\begin{aligned} & \eta_{0,m} \frac{u_{0,m}^{n+1} - 2u_{0,m}^n + u_{0,m}^{n-1}}{\Delta t^2} - \\ & \frac{1}{h} \left((\gamma_{\frac{1}{2},m} + \gamma_{-\frac{1}{2},m}) \frac{u_{1,m}^n - u_{0,m}^n}{h} - 2\gamma_{-\frac{1}{2},m} g(x_0, y_m, t_n) \right. \\ & \left. + \gamma_{0,m+\frac{1}{2}} \frac{u_{0,m+1}^n - u_{0,m}^n}{h} - \gamma_{0,m-\frac{1}{2}} \frac{u_{0,m}^n - u_{0,m-1}^n}{h} \right) = 0, \end{aligned} \quad (8.44)$$

which is actually a discrete variational form of the boundary condition.

Remarks

1. The above technique can be easily extended to the boundary condition at a corner of a rectangular domain.
2. Both Dirichlet and Neumann conditions can be treated in the same way for more complex second-order stencils, as defined in (4.16).

8.7.2 Higher-Order Schemes

The extension of such approximations to higher-order schemes is unfortunately not possible for the reason that we are going to illustrate by the following example:

Let us define the Dirichlet boundary condition as in (8.40) for the first approach of the fourth-order approximation. The value of $(\tilde{\Delta}_{h,4}^{(2q)} u_h^n)_{1,m}$ is then written as

$$\begin{aligned}
(\tilde{\Delta}_{h,2}^{(4)} u)_{1,m} = & \\
& \frac{4}{3h} \left(\gamma_{1+\frac{1}{2},m} \frac{u_{2,m} - u_{1,m}}{h} - \gamma_{1-\frac{1}{2},m} \frac{u_{1,m} - u_{0,m}}{h} \right. \\
& \left. + \gamma_{1,m+\frac{1}{2}} \frac{u_{1,m+1} - u_{1,m}}{h} - \gamma_{1,m-\frac{1}{2}} \frac{u_{1,m} - u_{1,m-1}}{h} \right) \\
& - \frac{1}{6h} \left(\gamma_{2,m} \frac{u_{3,m} - u_{1,m}}{2h} - \gamma_{0,m} \frac{u_{1,m} - u_{-1,m}}{2h} \right. \\
& \left. + \gamma_{1,m+1} \frac{u_{1,m+2} - u_{1,m}}{2h} - \gamma_{1,m-1} \frac{u_{1,m} - u_{1,m-2}}{2h} \right). \tag{8.45}
\end{aligned}$$

In this case, $u_{0,m}$ can be replaced by its value given in (8.40) but $u_{-1,m}$ has no value.

For the homogeneous condition ($g = 0$), this problem can be overcome by using the *image principle*, i.e. by considering the two approximated Dirichlet boundary conditions:

$$\frac{1}{2}(u_{1,m} + u_{-1,m}) = 0, \tag{8.46a}$$

$$\frac{1}{2}(u_{2,m} + u_{-2,m}) = 0. \tag{8.46b}$$

In other words, we add two exterior values $u_{-1,m}$ and $u_{-2,m}$ of u that we set equal to $-u_{1,m}$ and $-u_{2,m}$, respectively.

The non-homogeneous case, of course, cannot be treated by such a technique. In general, a boundary source defined by g will be shifted to an interior point of the domain, close to the boundary.

In the case of the Neumann condition, we also set, in the homogeneous case:

$$\frac{1}{2h}(u_{1,m} - u_{-1,m}) = 0, \tag{8.47a}$$

$$\frac{1}{4h}(u_{2,m} - u_{-2,m}) = 0, \tag{8.47b}$$

which set $u_{-1,m}$ and $u_{-2,m}$ equal to $u_{1,m}$ and $u_{2,m}$, respectively.

In the non-homogeneous case, the condition can also be shifted.

Such techniques can be extended to higher-order schemes by adding as many exterior values as the order of the method requires.

Remarks

1. Of course, the image principle could be used for the second-order approximation.
2. The impedance condition defined in (1.45) can be treated as the Neumann condition for second-order schemes by replacing $g(x_0, y_m, t_n)$ by $-(1/\alpha)(u_{0,m}^{n+1} - u_{0,m}^{n-1})/2\Delta t$ in (8.44) but, of course, not for higher-order schemes.

8.7.3 Extension to Systems

The Maxwell Equations. As we shall see later, the discrete values of \mathbf{E} in the Maxwell equations are actually the tangential components of this vector valued field. So, the perfectly conducting boundary condition defined in (1.46) can be treated as a Dirichlet condition on some components of the electric field. Of course, the same difficulties will be met for higher-order approximations and the image principle can also be used in that case.

The Silver-Müller conditions must be treated in the same way as the impedance condition but one must take into account the centered character of the scheme in a careful way. We shall not develop its approximation here.

The Elastics System. The displacement boundary condition defined in (1.49) is actually a Dirichlet boundary condition on each component of \mathbf{v} which can be treated as for the wave equation.

The traction boundary condition is not obvious to express when the elastics system is written in terms of \mathbf{v} only as in (1.34a) and (1.34c) and (1.35a) and (1.35b) but is a simple Dirichlet condition on the components of $\underline{\tau}\mathbf{n}$, where \mathbf{n} is the unit outward normal. As for the wave equation, the non-homogeneous boundary conditions will be shifted to an interior point of the domain, close to the boundary.

To all these difficulties of approximation of the boundary conditions must be added the fact that the approximation of a curved boundary is not so easy and introduces in general some parasitic reflections due to the staircase approximation of the curves.

We are now going to give some basis of analysis of the approximations defined in this chapter. We shall only consider the case of unbounded domains.

9. Stability by Energy Techniques

The study of the stability of the schemes obtained for heterogeneous media cannot be done by plane wave (or Fourier) analysis, since such a technique requires the invariance by translation of the medium. So, one must use energy techniques. These techniques are more general but also more complicated and their complexity often leads to partial results. In this chapter, we shall give some principles and examples of this kind of technique.

9.1 Positivity of the Discrete Operators

As in the homogeneous case, the positivity of the discrete operators is a necessary condition of stability of the schemes. So, before looking for the stability conditions by energy techniques, we study, as a first step, the positivity of the discrete operators in space.

Let A_h be a difference operator from V_0^d (defined in (4.7) and (4.11), d being the dimension in space of the problem) into itself. We associate to this operator the bilinear form¹:

$$a_h(u_h, v_h) = (A_h u_h, v_h)_0 \quad (u_h, v_h) \in (V_0^d)^2. \quad (9.1)$$

The positivity of A_h is then written as

$$a_h(u_h, u_h) \geq 0 \quad \forall u_h \in V_0^d. \quad (9.2)$$

9.1.1 Second-Order and Second Approach for Fourth-Order Approximations

The positivity of the operators obtained by the second-order approximation and the second approach of the fourth-order approximation is easy to obtain since, in these cases

$$A_h = -(D_{d,h}^{(q)})^* \circ (\gamma D_{d,h}^{(q)}), \quad (9.3)$$

where $q = 2$ or 4 .

¹ We recall that $(\cdot, \cdot)_0$ is the scalar product of V_0^d .

So, if $(\cdot, \cdot)_0$ is the scalar product of V_0^d (defined in (4.7) and (4.11)) and $(\cdot, \cdot)_{\frac{1}{2}}$ the scalar product of $V_{\frac{1}{2}}^d$ (defined in (4.8) and (4.12)), we can write the following relation:

$$\left(\tilde{\Delta}_{h,p}^{(q)} u_h, v_h \right)_0 = - \left(\gamma D_{d,h}^{(q)} u_h, D_{d,h}^{(q)} v_h \right)_{\frac{1}{2}}, \quad (9.4)$$

which ensures the positivity of the discrete operator $-\tilde{\Delta}_{h,d}^{(q)}$ as soon as $\gamma > 0$.

9.1.2 Fourth-Order: First Approach

For the first approach, the positivity is much less obvious because of the negative component of the combination of second-order approximations.

For reasons which will appear in the study, we shall assume that

$$\gamma_{\ell+\frac{1}{2}} = \frac{1}{h} \int_{\ell h}^{(\ell+1)h} \gamma(x) dx, \quad (9.5)$$

and

$$\gamma_\ell = \frac{1}{2h} \int_{(\ell-1)h}^{(\ell+1)h} \gamma(x) dx. \quad (9.6)$$

On the other hand, we shall replace, in (8.11), $\gamma_{\ell+\frac{1}{2}}$ by

$$\gamma_{\ell+\frac{1}{2}}^\theta = (1 - 2\theta)\gamma_{\ell+\frac{1}{2}} + \theta(\gamma_{\ell-\frac{1}{2}} + \gamma_{\ell+\frac{3}{2}}). \quad (9.7)$$

In the same way, one can replace, in (8.12), $\gamma_{\ell+\frac{1}{2},m}$ and $\gamma_{\ell,m+\frac{1}{2}}$ by

$$\gamma_{\ell+\frac{1}{2},m}^\theta = (1 - 2\theta)\gamma_{\ell+\frac{1}{2},m} + \theta(\gamma_{\ell-\frac{1}{2},m} + \gamma_{\ell+\frac{3}{2},m}), \quad (9.8a)$$

$$\gamma_{\ell,m+\frac{1}{2}}^\theta = (1 - 2\theta)\gamma_{\ell,m+\frac{1}{2}} + \theta(\gamma_{\ell,m-\frac{1}{2}} + \gamma_{\ell,m+\frac{3}{2}}). \quad (9.8b)$$

For this approximation, we can write:

$$\begin{aligned} a_h^\theta(u_h, v_h) = & \frac{4}{3} \sum_{\ell=-\infty}^{+\infty} \gamma_{\ell+\frac{1}{2}}^\theta \frac{u_{\ell+1} - u_\ell}{h} \frac{v_{\ell+1} - v_\ell}{h} h \\ & - \frac{1}{3} \sum_{\ell=-\infty}^{+\infty} \gamma_\ell \frac{u_{\ell+1} - u_{\ell-1}}{2h} \frac{v_{\ell+1} - v_{\ell-1}}{2h} h. \end{aligned} \quad (9.9)$$

Our study will be limited to the 1D case since its extension to the 2D case brings nothing basically new but is much more tedious.

Let us start from the identity:

$$a_h^\theta(u_h, u_h) = \frac{4}{3} \sum_{\ell=-\infty}^{+\infty} \gamma_{\ell+\frac{1}{2}}^\theta \left| \frac{u_{\ell+1} - u_\ell}{h} \right|^2 h - \frac{1}{3} \sum_{\ell=-\infty}^{+\infty} \gamma_\ell \left| \frac{u_{\ell+1} - u_{\ell-1}}{2h} \right|^2 h. \quad (9.10)$$

In order to estimate the second term of (9.10), we use the formula:

$$\left| \frac{u_{\ell+1} - u_{\ell-1}}{2h} \right|^2 \leq \frac{1}{2} \left| \frac{u_{\ell+1} - u_\ell}{h} \right|^2 + \frac{1}{2} \left| \frac{u_\ell - u_{\ell-1}}{h} \right|^2,$$

which provides

$$\sum_{\ell=-\infty}^{+\infty} \gamma_\ell \left| \frac{u_{\ell+1} - u_{\ell-1}}{2h} \right|^2 h \leq \sum_{\ell=-\infty}^{+\infty} \frac{\gamma_{\ell+1} + \gamma_\ell}{2} \left| \frac{u_{\ell+1} - u_\ell}{h} \right|^2 h. \quad (9.11)$$

By inserting (9.11) into (9.10), we obtain

$$a_h^\theta(u_h, u_h) \geq \sum_{\ell=-\infty}^{+\infty} \left\{ \frac{4}{3} \gamma_{\ell+\frac{1}{2}}^\theta - \frac{1}{3} \frac{(\gamma_{\ell+1} + \gamma_\ell)}{2} \right\} \left| \frac{u_{\ell+1} - u_\ell}{h} \right|^2 h. \quad (9.12)$$

Now, the definitions of γ_ℓ , $\gamma_{\ell+1}$ and $\gamma_{\ell+\frac{1}{2}}^\theta$ lead to

$$\begin{aligned} \frac{4}{3} \gamma_{\ell+\frac{1}{2}}^\theta - \frac{1}{3} \left(\frac{\gamma_{\ell+1} + \gamma_\ell}{2} \right) &= \left\{ \frac{4}{3}(1 - 2\theta) - \frac{1}{6} \right\} \gamma_{\ell+\frac{1}{2}} \\ &\quad + \left\{ \frac{4}{3}\theta - \frac{1}{12} \right\} (\gamma_{\ell+\frac{3}{2}} + \gamma_{\ell-\frac{1}{2}}). \end{aligned} \quad (9.13)$$

So, we obtain the following sufficient condition of positivity of (9.13):

$$\frac{1}{16} \leq \theta \leq \frac{7}{16}. \quad (9.14)$$

In this case, the right-hand side of (9.13) appears as a convex combination of $\gamma_{\ell-\frac{1}{2}}$, $\gamma_{\ell+\frac{1}{2}}$ and $\gamma_{\ell+\frac{3}{2}}$, such that

$$\gamma_{\ell+\frac{1}{2}} \geq \gamma_* \Rightarrow \frac{4}{3} \gamma_{\ell+\frac{1}{2}}^\theta - \frac{1}{3} \frac{(\gamma_{\ell+1} + \gamma_\ell)}{2} \geq \gamma_*. \quad (9.15)$$

By inserting this inequality into (9.12), we obtain:

$$a_h^\theta(u_h, u_h) \geq \gamma_* \sum_{\ell=-\infty}^{+\infty} \left| \frac{u_{\ell+1} - u_\ell}{h} \right|^2 h. \quad (9.16)$$

All this study can be summarized in the

Lemma 2. *If $1/16 \leq \theta \leq 7/16$, then, for any function $\gamma(x)$, the quadratic form $a_h(u_h, u_h)$ is positive and satisfies the coerciveness inequality:*

$$a_h^\theta(u_h, u_h) \geq \gamma_* \sum_{\ell=-\infty}^{+\infty} \left| \frac{u_{\ell+1} - u_\ell}{h} \right|^2 h.$$

Remarks

1. For $\theta = 0$, we obtain:

$$a_h^{(0)}(u_h, u_h) \geq \sum_{\ell=-\infty}^{+\infty} \left\{ \frac{4}{3} \gamma_{\ell+\frac{1}{2}} - \frac{1}{3} \left(\frac{\gamma_{\ell+1} + \gamma_\ell}{2} \right) \right\} \left| \frac{u_{\ell+1} - u_\ell}{h} \right|^2 h. \quad (9.17)$$

If $\gamma(x)$ varies quickly in the neighbourhood of $x_{\ell+1/2}$, it is easy to see that, since γ_ℓ and $\gamma_{\ell+1}$ depend on $\gamma_{\ell-1/2}$ and $\gamma_{\ell+3/2}$, it is not possible to control the term $(\gamma_\ell + \gamma_{\ell+1})$ by the use of only $\gamma_{\ell+1/2}$. This observation is the motivation of the new definition of $\gamma_{\ell+1/2}$ given in (9.7).

2. The inequality $\theta \geq 1/16$ is a numerical expression of the fact that the weight of the “extrapolation term” $\theta(\gamma_{\ell+3/2} + \gamma_{\ell-3/2})$ must be large enough. The inverse inequality $\theta \leq 7/16$ indicates that this weight must not be too large.

Of course, it would be natural to wonder whether the condition $1/6 \leq \theta \leq 7/16$ is necessary or not. We can actually formulate this question in the following way: when $\theta \notin [1/16, 7/16]$, does there exist a function $\gamma(x)$ and a discrete solution u_h which belongs to l^2 such that $a_h^\theta(u_h, u_h) < 0$?

The answer to this question is not so clear. We were only able to obtain the following result:

Lemma 3. *If θ belongs to the interval $[0, 1/32]$ there exists a function $\gamma(x)$ such that*

$$\exists u_h \in L_h^2 / a_h^\theta(u_h, u_h) < 0.$$

Proof. Let $u_h^{(l)}$ be the function defined by

$$\begin{cases} u_\ell^{(l)} = \frac{1}{h} & \text{if } \ell = l \\ u_\ell^{(l)} = 0 & \text{if } \ell \neq l. \end{cases}$$

A simple computation provides

$$\begin{aligned} h^3 a_h^\theta(u_h, u_h) = & \frac{1}{3} \left\{ (4\theta - \frac{1}{8}) \left(\frac{\gamma_{l+\frac{3}{2}} + \gamma_{l-\frac{3}{2}}}{2} \right) \right. \\ & \left. + \left(\frac{31}{8} - 4\theta \right) (\gamma_{l+\frac{1}{2}} + \gamma_{l-\frac{1}{2}}) \right\}. \end{aligned} \quad (9.18)$$

Now, if θ belongs to the interval $[0, 1/32]$, the coefficient $(4\theta - 1/8)$ is negative. This implies that if we define the two-layer medium

$$\begin{cases} \gamma(x) = \gamma_1 \text{ for } x < x_0, \\ \gamma(x) = \gamma_2 \text{ for } x > x_0, \end{cases}$$

we can choose θ such that $|x_0 - (1 + 1/2)h| < h$ and $\gamma_2 - \gamma_1$ large enough such that

$$a_h^\theta(u_h^{(1)}, u_h^{(1)}) < 0.$$

◇

The results of the two previous lemmas can be summarized in the following stability theorem:

Theorem 3. (i) When $\theta \in \left[\frac{1}{16}, \frac{7}{16}\right]$, the problem

$$\rho \frac{d^2 u_h}{dt^2} + \tilde{\Delta}_{h,p}^{(4)} u_h = 0 \quad (9.19)$$

is L^2 -stable.

(ii) On the other hand, when $\theta \in \left[0, \frac{1}{32}\right]$, (9.19) could be unstable for some functions $\gamma(x)$.

Remarks

1. Actually, when θ belongs to the interval $[0, 1/32]$, the Cauchy problem associated with (9.19) is unstable as soon as the function $\gamma(x)$ has large enough discontinuities. Moreover, even if we consider a function $\gamma(x)$ for which the linear form associated to the problem is positive, the coercivity inequality (9.16) is no longer true. One can show that we obtain an inequality of the form

$$a_h^\theta(u_h, u_h) \geq \alpha(\theta, \gamma) \sum_{\ell=-\infty}^{+\infty} \left| \frac{u_{\ell+1} - u_\ell}{h} \right|^2 h,$$

where $\alpha(\theta, \gamma)$ depends on the variation of the function $\gamma(x)$ and not only of its lower bound γ_* .

2. For solving inverse problems, in which the coefficients $\eta(x)$ and $\gamma(x)$ are unknown, it is safer to use numerical schemes independent of these coefficients.
3. Another approach to these questions of stability would be the use of the Kreiss techniques [60, 61, 76] to a two-layer medium. Although these techniques could lead to necessary and sufficient conditions of stability, they require very complicated computations, particularly in our case.

4. The stability of the scheme in higher-order dimensions is not so easy to obtain. It seems, in particular, that the modified equation approach could be unconditionally unstable if the contrast in velocities is too large.

9.2 Stability Conditions

9.2.1 A General Framework

We now look for a discrete equivalent of the energy identity given in (2.26).

Let us consider the following discrete problem:

$$\left(\frac{u_h^{n+1} - 2u_h^n + u_h^{n-1}}{\Delta t^2}, v_h \right)_0 + a_h(u_h^n, v_h) = 0 \quad u_h^n \in V_0^d, v_h \in V_0^d, \quad (9.20)$$

where $a_h(u_h^n, v_h)$ is associated to an approximation A_h of $-\Delta$.

By setting $v_h = \frac{u_h^{n+1} - u_h^{n-1}}{2\Delta t} = \frac{1}{2} \left(\frac{u_h^{n+1} - u_h^n}{\Delta t} + \frac{u_h^n - u_h^{n-1}}{\Delta t} \right)$ in (9.20), we obtain:

$$\frac{1}{2\Delta t} \left(\left\| \frac{u_h^{n+1} - u_h^n}{\Delta t} \right\|^2 - \left\| \frac{u_h^n - u_h^{n-1}}{\Delta t} \right\|^2 + a_h(u_h^{n+1}, u_h^n) - a_h(u_h^n, u_h^{n-1}) \right) = 0. \quad (9.21)$$

On the basis of the following discrete energy:

$$E_h^{n+\frac{1}{2}} = \frac{1}{2} \left\| \frac{u_h^{n+1} - u_h^n}{\Delta t} \right\|^2 + \frac{1}{2} a_h(u_h^{n+1}, u_h^n), \quad (9.22)$$

this identity can be written as

$$\frac{1}{\Delta t} (E_h^{n+\frac{1}{2}} - E_h^{n-\frac{1}{2}}) = 0. \quad (9.23)$$

Unfortunately, this definition of the discrete energy does not ensure the positivity of $a_h(u_h^{n+1}, u_h^n)$. Therefore, we must work on this formulation to obtain an adequate form.

We first notice that

$$\begin{aligned} a_h(u_h^{n+1}, u_h^n) &= a_h \left(\frac{u_h^{n+1} + u_h^n}{2}, \frac{u_h^{n+1} + u_h^n}{2} \right) \\ &\quad - \frac{\Delta t^2}{4} a_h \left(\frac{u_h^{n+1} - u_h^n}{\Delta t}, \frac{u_h^{n+1} - u_h^n}{\Delta t} \right). \end{aligned} \quad (9.24)$$

This equality implies that

$$\begin{aligned} E_h^{n+\frac{1}{2}} = & \left\| \frac{u_h^{n+1} - u_h^n}{\Delta t} \right\|^2 + a_h \left(\frac{u_h^{n+1} + u_h^n}{2}, \frac{u_h^{n+1} + u_h^n}{2} \right) \\ & - \frac{\Delta t^2}{4} a_h \left(\frac{u_h^{n+1} - u_h^n}{\Delta t}, \frac{u_h^{n+1} - u_h^n}{\Delta t} \right). \end{aligned} \quad (9.25)$$

Now, since we have

$$\|A_h\| = \sup_{u_h \in V_0^d} \frac{a_h(u_h, u_h)}{\|u_h\|^2}, \quad (9.26)$$

then

$$a_h \left(\frac{u_h^{n+1} - u_h^n}{\Delta t}, \frac{u_h^{n+1} - u_h^n}{\Delta t} \right) \leq \|A_h\| \left\| \frac{u_h^{n+1} - u_h^n}{\Delta t} \right\|^2. \quad (9.27)$$

Finally, the discrete energy can be written as

$$E_h^{n+\frac{1}{2}} \geq \left(1 - \frac{\Delta t^2}{4} \|A_h\| \right) \left\| \frac{u_h^{n+1} - u_h^n}{\Delta t} \right\|^2 + a_h(u_h^{n+\frac{1}{2}}, u_h^{n+\frac{1}{2}}). \quad (9.28)$$

A sufficient condition of positivity of $E_h^{n+\frac{1}{2}}$ is then:

$$\Delta t < \frac{2}{\sqrt{\|A_h\|}}. \quad (9.29)$$

So, the computation of the positivity condition is equivalent to the computation of the norm of A_h .

9.2.2 Computation of $\|A_h\|$ for the Second-Order Approximation

As an example, we shall now compute $\|A_h\|$ for the second-order scheme in 1D. A similar computation for fourth-order schemes is much more complicated. In [35, 102], one can find this computation for the two approaches of fourth-order schemes.

Let

$$\frac{d^2 u_h}{dt^2} + A_h u_h = 0 \quad (9.30)$$

be the semi-discrete second-order approximation in space of the wave equation, where

$$(A_h u_h)_\ell = \frac{1}{\eta_\ell h} \left(\gamma_{\ell-\frac{1}{2}} \frac{u_\ell - u_{\ell-1}}{h} - \gamma_{\ell+\frac{1}{2}} \frac{u_{\ell+1} - u_\ell}{h} \right). \quad (9.31)$$

A_h is self-adjoint for the scalar product of V_0^1 whose corresponding norm is

$$\|u_h\| = \sqrt{\sum_{\ell \in \mathbb{Z}} \eta_\ell |u_\ell|^2 h}. \quad (9.32)$$

From the above definitions, we obtain

$$a_h(u_h, u_h) = \sum_{\ell \in \mathbb{Z}} \gamma_{\ell + \frac{1}{2}} \left| \frac{u_{\ell+1} - u_\ell}{h} \right|^2 h. \quad (9.33)$$

Then, we have:

$$a_h(u_h, u_h) \leq \frac{1}{h^2} \sum_{\ell \in \mathbb{Z}} \gamma_{\ell + \frac{1}{2}} (|u_{\ell+1}|^2 + |u_\ell|^2) h. \quad (9.34)$$

After index translations and by using the fact that $\gamma_\ell = (\gamma_{\ell + \frac{1}{2}} + \gamma_{\ell - \frac{1}{2}})/2$, the right-hand side of this identity can be written as

$$\begin{aligned} \frac{1}{h^2} \sum_{\ell \in \mathbb{Z}} \gamma_{\ell + \frac{1}{2}} (|u_{\ell+1}|^2 + |u_\ell|^2) h &= \frac{2}{h^2} \sum_{\ell \in \mathbb{Z}} (\gamma_{\ell + \frac{1}{2}} + \gamma_{\ell - \frac{1}{2}}) |u_\ell|^2 h \\ &= \frac{4}{h^2} \sum_{\ell \in \mathbb{Z}} \gamma_\ell |u_\ell|^2 h \\ &= \frac{4}{h^2} \sum_{\ell \in \mathbb{Z}} \frac{\gamma_\ell}{\eta_\ell} \eta_\ell |u_\ell|^2 h. \end{aligned} \quad (9.35)$$

Now, by setting $c^* = \sup_{\ell \in \mathbb{Z}} \sqrt{\frac{\gamma_\ell}{\eta_\ell}}$, we obtain:

$$a_h(u_h, u_h) \leq \frac{4}{h^2} (c^*)^2 \|u_h\|^2, \quad (9.36)$$

which, by using (9.26), can be rewritten as

$$\|A_h\| \leq \frac{4}{h^2} (c^*)^2. \quad (9.37)$$

So, (9.37) provides the following stability condition:

$$\frac{c^* \Delta t}{h} < 1, \quad (9.38)$$

which is a natural extension of that obtained for a homogeneous medium.

As we said before, this kind of result is not always obvious for higher-order schemes. However, one can use, in practice, the condition obtained for a homogeneous medium in which c is replaced by c^* .

10. Reflection-Transmission Analysis

The purpose of this chapter is to study the effect of a discontinuity on the accuracy of the method. This effect will be measured by a plane wave analysis on a two-layer medium.

10.1 The 1D Case

10.1.1 The Continuous Problem

Let us consider the following two-layer medium:

$$(\gamma(x), \eta(x)) = \begin{cases} (\gamma_1, \eta_1) & \text{for } x < 0, \\ (\gamma_2, \eta_2) & \text{for } x > 0. \end{cases} \quad (10.1)$$

The velocity of a wave is $c_1 = \sqrt{\gamma_1/\eta_1}$ in the first layer and $c_2 = \sqrt{\gamma_2/\eta_2}$ in the second one.

In such a medium, an incident plane wave of amplitude 1 which propagates in the first layer will be decomposed, after crossing the interface between the two layers, into a transmitted wave of amplitude T and a reflected wave of amplitude R . Then, the solution can be written as

$$u(x, t) = \chi_{\mathbb{R}^-}(x) \left(e^{i(\omega t - K_1 x)} + R e^{i(\omega t + K_1 x)} \right) + T \chi_{\mathbb{R}^+}(x) e^{i(\omega t - K_2 x)}, \quad (10.2)$$

where $\chi_{\mathbb{R}^+}$ and $\chi_{\mathbb{R}^-}$ are the characteristic functions of the sets \mathbb{R}^+ and \mathbb{R}^- . On the other hand, K_1 and K_2 are defined by the dispersion relations in each layer

$$K_1 = \frac{\omega}{c_1}, \quad (10.3a)$$

$$K_2 = \frac{\omega}{c_2}. \quad (10.3b)$$

Since u and its normal derivative $\gamma \partial u / \partial n$ are continuous at the interface, we can write:

$$u(0^-) = e^{i\omega t} + Re^{i\omega t} = u(0^+) = Te^{i\omega t}, \quad (10.4a)$$

$$\gamma_1 u'(0^-) = -i\gamma_1 K_1 e^{i\omega t} + Ri\gamma_1 K_1 e^{i\omega t} = \gamma_2 u'(0^+) = -Ti\gamma_2 K_2 e^{i\omega t}. \quad (10.4b)$$

After simplification, we obtain:

$$1 + R = T, \quad (10.5a)$$

$$K_1 \gamma_1 R + K_2 \gamma_2 T = K_1 \gamma_1. \quad (10.5b)$$

By taking into account (10.3a) and (10.3b), this can be rewritten as

$$R = \frac{\sigma_1 - \sigma_2}{\sigma_1 + \sigma_2}, \quad (10.6a)$$

$$T = \frac{2\sigma_1}{\sigma_1 + \sigma_2}, \quad (10.6b)$$

where $\sigma_j = \sqrt{\eta_j \gamma_j}$, $j = 1, 2$ is the *acoustic impedance* of each layer.

10.1.2 Second-Order Approximation

We now apply this process to the semi-discrete approximation in space of the wave equation.

Let us first define a mesh of space-step equal to h on \mathbb{R} so that $x_0 = 0$. As in the continuous case, we obtain from the discrete dispersion relations in each layer, two wavenumbers K_1 and K_2 given by

$$K_1 = \frac{2}{h} \arcsin \left(\frac{\omega h}{2c_1} \right), \quad (10.7a)$$

$$K_2 = \frac{2}{h} \arcsin \left(\frac{\omega h}{2c_2} \right). \quad (10.7b)$$

Here also, we have one transmitted and one reflected wave. Then, the solution u_h can be written as

$$u(x, t) = e^{i(\omega t - \ell K_1 h)} + R_h e^{i(\omega t + \ell K_1 h)} \quad \text{for } \ell \leq 0, \quad (10.8a)$$

$$u(x, t) = T_h e^{i(\omega t - \ell K_2 h)} \quad \text{for } \ell \geq 0. \quad (10.8b)$$

By writing u_h at $\ell = 0$, we obtain (10.5a).

On the other hand, at this point, the semi-discrete wave equation can be written as

$$\frac{(\eta_1 + \eta_2)}{2} \frac{d^2 u_h}{dt^2} + \frac{1}{h} \left(\gamma_1 \frac{u_0 - u_{-1}}{h} - \gamma_2 \frac{u_1 - u_0}{h} \right) = 0. \quad (10.9)$$

By taking into account (10.8a) and (10.8b), this provides the following system in (R_h, T_h) :

$$1 + R_h = T_h, \quad (10.10a)$$

$$\begin{aligned} -(\eta_1 + \eta_2) \frac{T_h \omega^2}{2} + \frac{1}{h} \left(\frac{\gamma_1}{h} (1 + R_h - e^{iK_1 h} - R_h e^{-iK_1 h}) \right. \\ \left. - \frac{\gamma_2}{h} T_h (e^{-iK_2 h} - 1) \right) = 0. \end{aligned} \quad (10.10b)$$

From (10.10a) and (10.10b), we derive the following expression:

$$T_h = \frac{-2i\gamma_1 \sin K_1 h}{\gamma_2 e^{-iK_2 h} - \gamma_1 - \gamma_2 + \gamma_1 e^{-iK_1 h} + (\eta_1 + \eta_2) \frac{\omega^2 h^2}{2}}. \quad (10.11)$$

By taking into account (10.7a) and (10.7b) we get the following Taylor expansion of (10.11):

$$T_h = \frac{2\sigma_1}{\sigma_1 + \sigma_2} + \frac{\gamma_1 \eta_2 - \gamma_2 \eta_1}{4c_1 c_2 (\sigma_1 + \sigma_2)^2} \omega^2 h^2 + O(h^3). \quad (10.12)$$

This study shows that, for a second-order scheme, the error produced by a heterogeneity is of the same order as the scheme. Unfortunately, this will not always be the case for higher-order approximations.

10.1.3 Fourth-Order: First Approach

For the fourth-order scheme defined by using (8.11), a new phenomenon appears: for a given frequency ω and a velocity c , the dispersion relation

$$\omega^2 = \frac{4}{h^2} c^2 \sin^2 \frac{kh}{2} + \frac{4}{3h^2} c^2 \sin^4 \frac{kh}{2} \quad (10.13)$$

provides four values of k instead of two in the second-order case:

- Two real values $k = K(c, \omega h)$ and $k = -K(c, \omega h)$, where

$$K(c, \omega h) = \frac{2}{h} \arcsin \left(\frac{1}{2} \left(6 \left(1 + \frac{\omega^2 h^2}{3c^2} \right)^{\frac{1}{2}} - 6 \right)^{\frac{1}{2}} \right) \in [0, \frac{\pi}{h}],$$

- and two purely imaginary values $k = i\xi(c, \omega h)$ and $k = -i\xi(c, \omega h)$, where

$$\xi(c, \omega h) = \frac{2}{h} \operatorname{argsinh} \left(\frac{1}{2} \left(6 \left(1 + \frac{\omega^2 h^2}{3c^2} \right)^{\frac{1}{2}} + 6 \right)^{\frac{1}{2}} \right).$$

The waves corresponding to $k = \pm K(c, \omega h)$ propagate and are actually fourth-order approximations of the physical waves, so that

$$K(c, \omega h) = \frac{\omega}{c} \left(1 + \frac{\omega^4 h^4}{180c^4} + O(\omega^6 h^6) \right), \quad (10.14)$$

whereas the waves corresponding to $k = \pm\xi(c, \omega h)$ exist only in the half-space. They are stationary evanescent waves (which have no physical meaning and come from the approximation):

$$\exp i\omega t e^{-(\xi(c, \omega h)|x_\ell|)}, \quad (10.15)$$

whose penetration depth $p(h) = \xi(c, \omega h)^{-1}$ is asymptotically proportional to h when $h \rightarrow 0$. We have:

$$p(h) = \frac{h}{\beta}, \quad \beta = 2 \log_e(2 + \sqrt{3}). \quad (10.16)$$

The discrete solution is then of the form:

$$u_\ell(t) = e^{i(\omega t - K_1(\omega h)x_\ell)} + R_h e^{i(\omega t + K_1(\omega h)x_\ell)} + R'_h e^{i\omega t} e^{(-\xi_1(\omega h)|x_\ell|)}, \\ x_\ell \leq 0, \quad (10.17a)$$

$$u_\ell(t) = T_h e^{i(\omega t - K_2(\omega h)x_\ell)} + T'_h e^{i\omega t} e^{(-\xi_2(\omega h)|x_\ell|)}, \\ x_\ell \geq 0, \quad (10.17b)$$

where

$$K_j(\omega h) = K(c_j, \omega h) \quad j = 1, 2, \quad (10.18a)$$

$$\xi_j(\omega h) = \xi(c_j, \omega h) \quad j = 1, 2. \quad (10.18b)$$

R_h and T_h are the amplitudes of the propagating reflected and transmitted waves and R'_h and T'_h are those of the corresponding evanescent waves.

We must now find four relations from which we are able to derive the coefficients R_h, R'_h, T_h and T'_h . These relations are obtained, first by writing that the solution is continuous at $\ell = 0$, which can be written as

$$1 + R_h + R'_h = T_h + T'_h. \quad (10.19)$$

On the other hand, the three other equations are derived from the semi-discrete scheme at the points $\ell = -1, 0, 1$ (since, at the other points, the equation is written for homogeneous media):

$$\begin{aligned} \frac{(\eta_1 + \eta_2)}{2} \frac{d^2 u_0}{dt^2} &+ \frac{4}{3h} \left(((1 - \theta)\gamma_1 + \theta\gamma_2) \frac{u_0 - u_{-1}}{h} \right. \\ &\left. - (\theta\gamma_1 + (1 - \theta)\gamma_2) \frac{u_1 - u_0}{h} \right) \\ &- \frac{1}{6h} \left(\gamma_1 \frac{u_0 - u_{-2}}{2h} - \gamma_2 \frac{u_2 - u_0}{2h} \right) \quad \text{at } \ell = 0, \end{aligned} \quad (10.20)$$

$$\begin{aligned} \eta_1 \frac{d^2 u_{-1}}{dt^2} &+ \frac{4}{3h} \left(\gamma_1 \frac{u_{-1} - u_{-2}}{h} - ((1-\theta)\gamma_1 + \theta\gamma_2) \frac{u_0 - u_{-1}}{h} \right) \\ &- \frac{1}{6h} \left(\gamma_1 \frac{u_{-1} - u_{-3}}{2h} - \frac{(\gamma_1 + \gamma_2)}{2} \frac{u_1 - u_{-1}}{2h} \right) \end{aligned} \quad (10.21)$$

at $\ell = -1$,

$$\begin{aligned} \eta_2 \frac{d^2 u_1}{dt^2} &+ \frac{4}{3h} \left((\theta\gamma_1 + (1-\theta)\gamma_2) \frac{u_1 - u_0}{h} - \gamma_2 \frac{u_2 - u_1}{h} \right) \\ &- \frac{1}{6h} \left(\frac{(\gamma_1 + \gamma_2)}{2} \frac{u_1 - u_{-1}}{2h} - \gamma_2 \frac{u_3 - u_1}{2h} \right) \quad \text{at } \ell = 1. \end{aligned} \quad (10.22)$$

By inserting (10.17a) and (10.17b) into these three equations, we obtain the three other relations. Such a strategy naturally leads to a linear system in R_h, R'_h, T_h and T'_h that has the form:

$$M(\omega h) \begin{bmatrix} R_h \\ R'_h \\ T_h \\ T'_h \end{bmatrix} = F(\omega h), \quad (10.23)$$

where the matrix $M(\omega h)$ and the vector $F(\omega h)$ only depend on ωh and $(\eta_1, \gamma_1, \eta_2, \gamma_2)$. We do not give here their expressions which are very complicated.

We can use Maple or an equivalent software to solve the linear system (10.23) and to expand the solution for small values of ωh . We shall not develop here the details of these tedious calculations and simply give the following observations. Contrary to what happens for instance with the second-order scheme, the coefficients R_h, R'_h, T_h and T'_h are not real. This means that the phenomena of numerical transmission and reflection are accompanied by a phase-shift at the interface. Let us write:

$$R_h(\omega h) = \tilde{R}(\omega h) e^{i\varphi_R(\omega h)} \quad \tilde{R}(\omega h) > 0, \quad 0 \leq \varphi_R(\omega h) \leq \pi, \quad (10.24a)$$

$$R'_h(\omega h) = \tilde{R}'(\omega h) e^{i\varphi'_R(\omega h)} \quad \tilde{R}'(\omega h) > 0, \quad 0 \leq \varphi'_R(\omega h) \leq \pi, \quad (10.24b)$$

$$T_h(\omega h) = \tilde{T}(\omega h) e^{i\varphi_T(\omega h)} \quad \tilde{T}(\omega h) > 0, \quad 0 \leq \varphi_T(\omega h) \leq \pi, \quad (10.24c)$$

$$T'_h(\omega h) = \tilde{T}'(\omega h) e^{i\varphi'_T(\omega h)} \quad \tilde{T}'(\omega h) > 0, \quad 0 \leq \varphi'_T(\omega h) \leq \pi, \quad (10.24d)$$

where $\tilde{R}(\omega h)$ and $\tilde{R}'(\omega h)$ are the reflection coefficients, $\varphi_R(\omega h)$ and $\varphi'_R(\omega h)$ the corresponding phase-shifts, $\tilde{T}(\omega h)$ and $\tilde{T}'(\omega h)$ the transmission coefficients and $\varphi_T(\omega h)$ and $\varphi'_T(\omega h)$ the corresponding phase-shifts.

In fact, the phase-shifts $\varphi'_R(\omega h)$ and $\varphi'_T(\omega h)$ associated with the parasitic evanescent waves are not really troublesome since these waves do not propagate and their amplitudes ($R'(\omega h)$ and $T'(\omega h)$) go to 0, as we shall see, as ωh tends to 0. Conversely, the presence of the terms $\varphi_R(h)$ and $\varphi_T(h)$ can generate disturbing errors since they affect propagating waves of constant amplitude: in the case of several interfaces their effect could be likened to some kind of dispersion. Thus, we would like to decrease as much as possible these phase-shift terms. Let us describe in more detail the asymptotic results when ωh tends to 0. We set:

$$\phi(\theta) = \sqrt{3} \left\{ (16\theta - 1)^2 - 128\sqrt{3}\theta^2 \right\}, \quad (10.25a)$$

$$\Psi(\theta) = 128(4\sqrt{3} - 7\theta^2 + 32(2 - \sqrt{3})\theta + \sqrt{3} - 2), \quad (10.25b)$$

$$D(\theta) = (\gamma_2 - \gamma_1)^2 \Psi(\theta) + 8\gamma_1\gamma_2. \quad (10.25c)$$

Then one can show that concerning the reflection coefficients of the non-parasitic wave:

$$\tilde{R}(\omega h) = \frac{\sigma_1 - \sigma_2}{\sigma_1 + \sigma_2} (1 + O(\omega^2 h^2)), \quad (10.26a)$$

$$\tilde{T}(\omega h) = \frac{2\sigma_1}{\sigma_1 + \sigma_2} (1 + O(\omega^2 h^2)), \quad (10.26b)$$

which means that, as in the second order case, these coefficients are second order accurate. For the parasitic waves we have:

$$\tilde{R}'(\omega h) = \frac{\sigma_2(\gamma_2 - \gamma_1)}{3(\sigma_1 + \sigma_2)} \frac{\{(\gamma_1 - \gamma_2)\phi(\theta) - 48\theta\gamma_1\}}{D(\theta)} \frac{\omega h}{c_1} (1 + O(\omega^2 h^2)), \quad (10.27a)$$

$$\tilde{T}'(\omega h) = \frac{\sigma_1(\gamma_2 - \gamma_1)}{3(\sigma_1 + \sigma_2)} \frac{\{(\gamma_1 - \gamma_2)\phi(\theta) - 48\theta\gamma_1\}}{D(\theta)} \frac{\omega h}{c_2} (1 + O(\omega^2 h^2)). \quad (10.27b)$$

These coefficients tend to zero proportionally to ωh . As the penetration length of the corresponding evanescent waves is also $O(\omega h)$, the L^1 norm of the transmitted and reflected parasitic waves is $O(\omega^2 h^2)$.

Concerning the phase-shifts $\varphi_R(\omega h)$ and $\varphi_T(\omega h)$, we have:

$$\varphi_R(\omega h) = \frac{\eta_2(\gamma_1 - \gamma_2)^2(\gamma_1 + \gamma_2)}{3(\sigma_1 + \sigma_2)^2 D(\theta)} \phi(\theta) \frac{\omega h}{c_1} (1 + O(\omega^2 h^2)), \quad (10.28a)$$

$$\varphi_T(\omega h) = \frac{\eta_1(\gamma_1 - \gamma_2)^2(\gamma_1 + \gamma_2)}{3(\sigma_1 + \sigma_2)^2 D(\theta)} \phi(\theta) \frac{\omega h}{c_2} (1 + O(\omega^2 h^2)). \quad (10.28b)$$

These formulas show that, in the general case, the parasitic phase shifts go to 0 proportionally to ωh and that the proportionality constant is equal to 0 if and only if one of the two following conditions is satisfied:

$$(i) \quad \gamma_1 = \gamma_2, \quad (10.29a)$$

$$(ii) \quad \phi(\theta) = 0. \quad (10.29b)$$

(i) shows that there is no problem when γ is constant and (ii) proves that there is an optimal choice for the averaging coefficient λ corresponding to $\phi(\lambda) = 0$. This equation has only one solution in the interval $]0, 1/2[$ which is given by:

$$\theta = \theta^* = \frac{1}{16 \left(1 + \left(\frac{3}{4} \right)^{\frac{1}{4}} \right)} \simeq 0.032\,373\,274. \quad (10.30)$$

For this value, we obtain third-order phase-shifts for $\varphi_R(\omega h)$ and $\varphi_T(\omega h)$.

Remarks

1. One can see that

$$\frac{1}{32} < \theta^* < \frac{1}{16}.$$

In other words, θ^* is in the region for which we cannot conclude with certainty about the positivity of the discrete heterogeneous operator.

2. One could make θ depend on the node to which it is applied. This would lead to a more precise (but tedious) analysis.
3. When $\gamma_1 = \gamma_2$, φ_R and φ_T are in $O(h^3)$ for any value of θ .

To illustrate numerically the importance of the choice of λ on the reflection-transmission phenomena we have computed the coefficients $R(\omega h)$, $T(\omega h)$, $R'(\omega h)$ and $T'(\omega h)$ associated with the medium:

$$\eta_1 = 1, \quad \gamma_1 = 2, \quad (10.31a)$$

$$\eta_2 = 2, \quad \gamma_2 = 8, \quad (10.31b)$$

for $\theta = 0$ and $\theta = \theta^*$:

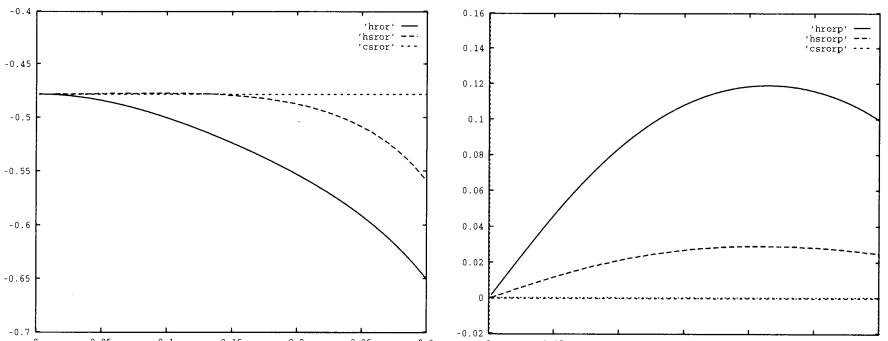


Fig. 10.1. $|R|$ (left) and $|R'|$ (right) for $\theta = 0$ (in continuous line) and $\theta = \theta^*$ (in dashed line). The straight line corresponds to the exact value

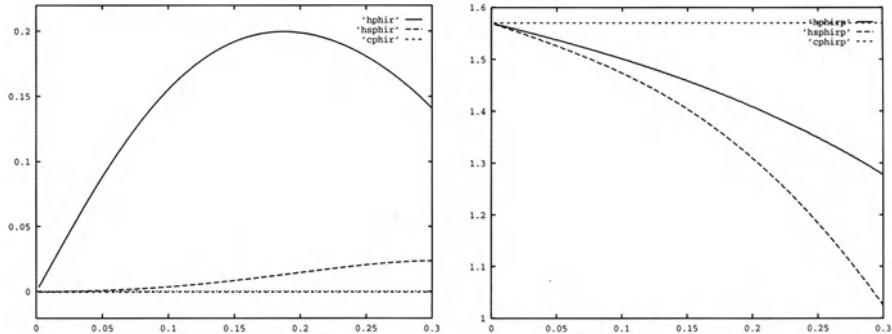


Fig. 10.2. φ_R (left) and φ'_R (right) for $\theta = 0$ (continuous line) and $\theta = \theta^*$ (in dashed line). The straight line corresponds to the exact value

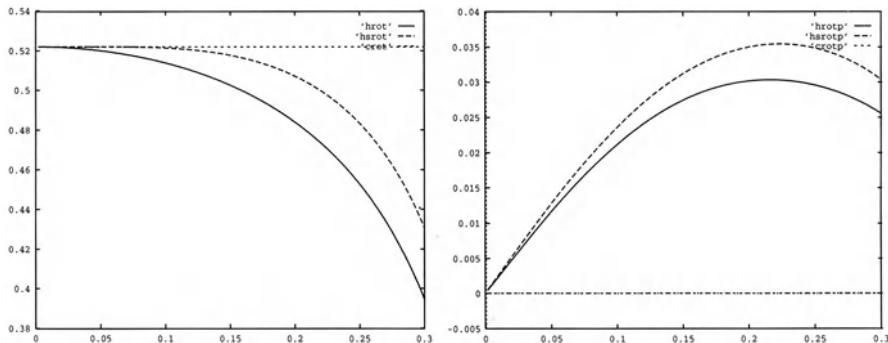


Fig. 10.3. $|T|$ (left) and $|T'|$ (right) for $\theta = 0$ (in continuous line) and $\theta = \theta^*$ (in dashed line). The straight line corresponds to the exact value

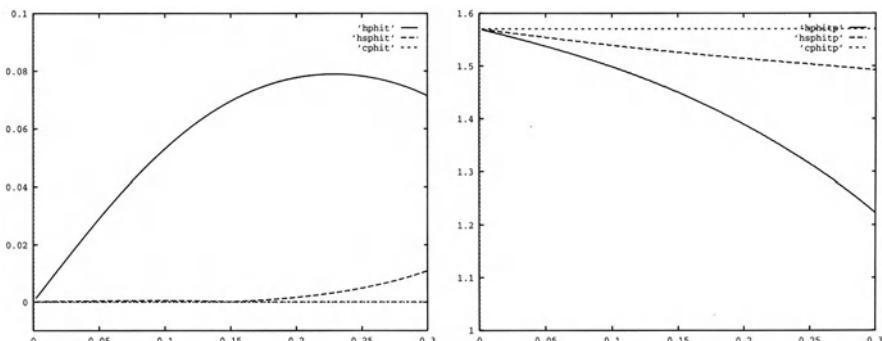


Fig. 10.4. φ_T (left) and φ'_T (right) for $\theta = 0$ (in continuous line) and $\theta = \theta^*$ (in dashed line). The straight line corresponds to the exact value

10.1.4 A Numerical Study

We consider the following model problem:

$$\eta(x) \frac{\partial^2 u}{\partial t^2} - \frac{\partial}{\partial x} (\gamma(x) \frac{\partial u}{\partial x}) = 0, \quad x \in]0, 12[, t \in]0, 48[, \quad (10.32a)$$

$$u(0, t) = u(12, t) = 0, \quad (10.32b)$$

$$u(x, 0) = u_0(x), \quad \frac{\partial u}{\partial t}(x, 0) = 0, \quad (10.32c)$$

where $\eta(x)$ and $\gamma(x)$ are piecewise constant in 6 intervals of equal length. The values of the coefficients are given in the following table:

Table 10.1. The values of η and γ for the six-layer experiment

Layers	1	2	3	4	5	6
$\eta(x)$	1	1	3	2	4	2
$\gamma(x)$	1	2	4	1	3	5
$c(x)$	1	1,414	1,155	0,707	0,866	1,581

The initial data is given by a (truncated) Gaussian function centered at $x = 6$. The minimal wavelength of this initial data is equal to $1/2$. As the mean velocity of our medium is approximately equal to 1, the wave propagates for a duration of 48 s, along approximatively 96 wavelengths so such an experiment corresponds to very large integration times for which it is well known that second-order schemes are, because of their dispersion, insufficiently accurate.

Moreover, as 96 wavelengths correspond to four times the total size of the integration domain, during the propagation the wave crosses approximately 20 interfaces. Such an example is thus a very good test to evaluate the importance of the approximation of transmission and reflection phenomena.

We have performed the following 10 numerical tests :

- | | |
|--------------------------------------|---|
| S_2^2
S_4^2
S_4^4
N | denotes the second-order scheme in space and time
denotes the second-order in time and fourth-order in space scheme
denotes the fourth-order in space and time scheme
denotes the number of grid points per wavelength |
|--------------------------------------|---|

Test No. 1 : Scheme $S_2^2, N = 7.5$.

Test No. 2 : Scheme $S_2^2, N = 15$.

Test No. 3 : Scheme $S_4^4, N = 7.5, \theta = 0$.

Test No. 4 : Scheme $S_4^4, N = 7.5, \theta = \theta^*$.

Test No. 5 : Scheme $S_4^4, N = 15, \theta = 0$.

Test No. 6 : Scheme $S_4^4, N = 15, \theta = \theta^*$.

Test No. 7 : Scheme $S_4^2, N = 7.5, \theta = 0, \alpha = 0.6$.

Test No. 8 : Scheme $S_4^2, N = 7.5, \theta = \theta^*, \alpha = 0.6$.

Test No. 9 : Scheme $S_4^2, N = 7.5, \theta = 0, \alpha = \alpha_{max}$.

Test No. 10 : Scheme $S_4^2, N = 7.5, \theta = \theta^*, \alpha = \alpha_{max}$.

A reference “exact solution” has been computed with the fourth-order scheme S_4^4 and 50 grid points per wavelength. In each case, except for tests No. 7 and No. 8, we have chosen the maximal time step allowed by stability condition. In Figs. 10.5–10.9 we present, for each of our experiments, two pictures representing the solution at time $t = 48$. In each case, the exact solution (continuous line) is compared with the exact one (dotted line).

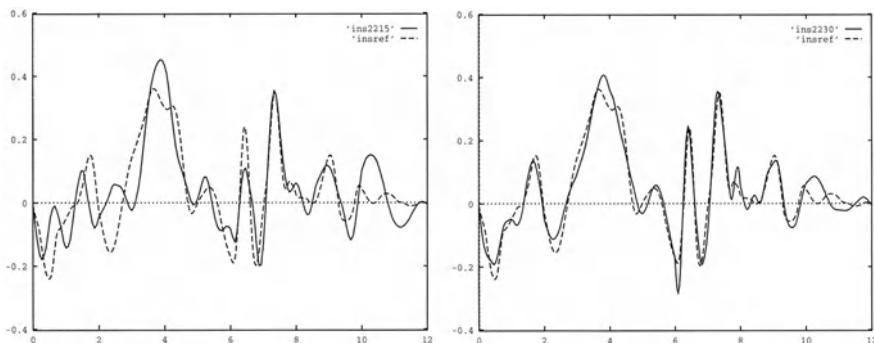


Fig. 10.5. Test No. 1 (*left*) and test No. 2 (*right*)

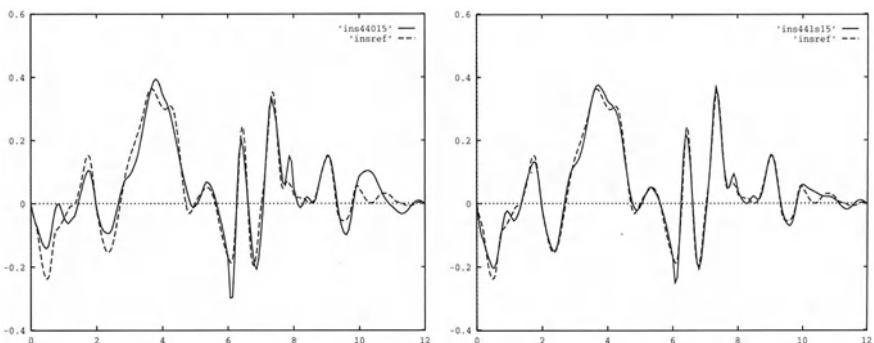


Fig. 10.6. Test No. 3 (*left*) and test No. 4 (*right*)

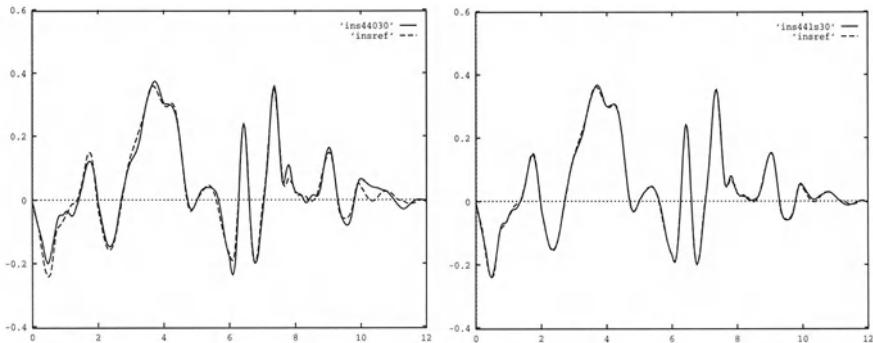


Fig. 10.7. Test No. 5 (left) and test No. 6 (right)

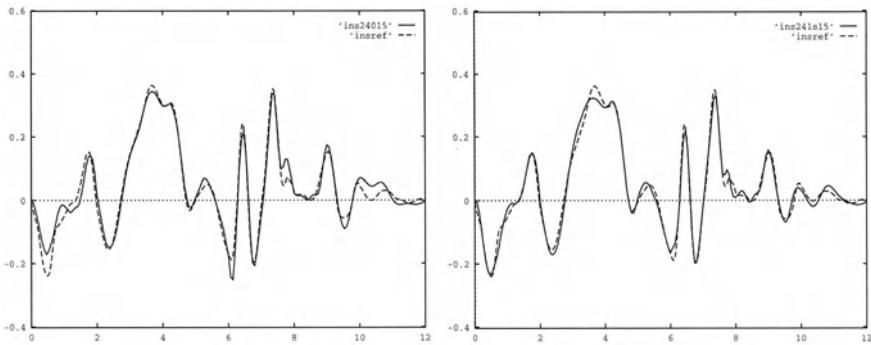


Fig. 10.8. Test No. 7 (left) and test No. 8 (right)

It clearly appears that the results of test Nos. 2 and 3 are roughly equivalent. This shows that for a given precision, one needs twice as many grid points with the second-order scheme than for the fourth-order one. By extrapolation, this would mean that the number of cells can be, when one uses, a fourth-order scheme instead of a second-order one, divided by 4 in dimension 2 and by 8 in dimension 3. As the time-step, because of the stability condition, is proportional to the space-step, this would imply a division of the total number of calculated values by 4 in dimension 1, by 8 in dimension 2 and by 16 in dimension 3.

On the other hand, tests Nos. 5 and 6 show that the change from $\theta = 0$ to $\theta = \theta^*$ gives a decrease of the error by a factor 2 and a stabilization of this error with respect to time (compare tests Nos. 4 and 5).

The comparison of tests 1, 3 and 7 shows that the precision obtained with the scheme S_4^2 for $\alpha = \alpha_M$ is intermediate between the one obtained with the scheme S_2^2 and the one given by the scheme S_4^4 . Moreover, the tests Nos. 7

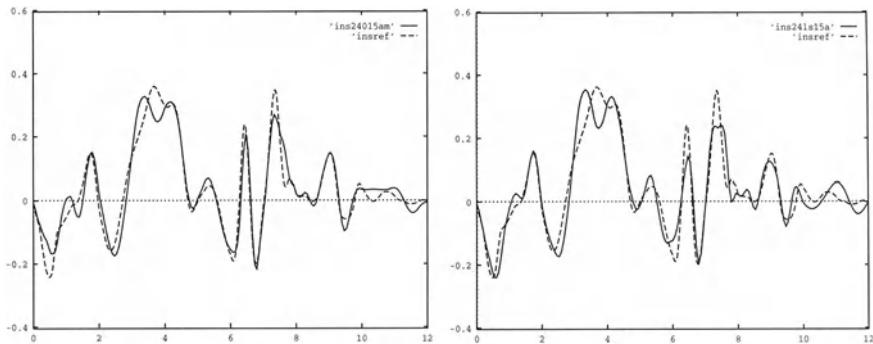


Fig. 10.9. Test No. 9 (*left*) and test No. 10 (*right*)

and 8 show that the gain of precision due to the change from $\theta = 0$ to $\theta = \theta^*$ is less sensitive. This means that the influence of θ is greater for the scheme S_4^4 than for the scheme S_4^2 . This is not surprising if we remember that the optimal value θ^* has been computed only for the semi-discrete scheme (i.e. for which the time remains continuous) which is clearly closer to scheme S_4^4 than to scheme S_4^2 . We can see that even with $\theta = 0$ (tests Nos. 3, 5 and 7) the calculations remained stable. This is due to the fact that the function $\gamma(x)$ does not present large discontinuities. It would be interesting (but tedious) to make the reflection-transmission analysis of the fully discretized scheme S_4^4 in order to see if θ^* is still an optimal value for the parameter θ .

10.2 The 2D Case

10.2.1 The Continuous Problem

Let us consider the following two-layer unbounded medium:

$$(\gamma(x, y), \eta(x, y)) = \begin{cases} (\gamma_1, \eta_1) & \text{for } x < 0, \\ (\gamma_2, \eta_2) & \text{for } x > 0, \end{cases} \quad (10.33)$$

and an incident wave whose wave vector is $\mathbf{K}_I = (K_{Ix}, K_{Iy})$ in the first layer. $\mathbf{K}_R = (K_{Rx}, K_{Ry})$ is the wave vector of the reflected wave and $\mathbf{K}_T = (K_{Tx}, K_{Ty})$ that of the transmitted wave.

Following Descarte's (or Snell's) law, we have

$$K_{Ix} = -K_{Rx}, \quad K_{Iy} = K_{Ry} \quad (10.34)$$

and, on the other hand, if we set $\mathbf{K}_I = \mathbf{K}_1$ and $\mathbf{K}_T = \mathbf{K}_2$, we know that

$$|\mathbf{K}_1| \sin \alpha_1 = |\mathbf{K}_2| \sin \alpha_2, \quad (10.35)$$

where α_j is defined by

$$\tan \alpha_j = \frac{K_{jy}}{K_{jx}} \quad j = 1, 2. \quad (10.36)$$

Then, the solution u of the problem can be written as

$$\begin{aligned} u(x, t) = & \chi_{\mathbb{R}^- \times \mathbb{R}}(x, y) \left(e^{i(\omega t - K_{1x}x - K_{1y}y)} + R e^{i(\omega t + K_{1x}x - K_{1y}y)} \right) \\ & + T \chi_{\mathbb{R}^+ \times \mathbb{R}}(x, y) e^{i(\omega t - K_{2x}x - K_{2y}y)}, \end{aligned} \quad (10.37)$$

where, as in 1D, R and T are the coefficients of reflection and transmission, $\chi_{\mathbb{R}^+ \times \mathbb{R}}$ and $\chi_{\mathbb{R}^- \times \mathbb{R}}$ are the characteristic functions of the sets $\mathbb{R}^+ \times \mathbb{R}$ and $\mathbb{R}^- \times \mathbb{R}$. Moreover, $\mathbf{K}_1 = (K_{1x}, K_{1y})$ and $\mathbf{K}_2 = (K_{2x}, K_{2y})$ satisfy the following dispersion relations in each layer:

$$K_{1x}^2 + K_{1y}^2 = \frac{\omega^2}{c_1^2}, \quad (10.38a)$$

$$K_{2x}^2 + K_{2y}^2 = \frac{\omega^2}{c_2^2}. \quad (10.38b)$$

From (10.36) and (10.38a) and (10.38b), we obtain:

$$K_{1x}^2 = \frac{\omega^2 \cos^2 \alpha_1}{c_1^2}, \quad (10.39a)$$

$$K_{2x}^2 = \frac{\omega^2 \cos^2 \alpha_2}{c_2^2}. \quad (10.39b)$$

On the other hand, we obtain, from (10.35) and (10.38a) and (10.38b), the Descarte's law:

$$\frac{\sin^2 \alpha_1}{c_1^2} = \frac{\sin^2 \alpha_2}{c_2^2}, \quad (10.40)$$

from which we derive, by using (10.39a) and (10.39b):

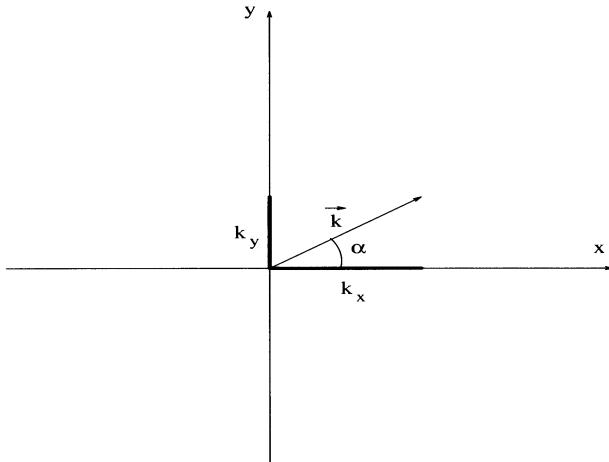
$$K_{1x}^2 = \frac{\omega^2 (1 - \sin^2 \alpha_1)}{c_1^2}, \quad (10.41a)$$

$$K_{2x}^2 = \frac{\omega^2}{c_1^2 c_2^2} (c_1^2 - c_2^2 \sin^2 \alpha_1). \quad (10.41b)$$

The continuity of u and that of the trace of $\gamma \mathbf{grad} u \cdot \mathbf{n}$ at the interface provide the following system:

$$1 + R = T, \quad (10.42a)$$

$$\gamma_1 K_{1x} (1 - R) = \gamma_2 K_{2x} T. \quad (10.42b)$$

**Fig. 10.10.** Definition of α

After some computations, one can show, by using (10.41a) and (10.41b) and (10.42a) and (10.42b), that the reflection and transmission coefficients are given by the following relations:

$$R = \frac{\sigma_1 - \sigma_2 \sqrt{\frac{1 - (c_2/c_1)^2 \sin^2 \alpha_1}{1 - \sin^2 \alpha_1}}}{\sigma_1 + \sigma_2 \sqrt{\frac{1 - (c_2/c_1)^2 \sin^2 \alpha_1}{1 - \sin^2 \alpha_1}}}, \quad (10.43a)$$

$$T = \frac{2\sigma_1}{\sigma_1 + \sigma_2 \sqrt{\frac{1 - (c_2/c_1)^2 \sin^2 \alpha_1}{1 - \sin^2 \alpha_1}}}, \quad (10.43b)$$

where σ_1 and σ_2 are defined as in the 1D case.

One can easily see that there exist a transmitted wave for an angle $\alpha_1 < \alpha_L$, where $\alpha_L = \arcsin(c_1/c_2)$, only if the two velocities satisfy the inequality: $c_1/c_2 \leq 1$.

Of course, for $\alpha_1 = 0$, we obtain the 1D case.

10.2.2 Second-Order Scheme

We are under the conditions defined in (10.33) and, with the same notation, relations (10.34), (10.35) and (10.36) remain true. Then, as in the 1D case, the solution can be written as

$$u(x, t) = e^{i(\omega t - \ell K_{1x} h - m K_{1y} h)} + R_h e^{i(\omega t + \ell K_{1x} h - m K_{1y} h)} \text{ for } \ell \leq 0, \quad (10.44a)$$

$$u(x, t) = T_h e^{i(\omega t - \ell K_{2x} h - m K_{2y} h)} \text{ pour } \ell \geq 0. \quad (10.44b)$$

K₁ = (K_{1x}, K_{1y}) and **K₂** = (K_{2x}, K_{2y}) satisfy, in each layer, the following dispersion relations:

$$\omega^2 = \frac{4c^2}{h^2} \left(\sin^2 \frac{K_{1x} h}{2} + \sin^2 \frac{K_{1y} h}{2} \right), \quad (10.45a)$$

$$\omega^2 = \frac{4c^2}{h^2} \left(\sin^2 \frac{K_{2x} h}{2} + \sin^2 \frac{K_{2y} h}{2} \right). \quad (10.45b)$$

By taking into account (10.36), these relations can be written as

$$\omega^2 = \frac{4c^2}{h^2} \left(\sin^2 \frac{K_{1x} h}{2} + \sin^2 \frac{K_{1x} \tan \alpha_1 h}{2} \right), \quad (10.46a)$$

$$\omega^2 = \frac{4c^2}{h^2} \left(\sin^2 \frac{K_{2x} h}{2} + \sin^2 \frac{K_{2x} \tan \alpha_2 h}{2} \right). \quad (10.46b)$$

Unfortunately, these relations are implicit in K_{1x} and K_{2x} and, therefore, these quantities cannot be expressed in terms of α₁, α₂ and h. However, since our purpose is only to obtain a Taylor expansion in h of the reflection-transmission coefficients, it is sufficient to deduce it from (10.46a) and (10.46b). We obtain:

$$\begin{aligned} K_{jx} = & \frac{\omega \cos \alpha_j}{c_j} \left(1 + \frac{\omega^2}{24c_j^2} (1 - 2 \cos^2 \alpha_j + 2 \cos^4 \alpha_j) h^2 \right. \\ & \left. + \frac{\omega^4}{5760c_j^4} (27 - 116 \cos^2 \alpha_j + 256 \cos^4 \alpha_j \right. \\ & \left. - 280 \cos^6 \alpha_j + 140 \cos^8 \alpha_j) h^4 \right) + O(h^6) \quad \forall j = 1, 2. \end{aligned} \quad (10.47)$$

The semi-discrete equation at ℓ = 0 can be written as

$$\begin{aligned} \frac{(\eta_1 + \eta_2)}{2} \frac{d^2 u_h}{dt^2} + \frac{1}{h} \left(\gamma_1 \frac{u_{0,0} - u_{-1,0}}{h} - \gamma_2 \frac{u_{1,0} - u_{0,0}}{h} \right. \\ \left. + \frac{(\gamma_1 + \gamma_2)}{2} \frac{u_{0,0} - u_{0,-1}}{h} \right. \\ \left. - \frac{(\gamma_1 + \gamma_2)}{2} \frac{u_{0,1} - u_{0,0}}{h} \right) = 0. \end{aligned} \quad (10.48)$$

By replacing u by its value defined in (10.44a) and (10.44b) and by writing the continuity condition at ℓ = 0, we obtain, after some computations (aided by Maple):

$$\begin{aligned}
T_h = & -4i\gamma_1 \sin(K_{1x} h) (16\gamma_2 \cos^4 \xi \\
& + 16\gamma_1 \cos^4 \xi + \omega^2 h^2 \eta_2 - 2i\gamma_2 \sin(K_{2x} h) \\
& - 16\gamma_1 \cos^2 \xi + 2\gamma_2 \cos(K_{2x} h) - 2\gamma_1 - 2\gamma_2 \\
& - 2i\gamma_1 \sin(K_{1x} h) + 2\gamma_1 \cos(K_{1x} h) \\
& - 16\gamma_2 \cos^2 \xi + \omega^2 h^2 \eta_1)^{-1}
\end{aligned} \tag{10.49}$$

where: $\xi = \frac{K_{1x} h}{4} \tan \alpha_1$,

whose Taylor expansion, after taking into account (10.47), can be written as

$$T_h = C_0 + C_1 h + C_2 h^2, \tag{10.50}$$

where

$$C_0 = \frac{2\sigma_1 \cos \alpha_1}{d_1},$$

$$C_1 = -i \frac{\omega \cos \alpha_1}{c_1 \gamma_1 \gamma_2 d_1^2} \left(\frac{\sin^2 \alpha_2}{c_2^2} - \frac{\sin^2 \alpha_1}{c_1^2} \right)$$

and

$$\sigma_j = \sqrt{\eta_j \gamma_j}, \quad j = 1, 2,$$

$$d_1 = \sigma_1 \cos \alpha_1 + \sigma_2 \cos \alpha_2.$$

Since our purpose is to obtain an error in $O(h^2)$, we shall not give the expression of C_2 , which is rather complicated.

The presence of C_1 would mean that the global error on T_h is in $O(h)$. Actually, the first-order scheme will be cancelled by using the following property:

Let us write a Taylor expansion of the dispersion relations given in (10.45a) and (10.45b):

$$K_{1x}^2 + K_{1y}^2 - \frac{\omega^2}{c_1^2} + \frac{1}{12} (K_{1x}^4 + K_{1y}^4) h^2 + O(h^4), \tag{10.51a}$$

$$K_{2x}^2 + K_{2y}^2 - \frac{\omega^2}{c_2^2} + \frac{1}{12} (K_{2x}^4 + K_{2y}^4) h^2 + O(h^4). \tag{10.51b}$$

By multiplying (10.51a) by $\sin^2 \alpha_1$ and (10.51b) by $\sin^2 \alpha_2$ and taking into account (10.35), we obtain:

$$\begin{aligned} \frac{\sin^2 \alpha_1}{c_1^2} = & \frac{\sin^2 \alpha_2}{c_2^2} + \frac{1}{12\omega^2} ((K_{1x}^4 + K_{1y}^4) \sin^2 \alpha_1 \\ & + (K_{2x}^4 + K_{2y}^4) \sin^2 \alpha_2) h^2 + O(h^4). \end{aligned} \quad (10.52)$$

Finally, by inserting this expression into C_1 , one can easily see that $C_1 = O(h^2)$ and, therefore, we actually obtain a second-order truncation error on T_h .

Remarks

1. Equation (10.52) shows that this scheme provides a second-order approximation of the Descarte law.
2. Since $C_1 = O(h^2)$, an additional term in h^3 is added to the third-order term.
3. Equation (10.47) shows that, in fact, one can replace $K_{jx}^4 + K_{jy}^4$ by $(\omega^4 \cos^4 \alpha_j / c_j^4)(1 + \tan^4 \alpha_j)$, $j = 1, 2$ in (10.52).
4. The extension of such a study to fourth-order approximations requires us to know the Taylor expansion of the wavenumbers associated with the physical and evanescent waves. The first expansion is easy to obtain but the second one leads to an implicit equation which is impossible to solve analytically. So, the reflection and transmission coefficients must then be computed numerically.

Besides the problem of the order of the reflection and transmission amplitudes, which is generally much lower than the order of the method in homogenous media, in 2D and 3D, another drawback appears. Since finite difference methods are constructed on regular grids, the interfaces which are not parallel to the axes are approximated by staircases which generate numerical diffracted waves. This diffraction increases when one uses higher-order methods because these methods require less points per wavelength and so, the staircases are larger. A classical technique mainly used by geophysicists consists in averaging the heterogeneities along the interfaces [87, 122], i.e. in computing the mean values of the coefficients on the cells crossed by an interface.

For the case described in Fig. 10.11, for instance, the value at the center of the cell which would be ρ_2 is replaced by the mean value:

$$\frac{1}{h^2} \int_C \rho(\mathbf{x}) d\mathbf{x} = \frac{1}{h^2} (\text{mes}(A_1)\rho_1 + \text{mes}(A_2)\rho_2). \quad (10.53)$$

This technique is rather efficient, but it is not obvious to implement and, moreover, it does not provide any control on the order of the transmitted and reflected waves.

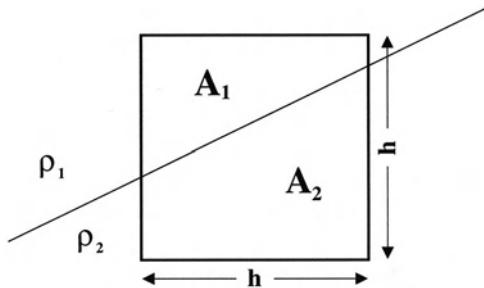


Fig. 10.11. An interface in ρ crossing a 2D cell \mathcal{C}

10.2.3 A Numerical Experiment

In conclusion, we shall present a numerical experiment in a 2D non-homogeneous acoustic medium. Using this model, we compare a second-order and a fourth-order approximation.

On a square domain $[0, 10] \times [0, 10]$ in which $\eta = \gamma = 1$ with, at its center, a disk of diameter 4 and in which $\eta = \gamma = 4$, we propagate a wavelet of frequency 4 at a source of coordinates (1,5). The values of η and γ show that the velocity is equal to 1 in the whole domain (Fig. 10.12).

On the upper part of the boundary, we have $u = 0$ and a second-order absorbing boundary condition (approximated by an uncentered finite difference scheme) is applied to the three other parts of the boundary. Here, $\theta = \theta^*$ but taking $\theta = 0$ does not make a significant difference in this experiment. Actually, it seems that the 2D case is less sensitive to the value of θ than the 1D case.

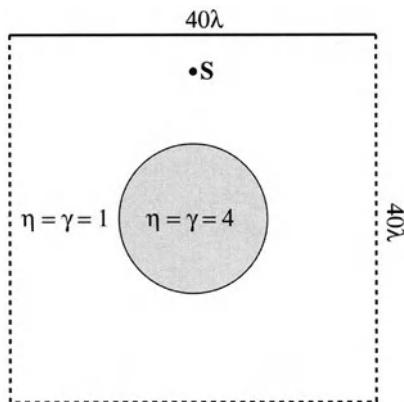


Fig. 10.12. The configuration of the experiment

In Fig. 10.13, we show four snapshots of the solution obtained by a fourth-order approximation in space based on the first approach of the modified equation.

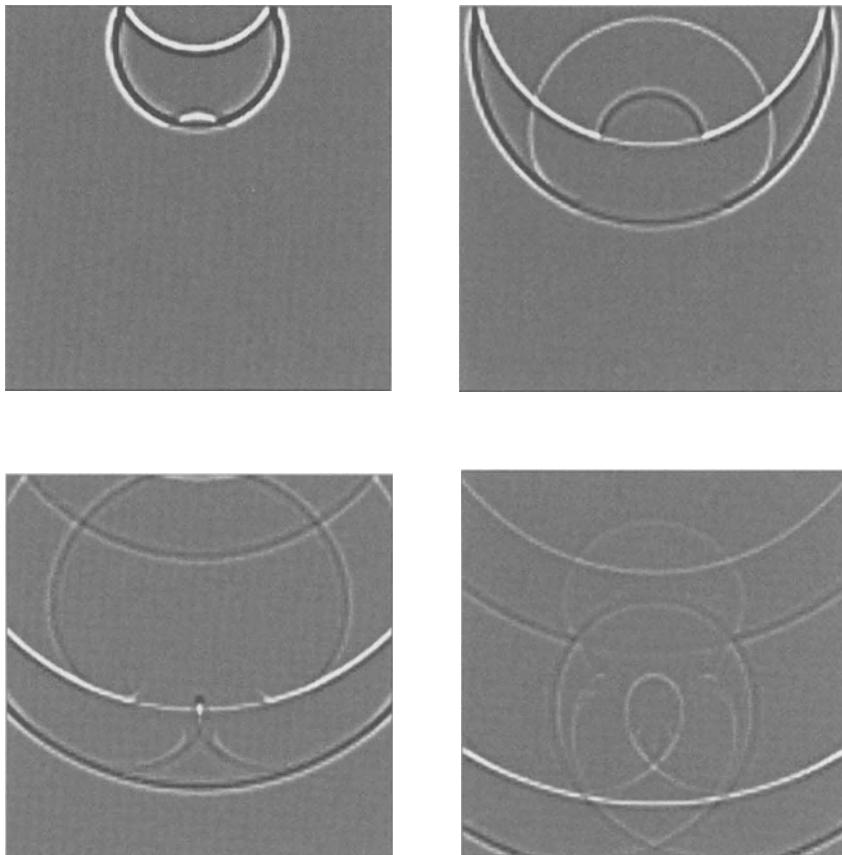


Fig. 10.13. Snapshots of the solution for $t = 2.5, 5, 7.5, 10$ with a fourth-order approximation in space and time and 10 points per wavelength

In Figs. 10.14–10.16, we give the *seismograms*¹ (i.e. the curve giving the solution versus the time) of the solution at the point $(5,9)$ (i.e. when the wave has propagated across almost the whole domain) for different approximations (the curves are continuous lines) compared with the “exact” solution (dashed lines) obtained by the fourth-order scheme with 15 points per wavelength.

¹ In geophysics, one would say the “traces of the seismogram”.

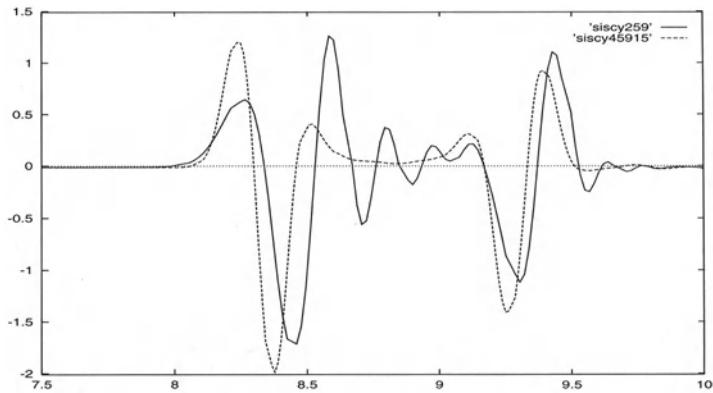


Fig. 10.14. Seismogram for the second-order scheme and 10 points per wavelength

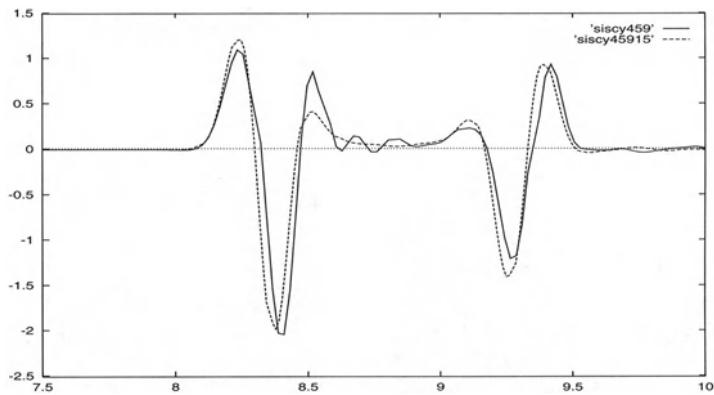


Fig. 10.15. Seismogram for the fourth-order scheme and 8 points per wavelength. The CPU time (on a DEC station) is almost the same as that of the the second-order scheme and 10 points per wavelength

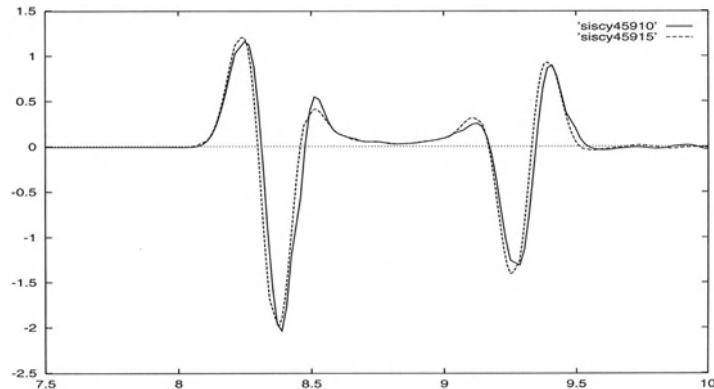


Fig. 10.16. Seismogram for the fourth-order scheme and 10 points per wavelength

Part III

Finite Element Methods

Introduction

It is evident that finite difference methods are inappropriate for treating complex geometries. As we said previously, for the wave equation, a staircase approximation of a reflecting boundary generates parasitic diffraction phenomena which can seriously damage the solution. Moreover, modeling anisotropic media on staggered grids is a very difficult problem.

Of course, the solution to all these problems is the use of finite element methods (FEM) for solving wave equations. However, for a long time, such an approach seemed unrealistic, at least for higher-order approximations, since such methods introduce a mass-matrix. This could be stored as an n -diagonal band matrix, but whose number of diagonals increases with the dimension in space and the order of the approximation. This non-negligible matrix has to be inverted at each time-step, which substantially increases the cost of the FEM.

The first step towards the solution of this difficult problem was found by Hennart [65] for the heat equation and, independently, by Young [121] for reservoir simulation, at the end of the 1980s and consisted in using the Gauss-Lobatto quadrature formula for computing the mass-matrix. The theoretical justification of this approach is based on some exercises in Ciarlet's basic book on finite element theory [24].

The method was found and it remained to apply it to the wave equation. This was done thanks to the annual visit of Jean-Pierre Hennart to our group in 1990, during which I asked him if he had an idea about mass-lumping for the wave equation. Of course, the analysis of the method for the wave equation had to be done and was partly developed in Nathalie Tordjman's thesis [115] in which this concept was also extended to triangles. The extension to tetrahedra was studied by Mulder [88] and its application to the elastics system was realized by Komatitsch and Vilotte [75].

Since the p -version of this method appeared at the end of the 1990s in the spectral methods community under the name of "spectral elements" [83], this name became the canonical one for the method.

When I met Peter Monk at the Wave Conference I organized in 1991, he asked me to extend this approach to the Maxwell equations. The proposal was challenging since we had to use mixed finite elements for these equations. A first step was done by using the so-called "first family of Nédélec's elements"

[92]. But the method obtained held only on orthogonal meshes and was not able to treat anisotropy [38]. The ultimate step was reached by using the “second family of Nédélec’s elements” [93] which enabled us to overcome these difficulties and led to a very cheap formulation of mixed finite elements [39]. In parallel, Alexandre Elmkes, in his thesis directed by Patrick Joly, studied the extension of this approach to triangles and tetrahedra [50, 51, 52].

Finally, the efficiency of the mixed FEM obtained for the Maxwell equations suggested to me to extend this approach to the wave equation expressed as a system and to the elastics system. This idea, developed in Sandrine Fauqueux’s thesis [55], led to a mathematical decomposition of the spectral element method which provided a cheap algorithm and enabled us to treat unbounded domains in an easy way [31]. Moreover, Sandrine Fauqueux completed the analysis of the spectral elements for wave equations.

The eleventh chapter will be devoted to the construction and the analysis of mass-lumped continuous finite elements for the 1D wave equation. Although a priori simple, this problem will point out the main difficulties encountered for this approximation and will provide the fundamental tools for constructing and analyzing spectral elements in 2D and 3D. The twelfth chapter will extend these results to quadrilateral and hexahedral elements which are the actual spectral elements. The non-obvious extension of the mass-lumping technique to triangles and tetrahedra will be given in its last section. The purpose of the following chapter will be to study mass-lumping for mixed finite elements, first for Maxwell equations, then for the acoustics equation and the elastics system. Finally, the fourteenth and concluding chapter will deal with modeling unbounded domains.

Although error analysis is the classical way for studying finite element methods, we preferred to use plane wave analysis because the results given by error analysis are too general to obtain precise informations on the methods in the case of wave equations.

11. Mass-Lumping in 1D

11.1 Basic Approximations

11.1.1 Construction of Mass-Lumped Finite Elements

Let us consider the following 1D problem:

$$\left\{ \begin{array}{l} \text{Find } u : \mathbb{R} \times]0, T[\rightarrow \mathbb{R} \text{ such that:} \\ \frac{\partial^2 u}{\partial t^2}(x, t) - c^2 \frac{\partial^2 u}{\partial x^2}(x, t) = 0 \quad \text{in } \mathbb{R} \times]0, T[, \\ u(x, 0) = u_0(x), \quad \frac{\partial u}{\partial t}(x, 0) = u_1(x) \quad \text{in } \mathbb{R}, \end{array} \right. \quad (11.1)$$

whose variational formulation is:

$$\left\{ \begin{array}{l} \text{Find } u(., t) \in H^1(\mathbb{R}), \quad t \in]0, T[\quad \text{such that:} \\ \frac{d^2}{dt^2} \int_{\mathbb{R}} u v \, dx + c^2 \int_{\mathbb{R}} \frac{\partial u}{\partial x} \frac{\partial v}{\partial x} \, dx = 0 \quad \forall v \in H^1(\mathbb{R}), \\ u(x, 0) = u_0(x), \quad \frac{\partial u}{\partial t}(x, 0) = u_1(x) \quad \text{in } \mathbb{R}. \end{array} \right. \quad (11.2)$$

In this section, we shall develop the semi-discretization in space of this problem, which is the main point of this study.

Let $V_h^r(\mathbb{R}) = \{v \in C^0(\mathbb{R}) \mid \forall p \in \mathbb{Z}, v|_{[x_p, x_{p+1}]} \in P_r\}$, where P_r is the space of polynomials of degree r or less on \mathbb{R} , be the Lagrange (or continuous) finite element space of r th-order associated with a mesh $\{[x_p, x_{p+1}]\}_{p \in \mathbb{Z}}$ of \mathbb{R} . We have $V_h^r(\mathbb{R}) \subset H^1(\mathbb{R})$ [24]. The semi-discretized formulation in space of the problem can be written as follows:

$$\left\{ \begin{array}{l} \text{Find } u_h(., t) \in V_h^r(\mathbb{R}), t \in]0, T[\text{ such that:} \\ \frac{d^2}{dt^2} \int_{\mathbb{R}} u_h v_h dx + c^2 \int_{\mathbb{R}} \frac{\partial u_h}{\partial x} \frac{\partial v_h}{\partial x} dx = 0, \quad \forall v_h \in V_h^r(\mathbb{R}), \\ u_h(x, 0) = u_{0,h}(x), \quad \frac{\partial u_h}{\partial t}(x, 0) = u_{1,h}(x) \quad \text{in } \mathbb{R}, \end{array} \right. \quad (11.3)$$

where $u_{0,h}$ and $u_{1,h}$ are suitable approximations of u_0 and u_1 .

For each segment $[x_p, x_{p+1}]$ of \mathbb{R} , we have a set of $r + 1$ interpolations points which are its two ends x_p and x_{p+1} and $r - 1$ regularly spaced interior points denoted $x_{p,j}$, such that $x_{p,j} = x_p + j(x_{p+1} - x_p)/r$, $j = 1..r - 1$. The degrees of freedom of the FEM are the values of the functions of $V_h^r(\mathbb{R})$ at these interpolation points.

Let $(\lambda_q)_{q \in \mathbb{Z}}$ be a basis of $V_h^r(\mathbb{R})$. The basis functions are of two kinds:

- If λ_q corresponds to a point x_p , its support is $[x_{p-1}, x_{p+1}]$ and its restrictions to $[x_{p-1}, x_p]$ and $[x_p, x_{p+1}]$ are polynomials of P_r . Moreover, we have $\lambda_q(x_p) = 1$ and $\lambda_q = 0$ for all the other interpolation points of $[x_{p-1}, x_{p+1}]$. In this case we set $\lambda_q = \lambda_p$.
- If λ_q corresponds to an interior point $x_{p,j}$, it is a polynomial of P_r on its support $[x_p, x_{p+1}]$ such that $\lambda_q(x_{p,j}) = 1$ and $\lambda_q = 0$ for all the other interpolation points of $[x_p, x_{p+1}]$. In this case we set $\lambda_q = \lambda_{p,j}$.

With the above notations, we can write:

$$u_h = \sum_{p \in \mathbb{Z}} u_p \lambda_p + \sum_{p \in \mathbb{Z}} \sum_{\ell=1}^{r-1} u_{p,\ell} \lambda_{p,\ell}. \quad (11.4)$$

From (11.4), we get:

$$\begin{aligned} \int_{\mathbb{R}} u_h \lambda_p dx &= \sum_{\ell=1}^{r-1} \left(u_{p-1,\ell} \int_{x_{p-1}}^{x_p} \lambda_{p-1,\ell} \lambda_p dx + u_{p,\ell} \int_{x_p}^{x_{p+1}} \lambda_{p,\ell} \lambda_p dx \right) \\ &\quad + u_{p-1} \int_{x_{p-1}}^{x_p} \lambda_{p-1} \lambda_p dx + u_{p+1} \int_{x_p}^{x_{p+1}} \lambda_{p+1} \lambda_p dx \\ &\quad + u_p \int_{x_{p-1}}^{x_{p+1}} \lambda_p^2 dx, \end{aligned} \quad (11.5a)$$

$$\begin{aligned} \int_{\mathbb{R}} u_h \lambda_{p,j} dx &= \sum_{\ell=1}^{r-1} u_{p,\ell} \int_{x_p}^{x_{p+1}} \lambda_{p,\ell} \lambda_{p,j} dx \\ &\quad + u_p \int_{x_p}^{x_{p+1}} \lambda_p \lambda_{p,j} dx + u_{p+1} \int_{x_p}^{x_{p+1}} \lambda_{p+1} \lambda_{p,j} dx. \end{aligned} \quad (11.5b)$$

Then (11.3) is equivalent to the following (infinite¹) system of ordinary differential equations (the λ_ℓ and λ_m are taken with their initial meaning):

$$M_{1,r} \frac{d^2 \mathbf{U}}{dt^2}(t) + K_{1,r} \mathbf{U}(t) = 0,$$

$$\text{with } \left\{ \begin{array}{l} (M_{1,r})_{\ell,m} = \int_{\mathbb{R}} \lambda_\ell(x) \lambda_m(x) dx, \\ (K_{1,r})_{\ell,m} = \int_{\mathbb{R}} \frac{\partial \lambda_\ell}{\partial x} \frac{\partial \lambda_m}{\partial x} dx, (\ell, m) \in \mathbb{Z}^2, \\ \mathbf{U}^T = (u_q, u_{q,1}, \dots, u_{q,r-1})_{q \in \mathbb{Z}}. \end{array} \right. \quad (11.6)$$

$M_{1,r}$ is the *mass matrix* and $K_{1,r}$ is the *stiffness matrix* of the discrete problem.

If one chooses the canonical basis of $V_h^r(\mathbb{R})$, $M_{1,r}$ is a symmetric positive $(2r+1)$ -diagonal matrix. Since the inverse of this matrix is full and then cannot be stored, after discretization in time, this mass matrix must be inverted at each time-step. Of course, this matrix is even larger in the 2D and 3D cases.

All these considerations explain the need of eliminating the mass matrix in order to hope to compete with finite difference methods. A good approach is to use a *mass-lumping* technique which consists in computing an approximated value of the mass integrals by using a numerical integration formula (or quadrature rule). However, such a formula must satisfy the following two properties:

1. It must keep the order of approximation of the method.
2. Its points must coincide with the degrees of freedom – which correspond to the Lagrange interpolation points – of the finite element method on each interval $[x_p, x_{p+1}]$.

¹ Which is finite in a bounded domain with boundary conditions.

The first condition is natural. The second condition is derived from the following:

If we replace the integrals in (11.5a) and (11.5b) by a quadrature rule whose weights are ω_m^p and whose points ξ_m^p , $m = 1..n$ on $[x_p, x_{p+1}]$, we obtain²:

$$\begin{aligned} \int_{\mathbb{R}} u_h \lambda_p dx &\simeq \sum_{\ell=1}^{r-1} u_{p-1,\ell} \left(\sum_{m=1}^n \omega_m^{p-1} \lambda_{p-1,\ell}(\xi_m^{p-1}) \lambda_p(\xi_m^{p-1}) \right) \\ &+ \sum_{\ell=1}^{r-1} u_{p,\ell} \left(\sum_{m=1}^n \omega_m^p \lambda_{p,\ell}(\xi_m^p) \lambda_p(\xi_m^p) \right) \\ &+ u_{p-1} \sum_{m=1}^n \omega_m^{p-1} \lambda_{p-1}(\xi_m^{p-1}) \lambda_p(\xi_m^{p-1}) \\ &+ u_{p+1} \sum_{m=1}^n \omega_m^p \lambda_{p+1}(\xi_m^p) \lambda_p(\xi_m^p) \end{aligned} \quad (11.7a)$$

$$\begin{aligned} \int_{\mathbb{R}} u_h \lambda_{p,j} dx &\simeq \sum_{\ell=1}^{r-1} u_{p,\ell} \sum_{m=1}^n \omega_m^p \lambda_{p,\ell}(\xi_m^p) \lambda_{p,j}(\xi_m^p) \\ &+ u_p \sum_{m=1}^n \omega_m^p \lambda_p(\xi_m^p) \lambda_{p,j}(\xi_m^p) \\ &+ u_{p+1} \sum_{m=1}^n \omega_m^p \lambda_{p+1}(\xi_m^p) \lambda_{p,j}(\xi_m^p). \end{aligned} \quad (11.7b)$$

Now, if $n = r + 1$, $\xi_{m+1}^p = x_{p,m}$ $\forall m = 1..r - 1$ and $\xi_1^p = x_p$ and $\xi_{r+1}^p = x_{p+1}$, all the approximate integrals defined in (11.7a) and (11.7b) are equal to 0 except those of λ_p^2 and $\lambda_{p,j}^2$ (since all the basis functions are equal to 0 at all the interpolation points except for that to which they correspond).

So, (11.7a) and (11.7b) simply become:

² As we shall see later, these points and weights are derived from the quadrature points and weight on the interval $[0, 1]$ by an affine mapping depending on $[x_p, x_{p+1}]$.

$$\int_{\mathbb{R}} u_h \lambda_p dx \simeq (\omega_{r+1}^{p-1} + \omega_1^p) u_p, \quad (11.8a)$$

$$\int_{\mathbb{R}} u_h \lambda_{p,j} dx \simeq \omega_{j+1}^p u_{p,j}, \quad (11.8b)$$

which actually provides a diagonal mass matrix.

The first condition can be derived from [24, 25], in which it is said that, in order to keep, for the mass-lumped scheme, the same accuracy as that of the classical scheme, one must use, for P_r finite elements, a quadrature rule exact for P_{2r-1} .

For $r = 1$ and $r = 2$, both conditions are realized by the trapezoidal and the Simpson quadrature rules respectively since the trapezoidal rule is exact for P_1 and has its points at the ends of the interval of integration and the Simpson rule is exact for P_3 and its quadrature points are the two ends and the midpoint of the interval which are the interpolation points for P_2 .

For $r > 2$, for which one uses classically regularly spaced degrees of freedom, a Newton-Cotes quadrature rule [45] would not satisfy the first condition since such a rule with r quadrature points will be exact for P_r only³. A Gauss quadrature rule could satisfy the first condition but never the second one since its set of quadrature points never contains the ends of the interval.

Another candidate is the Gauss-Lobatto rule. This rule has the feature of including the ends of the interval of integration in its quadrature points and to be exact for P_{2r-1} with $r + 1$ quadrature points ($r - 1$ interior points + the ends of the interval).

So, a Gauss-Lobatto quadrature rule could be a good candidate, since it is exact for P_{2r-1} . On the other hand, it has the same number of points as the interpolation points for a P_r FEM on an interval $[x_p, x_{p+1}]$ and the ends of the interval are quadrature points. In order to achieve the second condition we simply have to shift the interior interpolation points to the Gauss-Lobatto quadrature points. So that, for $r > 2$, the degrees of freedom must be modified in order to coincide with the quadrature points as shown in Fig. 11.1.

Actually, the quadrature points of the r th-order Gauss-Lobatto rule on the interval $]0, 1[$ are the zeroes of the derivative of $L_r(2x - 1)$, where L_r is the r th Legendre polynomial [45, 96] of which we recall the definition:

³ Another drawback of these rules lies in the fact that they could contain negative weights for $r \geq 7$ which leads, as we shall see later, to unstable schemes.

$$\left\{ \begin{array}{l} L_0(x) = 1, \\ L_1(x) = x, \\ L_r(x) = \frac{(2r-1)}{r}xL_{r-1}(x) - \frac{(r-1)}{r}L_{r-2}(x) \quad \forall r > 1. \end{array} \right. \quad (11.9)$$

Of course, one must add the points 0 and 1 to these zeroes.

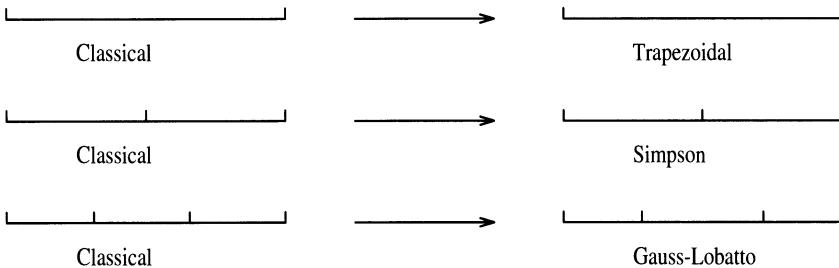


Fig. 11.1. Q_1 : trapezoidal quadrature rule: $\hat{x}_1 = \hat{\xi}_1 = 0$, $\hat{x}_2 = \hat{\xi}_2 = 1$ (top). Q_2 : Simpson quadrature rule: $\hat{x}_1 = \hat{\xi}_1 = 0$, $\hat{x}_2 = \hat{\xi}_2 = 1/2$, $\hat{x}_3 = \hat{\xi}_3 = 1$ (middle). Q_3 : Gauss-Lobatto quadrature rule: $\hat{x}_1 = \hat{\xi}_1 = 0$, $\hat{x}_2 = 1/3$, $\hat{\xi}_2 = (5 - \sqrt{5})/10$, $\hat{x}_3 = 2/3$, $\hat{\xi}_3 = (5 + \sqrt{5})/10$, $\hat{x}_4 = \hat{\xi}_4 = 1$ (bottom), where \hat{x}_j are the degrees of freedom for classical FEM and $\hat{\xi}_j$ are the Gauss-Lobatto points on the interval $[0, 1]$

Equation (11.9) shows that, in particular, the trapezoidal and the Simpson rules are the Gauss-Lobatto rules of first and second orders⁴.

In Table 11.1, we give the exact values of the abscissae of the Gauss-Lobatto points $\hat{\xi}_j$ and the corresponding weights $\hat{\omega}_j$, $j = 1..r + 1$ on the interval $[0, 1]$. These values will be given up to $r = 6$ because the exact values for higher orders are too complicated. Since the points and the weights are symmetric versus the middle of the interval (i.e. $\hat{\xi}_{r+2-j} = 1 - \hat{\xi}_j$ and $\hat{\omega}_{r+2-j} = \hat{\omega}_j$), we only give the values on $[0, 1/2]$.

The approximated values for the points and the weights on the semi-interval $[0, 1/2]$ are given below for $r = 1$ to 10.

- $r = 1$
- Point: $\hat{\xi}_1 = 0$.
- Weight: $\hat{\omega}_1 = 0.5$.

⁴ This explains why the Simpson rule, which should be exact for P_2 as a Newton-Cotes quadrature rule, is actually exact for P_3 .

– $r = 2$

Points: $\hat{\xi}_1 = 0, \hat{\xi}_2 = 0.5$.

Weights: $\hat{\omega}_1 = 0.166\ 666\ 667, \hat{\omega}_2 = 0.666\ 666\ 667$.

– $r = 3$

Points: $\hat{\xi}_1 = 0, \hat{\xi}_2 = 0.276\ 393\ 202$.

Weights: $\hat{\omega}_1 = 0.0833\ 333\ 333, \hat{\omega}_2 = 0.416\ 666\ 667$.

– $r = 4$

Points: $\hat{\xi}_1 = 0, \hat{\xi}_2 = 0.172\ 673\ 165, \hat{\xi}_3 = 0.5$.

Weights: $\hat{\omega}_1 = 0.05, \hat{\omega}_2 = 0.272\ 222\ 222, \hat{\omega}_3 = 0.355\ 555\ 556$.

– $r = 5$

Points: $\hat{\xi}_1 = 0, \hat{\xi}_2 = 0.117\ 472\ 338, \hat{\xi}_3 = 0.357\ 384\ 242$.

Weights: $\hat{\omega}_1 = 0.033\ 333\ 333, \hat{\omega}_2 = 0.189\ 237\ 478, \hat{\omega}_3 = 0.277\ 429\ 189$.

– $r = 6$

Points: $\hat{\xi}_1 = 0, \hat{\xi}_2 = 0.084\ 888\ 052, \hat{\xi}_3 = 0.265\ 575\ 603, \hat{\xi}_4 = 0.5$.

Weights: $\hat{\omega}_1 = 0.023\ 809\ 524, \hat{\omega}_2 = 0.138\ 413\ 024, \hat{\omega}_3 = 0.215\ 872\ 691, \hat{\omega}_4 = 0.243\ 809\ 524$.

– $r = 7$

Points: $\hat{\xi}_1 = 0, \hat{\xi}_2 = 0.064\ 129\ 926, \hat{\xi}_3 = 0.204\ 149\ 909, \hat{\xi}_4 = 0.395\ 350\ 391$.

Weights: $\hat{\omega}_1 = 0.017\ 857\ 143, \hat{\omega}_2 = 0.105\ 352\ 114, \hat{\omega}_3 = 0.170\ 561\ 346, \hat{\omega}_4 = 0.206\ 229\ 397$.

– $r = 8$

Points: $\hat{\xi}_1 = 0, \hat{\xi}_2 = 0.055\ 012\ 100\ 2, \hat{\xi}_3 = 0.161\ 406\ 860, \hat{\xi}_4 = 0.318\ 441\ 268, \hat{\xi}_5 = 0.5$.

Weights: $\hat{\omega}_1 = 0.0138\ 888\ 889, \hat{\omega}_2 = 0.082\ 747\ 681, \hat{\omega}_3 = 0.137\ 269\ 356, \hat{\omega}_4 = 0.173\ 214\ 255, \hat{\omega}_5 = 0.185\ 759\ 637$.

– $r = 9$

Points: $\hat{\xi}_1 = 0, \hat{\xi}_2 = 0.040\ 233\ 046, \hat{\xi}_3 = 0.130\ 613\ 067, \hat{\xi}_4 = 0.261\ 037\ 525, \hat{\xi}_5 = 0.417\ 360\ 521$.

Weights: $\hat{\omega}_1 = 0.011\ 111\ 111, \hat{\omega}_2 = 0.066\ 652\ 995, \hat{\omega}_3 = 0.112\ 444\ 671, \hat{\omega}_4 = 0.146\ 021\ 342, \hat{\omega}_5 = 0.163\ 769\ 881$.

– $r = 10$

Points: $\hat{\xi}_1 = 0, \hat{\xi}_2 = 0.032\ 999\ 285, \hat{\xi}_3 = 0.107\ 758\ 263, \hat{\xi}_4 = 0.217\ 382\ 337, \hat{\xi}_5 = 0.352\ 120\ 932, \hat{\xi}_6 = 0.5$.

Weights: $\hat{\omega}_1 = 0.009\ 090\ 909, \hat{\omega}_2 = 0.054\ 806\ 137, \hat{\omega}_3 = 0.093\ 584\ 941, \hat{\omega}_4 = 0.124\ 024\ 052, \hat{\omega}_5 = 0.143\ 439\ 562, \hat{\omega}_6 = 150\ 108\ 798$.

Table 11.1. Exact values of the abscissae of the points and of the weights of the Gauss-Lobatto quadrature rules for the semi-interval $[0, 1/2]$

	$j = 1$	$j = 2$	$j = 3$	$j = 4$
$r = 1 (\hat{\xi})$	0			
$r = 1 (\hat{\omega})$	$\frac{1}{2}$			
$r = 2 (\hat{\xi})$	0	$\frac{1}{2}$		
$r = 2 (\hat{\omega})$	$\frac{1}{6}$	$\frac{2}{3}$		
$r = 3 (\hat{\xi})$	0	$\frac{5 - \sqrt{5}}{10}$		
$r = 3 (\hat{\omega})$	$\frac{1}{12}$	$\frac{5}{12}$		
$r = 4 (\hat{\xi})$	0	$\frac{7 - \sqrt{5}}{14}$	$\frac{1}{2}$	
$r = 4 (\hat{\omega})$	$\frac{1}{20}$	$\frac{16}{45}$	$\frac{49}{180}$	
$r = 5 (\hat{\xi})$	0	$\frac{21 - \sqrt{147 + 42\sqrt{7}}}{42}$	$\frac{21 - \sqrt{147 - 42\sqrt{7}}}{42}$	
$r = 5 (\hat{\omega})$	$\frac{1}{30}$	$\frac{14 - \sqrt{7}}{60}$	$\frac{14 + \sqrt{7}}{60}$	
$r = 6 (\hat{\xi})$	0	$\frac{33 - \sqrt{95 + 66\sqrt{15}}}{66}$	$\frac{33 - \sqrt{95 - 66\sqrt{15}}}{66}$	$\frac{1}{2}$
$r = 6 (\hat{\omega})$	$\frac{1}{42}$	$\frac{31}{175} - \frac{\sqrt{15}}{100}$	$\frac{31}{175} + \frac{\sqrt{15}}{100}$	$\frac{128}{525}$

The Gauss-Lobatto points on an interval $[x_p, x_{p+1}]$ are deduced from those on the unit interval $[0, 1]$ by the following affine mapping:

$$F_p(\hat{x}) = (x_{p+1} - x_p)\hat{x} + x_p \quad (11.10)$$

and the weights are multiplied by $(x_{p+1} - x_p)$.

Now, on the interval $[0, 1]$, the interpolation function $\hat{\varphi}_j$ of r th order is defined as

$$\hat{\varphi}_j(\hat{x}) = \frac{\prod_{\substack{p=1, p \neq j \\ r+1}}^{r+1} (\hat{x} - \hat{\xi}_p)}{\prod_{p=1, p \neq j}^{r+1} (\hat{\xi}_j - \hat{\xi}_p)}, \quad j = 1..r+1. \quad (11.11)$$

The interpolation functions φ_j on an interval $[x_p, x_{p+1}]$ are derived from (11.11) in the following way:

$$\varphi_{p,j} = \hat{\varphi}_j \circ F_p^{-1}, \quad \forall j = 1..r+1. \quad (11.12)$$

With this notation, the basis functions of $V_h^r(\mathbb{R})$ can be written as

$$\lambda_p = \varphi_{p-1,r+1} \chi_p + \varphi_{p,1} \chi_{p+1}, \quad (11.13a)$$

$$\lambda_{p,j} = \varphi_{p,j+1} \chi_p \quad \forall j = 1..r-1, \quad (11.13b)$$

where χ_p is the characteristic function of $[x_p, x_{p+1}]$.

Remarks

1. One can easily check that this technique of mass-lumping holds for the inhomogeneous wave equation defined in (1.1).
2. The stiffness matrix $K_{1,r}$ can also be computed by using the Gauss-Lobatto rule. In the 1D case, such a computation provides the exact matrix but this will not be the case for higher dimensions in space.
3. The mass-lumping by the Gauss-Lobatto quadrature rule for $r = 1$ on a regular mesh provides exactly the second-order FDM approximation.
4. Another approach to spectral elements, based on the zeroes of Chebyshev polynomials as quadrature points, does not lead to mass-lumping [103].

11.1.2 Approximation in Time

A natural idea would be to approximate the time derivative by a finite element method. Unfortunately, even when mass-lumped, FEM techniques are not convenient for the approximation in time, since they provide implicit

schemes which are not suitable for the wave equations. So, in this case also, we shall use finite difference approximations of the derivative in time.

It is easy to see that the leaprog and symmetric schemes can be applied to this approach in the same way as for the finite difference approximation. This is not the case for the modified equation approach which cannot be applied in the same form as for finite difference methods for the following reason:

In our case, the modified equation approach corresponding to finite difference schemes would be, in dimension d :

$$M_{1,r} \frac{\mathbf{U}^{n+1} - 2\mathbf{U}^n + \mathbf{U}^{n-1}}{\Delta t^2} - \frac{c^4 \Delta t^2}{12} (\Delta^2)_h \mathbf{U}^n - c^2 K_{d,r} \mathbf{U}^n = 0, \quad (11.14)$$

where $K_{d,r}$ is obtained by applying a C^0 finite element method to Δ and $(\Delta^2)_h$ would be a finite element approximation of Δ^2 .

Unfortunately, a suitable approximation of Δ^2 by a finite element method would use an H^2 approximation, i.e. C^1 finite elements based on Hermite interpolation. Since we cannot mix C^0 and C^1 elements in the same variational formulation and, on the other hand, the use of C^1 elements is very complicated, besides the fact that mass-lumping is not obvious for such elements, this formulation is not worthwhile in our case.

An alternative to this formulation can be obtained by applying the modified equation approach to the discrete equation:

$$\frac{\mathbf{U}^{n+1} - 2\mathbf{U}^n + \mathbf{U}^{n-1}}{\Delta t^2} + c^2 N_{d,r} \mathbf{U}^n = 0, \quad (11.15)$$

where

$$N_{d,r} = M_{d,r}^{-1} K_{d,r}. \quad (11.16)$$

By using the technique described in (4.36)–(4.39), we obtain:

$$\frac{\mathbf{U}^{n+1} - 2\mathbf{U}^n + \mathbf{U}^{n-1}}{\Delta t^2} - \frac{c^4 \Delta t^2}{12} (N_{d,r})^2 \mathbf{U}^n + c^2 N_{d,r} \mathbf{U}^n = 0, \quad (11.17)$$

in which we first approximate Δ by using C^0 elements, and then we take the square of the approximated operator obtained.

Such an approach, which leads to very dispersive schemes in the case of finite difference approximations, provides satisfying results in the case of FEM. However, the factorized form of (11.17):

$$\frac{\mathbf{U}^{n+1} - 2\mathbf{U}^n + \mathbf{U}^{n-1}}{\Delta t^2} - c^2 N_{d,r} \left(\frac{c^2 \Delta t^2}{12} N_{d,r} \mathbf{U}^n - \mathbf{U}^n \right) = 0 \quad (11.18)$$

shows that this approach requires two computations of $N_{d,r}$ applied to a vector, which multiplies the computation time by two. This drawback is balanced by the fact that the stability condition is actually almost twice that of the leapfrog scheme. However this is true for the fourth-order scheme in time defined in (11.17) but not for higher-order approximations obtained as in (4.75), which can be written as follows:

$$\begin{aligned} \frac{\mathbf{U}^{n+1} - 2\mathbf{U}^n + \mathbf{U}^{n-1}}{\Delta t^2} &+ c^2 N_{d,r} \mathbf{U}^n \\ &- 2 \sum_{j=2}^m (-1)^j \frac{c^{2j} \Delta t^{2j-2}}{(2j)!} (N_{d,r})^j \mathbf{U}^n = 0, \end{aligned} \quad (11.19)$$

for which, as we shall see in a next section, the stability condition is not large enough to balance the number of computations of $N_{d,r}$.

11.2 Dispersion Relations

In the following sections, we shall suppose that all the points x_p are regularly spaced and we shall set:

$$h = x_{p+1} - x_p, \quad \forall p \in \mathbb{Z}. \quad (11.20)$$

11.2.1 P_2 Finite Elements

Even on a regular mesh, the finite element approximation has the following feature which does not appear in finite difference methods: to each kind of degree of freedom corresponds a different discrete equation. For instance, in P_2 , we obtain two discrete equations corresponding to the ends x_p and the midpoints $x_{p,1}$ of the intervals:

$$\frac{d^2 u_p}{dt^2}(t) = -\frac{c^2}{h^2} (14u_p(t) - 8(u_{p,1}(t) + u_{p-1,1}(t)) + u_{p+1}(t) + u_{p-1}(t)), \quad \forall p \in \mathbb{Z}, \quad (11.21a)$$

$$\frac{d^2 u_{p,1}}{dt^2}(t) = \frac{4c^2}{h^2} (u_p(t) - 2u_{p,1}(t) + u_{p+1}(t)), \quad \forall p \in \mathbb{Z}. \quad (11.21b)$$

Since we have here two classes of degrees of freedom invariable by translation, as indicated in Fig. 11.2, the plane wave solution is defined as follows:

$$u_p = \alpha_1 e^{i(pkh - \omega_h t)}, \quad (11.22a)$$

$$u_{p,1} = \alpha_2 e^{i((p+\frac{1}{2})kh - \omega_h t)}, \quad (11.22b)$$

where $(\alpha_1, \alpha_2)^T \in \mathbb{C}^2$.

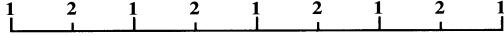


Fig. 11.2. The two classes of degrees of freedom for P_2

By inserting (11.22a) and (11.22b) into (11.21a) and (11.21b) we obtain:

$$\omega_h^2 \alpha_1 = \frac{c^2}{h^2} (14\alpha_1 - 8\alpha_2(e^{ikh/2} + e^{-ikh/2}) + \alpha_1(e^{ikh} + e^{-ikh})), \quad (11.23a)$$

$$\omega_h^2 \alpha_2 = -\frac{4c^2}{h^2} (\alpha_1 e^{-ikh/2} - 2\alpha_2 + \alpha_1 e^{ikh/2}), \quad (11.23b)$$

which can be written, after simplification, as the following generalized eigenvalue problem:

$$\widehat{N}_{1,2} \mathbf{U}_\alpha = \omega_h^2 \mathbf{U}_\alpha, \quad (11.24)$$

where

$$\mathbf{U}_\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}, \quad \widehat{N}_{1,2} = \widehat{M}_{1,2}^{-1} \widehat{K}_{1,2},$$

$$\widehat{M}_{1,2} = \frac{1}{3} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \quad \widehat{K}_{1,2}(k) = \frac{2c^2}{3h^2} \begin{pmatrix} 7 + \cos kh & -8 \cos \frac{kh}{2} \\ -8 \cos \frac{kh}{2} & 8 \end{pmatrix}.$$

$\widehat{M}_{1,2}$ and $\widehat{K}_{1,2}$ are the matricial symbols of the mass matrix $M_{1,2}$ and the opposite of the stiffness matrix $-K_{1,2}$. These symbols satisfy the same properties as the scalar symbols obtained for finite difference methods.

After setting $w = \sin kh/2$, the characteristic equation of (11.24) can be written as

$$h^4 \omega_h^4 + 4c^2 h^2 (w^2 - 6) \omega_h^2 + 96c^4 w^2 = 0. \quad (11.25)$$

The two zeroes of this quartic equation

$$\omega_{h,1}^2 = \frac{c^2}{h^2} (12 - 2w^2 - 2\sqrt{36 - 36w^2 + w^4}), \quad (11.26a)$$

$$\omega_{h,2}^2 = \frac{c^2}{h^2} (12 - 2w^2 + 2\sqrt{36 - 36w^2 + w^4}), \quad (11.26b)$$

are the two dispersion relations of the problem.

Now, if we compute the Taylor expansions of the two relations, we obtain:

$$\omega_{h,1}^2 = c^2 k^2 \left(1 - \frac{k^4 h^4}{1440} - \frac{k^6 h^6}{48384} + O(k^8 h^8) \right), \quad (11.27a)$$

$$\omega_{h,2}^2 = c^2 k^2 \left(\frac{24}{k^2 h^2} - 2 + \frac{k^2 h^2}{12} - \frac{k^4 h^4}{480} + O(k^6 h^6) \right). \quad (11.27b)$$

The first relation is a fourth-order approximation of the dispersion relation of the continuous wave equation and the second one reveals the presence of a parasitic wave whose velocity tends to infinity when h tends to zero. As we shall see later, the amplitude of this wave is in $O(h^4)$.

11.2.2 P_3 and Higher-Order Finite Elements

For P_3 finite elements, we have the three following equations which correspond to three degrees of freedom located at the points x_p , $x_{p,1} = x_p + h(5 - \sqrt{5})/10$ and $x_{p,2} = x_p + h(5 + \sqrt{5})/10$:

$$\begin{aligned} \frac{d^2 u_p}{dt^2}(t) &+ \frac{c^2}{h^2} [52u_p(t) - \frac{5}{2}(5 + 3\sqrt{5})(u_{p-1,2}(t) + u_{p,1}(t)) \\ &+ \frac{5}{2}(3\sqrt{5} - 5)(u_{p-1,1}(t) + u_{p,2}(t)) \\ &- (u_{p-1}(t) + u_{p+1}(t))] = 0, \quad \forall p \in \mathbb{Z}, \end{aligned} \quad (11.28a)$$

$$\begin{aligned} \frac{d^2 u_{p,1}}{dt^2}(t) &+ \frac{c^2}{h^2} [-(5 + 3\sqrt{5})u_p(t) + 20u_{p,1}(t) - 10u_{p,2}(t) \\ &+ (3\sqrt{5} - 5)u_{p+1}(t)] = 0, \quad \forall p \in \mathbb{Z}, \end{aligned} \quad (11.28b)$$

$$\begin{aligned} \frac{d^2 u_{p,2}}{dt^2}(t) &+ \frac{c^2}{h^2} [(3\sqrt{5} - 5)u_p(t) - 10u_{p,1}(t) + 20u_{p,2}(t) \\ &- (5 + 3\sqrt{5})u_{p+1}(t)] = 0, \quad \forall p \in \mathbb{Z}. \end{aligned} \quad (11.28c)$$

Here, as shown if Fig. 11.3, the plane wave solution is subdivided into three classes represented by

$$u_p = \alpha_1 e^{i(pkh - \omega_h t)}, \quad (11.29a)$$

$$u_{p,1} = \alpha_2 e^{i((p+\mu)kh - \omega_h t)}, \quad (11.29b)$$

$$u_{p,2} = \alpha_3 e^{i((p+\nu)kh - \omega_h t)}, \quad (11.29c)$$

where $\mu = (5 - \sqrt{5})/10$, $\nu = (5 + \sqrt{5})/10$ and $(\alpha_1, \alpha_2, \alpha_3)^T \in \mathbb{C}^3$.



Fig. 11.3. The three classes of degrees of freedom for P_3

By inserting (11.29a)–(11.29c) into (11.28a)–(11.28c), we obtain, after simplification:

$$\begin{aligned}
\alpha_1 \omega_h^2 + & \frac{c^2}{h^2} [52\alpha_1 - \frac{5}{2}(5+3\sqrt{5})(\alpha_3 e^{-ikh} + \alpha_2 e^{ikh}) \\
& + \frac{5}{2}(3\sqrt{5}-5)(\alpha_2 e^{-ikh} + \alpha_3 e^{ikh}) \\
& - \alpha_1(e^{-ikh} + e^{ikh})] = 0,
\end{aligned} \tag{11.30a}$$

$$\begin{aligned}
\alpha_2 \omega_h^2 + & \frac{c^2}{h^2} [-(5+3\sqrt{5})\alpha_1 e^{-ikh} + 20\alpha_2 - 10\alpha_3 e^{ikh} \\
& + (3\sqrt{5}-5)\alpha_1 e^{ikh}] = 0,
\end{aligned} \tag{11.30b}$$

$$\begin{aligned}
\alpha_3 \omega_h^2 + & \frac{c^2}{h^2} [(3\sqrt{5}-5)\alpha_1 e^{-ikh} - 10\alpha_2 e^{-ikh} + 20\alpha_3 \\
& - (5+3\sqrt{5})\alpha_1 e^{ikh}] = 0,
\end{aligned} \tag{11.30c}$$

which provides the following three-dimensional eigenvalue problem:

$$\widehat{N}_{1,3} \mathbf{U}_\alpha = \omega_h^2 \mathbf{U}_\alpha, \tag{11.31}$$

where

$$\mathbf{U}_\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}, \quad \widehat{N}_{1,3}(k) = \widehat{M}_{1,3}^{-1} \widehat{K}_{1,3},$$

with:

$$\widehat{M}_{1,3} = \frac{1}{12} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 5 \end{bmatrix},$$

$$\begin{aligned}
\frac{h^2}{12c^2} \widehat{K}_{1,3} = & \\
& \left[\begin{array}{ccc} 26 - \cos kh & \frac{-5ae^{ikh\mu} + 5be^{-ikh\nu}}{4} & \frac{-5ae^{-ikh\mu} + 5be^{ikh\nu}}{4} \\
\frac{-5ae^{-ikh\mu} + 5be^{ikh\nu}}{4} & 25 & \frac{-25e^{i\frac{\sqrt{5}}{5}kh}}{2} \\
\frac{-5ae^{ikh\mu} + 5be^{-ikh\nu}}{4} & \frac{-25e^{-i\frac{\sqrt{5}}{5}kh}}{2} & 25 \end{array} \right],
\end{aligned}$$

where $a = 5 + 3\sqrt{5}$ and $b = -5 + 3\sqrt{5}$. Here also, $\widehat{M}_{1,3}$ and $\widehat{K}_{1,3}$ are the symbols of the mass matrix and the stiffness matrix.

By setting $w = \cos kh$, the characteristic polynomial of (11.31) can be written as

$$h^6\omega_h^6 + 2c^2h^4(w-46)\omega_h^4 + 120c^4h^2(w+14)\omega_h^2 + 3600c^6(w-1) = 0. \quad (11.32)$$

After a computation aided by Maple, we obtain the following three roots of this equation:

$$\omega_{h,1}^2 = \frac{c^2}{6h^2}(4a_1^{1/3} - 9a_2 - 4w + 184), \quad (11.33a)$$

$$\omega_{h,2}^2 = \frac{c^2}{12h^2}(-4a_1^{1/3} + 9a_2 - 8w + 368 + i\sqrt{3}(4a_1^{1/3} + 9a_2)), \quad (11.33b)$$

$$\omega_{h,3}^2 = \frac{c^2}{12h^2}(-4a_1^{1/3} + 9a_2 - 8w + 368 - i\sqrt{3}(4a_1^{1/3} + 9a_2)), \quad (11.33c)$$

where

$$\begin{aligned} a_1 &= -w^3 + 273w^2 - 16743w + 16471 \\ &\quad + 45\sqrt{3(w^4 - 364w^3 + 33261w^2 - 24934w - 58589)}, \\ a_2 &= \frac{-4w^2 + 728w - 3424}{9a_1^{1/3}}. \end{aligned}$$

Actually, these roots are all real and their Taylor expansions are

$$\omega_{h,1}^2 = c^2k^2\left(\frac{60}{k^2h^2} + 5 - \frac{7}{6}k^2h^2 + \frac{23}{60}k^4h^4 - \frac{1899}{11200}k^6h^6 + O(h^8)\right), \quad (11.34a)$$

$$\omega_{h,2}^2 = c^2k^2\left(1 - \frac{1}{302400}k^6h^6 - \frac{1}{427680000}k^{10}h^{10} + O(h^{12})\right), \quad (11.34b)$$

$$\begin{aligned} \omega_{h,3}^2 &= c^2k^2\left(\frac{30}{k^2h^2} - 5 + \frac{13}{12}k^2h^2 - \frac{137}{360}k^4h^4\right. \\ &\quad \left.+ \frac{51259}{302400}k^6h^6 + O(h^8)\right). \end{aligned} \quad (11.34c)$$

Here also, we obtain one physical dispersion relations and two other corresponding to parasitic waves whose amplitudes are in $O(h^6)$ or $O(h^5)$ [115].

Of course, such an approach, which provides the explicit form of the eigenvalues, is difficult to apply to higher-order methods since the degree of the

characteristic polynomial increases with the order of the method. However, the characteristic equation (or its Taylor expansion) can be obtained for rather high orders. In that case, the Taylor expansions of the eigenvalues can be computed by a recursive algorithm. As an example, we compute, in the following, the Taylor expansions of the eigenvalues (11.33a)–(11.33c) by this algorithm.

We first look for a sixth-order Taylor expansion around $h = 0$ of the characteristic polynomial (11.32), in which we set $\lambda = \omega_h^2$. We obtain:

$$\begin{aligned} & \frac{1800(c^6k^2 - c^4\lambda)}{h^4} + \frac{-150c^6k^4 + 90c^2\lambda^2 + 60c^4k^2\lambda}{h^2} \\ & -\lambda^3 + c^2k^2\lambda^2 - 5c^4k^4\lambda + 5c^6k^6 \\ & + \left(-\frac{1}{12}c^2k^4\lambda^2 + \frac{1}{6}c^4k^6\lambda - \frac{5}{56}c^6k^8\right)h^2 \\ & + \left(\frac{1}{360}c^2k^6\lambda^2 - \frac{1}{336}c^4k^8\lambda + \frac{1}{1008}c^6k^{10}\right)h^4 \\ & + \left(-\frac{1}{20160}c^2k^8\lambda^2 + \frac{1}{30240}c^4k^{10}\lambda - \frac{1}{133056}c^6k^{12}\right)h^6 = 0. \end{aligned} \quad (11.35)$$

Then, we look for a solution of this Taylor expansion of the form

$$\lambda = c^2 \frac{\gamma}{h^2}, \quad (11.36)$$

that we insert into (11.35).

The lowest-order term of the result is

$$\frac{c^6}{h^6}(-\gamma^3 + 90\gamma^2 - 1800\gamma), \quad (11.37)$$

which has three roots equal to 0, 30 and 60. Of course, the root equal to 0 corresponds to the physical solution and the other two roots to the parasitic waves. We take $\gamma = 30$. Then, we look for a solution of (11.35) of the form:

$$\lambda = c^2 \left(\frac{30}{h^2} + \gamma k^2 \right). \quad (11.38)$$

The lowest-order term of (11.35) becomes

$$\frac{c^6}{h^4}(900\gamma k^2 + 4500k^2), \quad (11.39)$$

whose unique root is $\gamma = -5$.

Once more we look for a solution of the form

$$\lambda = c^2 \left(\frac{30}{h^2} - 5k^2 + \gamma k^4 h^2 \right). \quad (11.40)$$

After inserting (11.40) into (11.35), we obtain its lowest-order term:

$$\frac{c^6}{h^2} (900k^4\gamma - 975k^4). \quad (11.41)$$

Its root is $\gamma = 13/12$.

So, let us take

$$\lambda = c^2 \left(\frac{30}{h^2} - 5k^2 + \frac{13}{12}k^4 h^2 + \gamma k^6 h^4 \right) \quad (11.42)$$

and insert it into (11.35). Then, its lowest-order term is

$$c^6 \left(900k^6\gamma - \frac{685}{2}k^6 \right), \quad (11.43)$$

whose unique root is $\gamma = -137/360$.

Finally, we look for an eigenvalue of the form

$$\lambda = c^2 \left(\frac{30}{h^2} - 5k^2 + \frac{13}{12}k^4 h^2 - \frac{137}{360}k^6 h^4 + \gamma k^8 h^6 \right). \quad (11.44)$$

The lowest-order term of (11.35), into which we inserted (11.44), is

$$c^6 h^2 \left(900k^8\gamma - \frac{51259}{336}k^8 \right) \quad (11.45)$$

and its unique root is $\gamma = 51259/302400$.

We can iterate the process as long as necessary, but we have reached the maximal accuracy for a sixth-order Taylor expansion of the characteristic polynomial. Higher accuracy requires adding to this expansion as many terms as the additional terms of the eigenvalue.

The rest of the Taylor expansion of (11.32) in which we insert the sixth-order expansion of λ is in $O(h^4)$, which means that we solved (11.32) up to the fourth-order.

Of course, we would obtain (11.34a) or (11.34b) by starting with $\gamma = 60$ or $\gamma = 0$.

If we did not know the form of the first term, we could start with any (even) negative power of h and we would find that the lowest term is proportional to γ , which would imply that $\gamma = 0$.

We could also take the unknown term of λ just proportional to h^{2r} without taking into account the term in k^{2r+2} . Then, we would obtain γ in terms of k^{2r+2} .

So, we have given a general algorithm which enables us to obtain the Taylor expansions of the dispersion relations as soon as we know how to compute a sufficiently accurate Taylor expansion of the characteristic polynomial of the eigenvalue problem derived from the plane wave solution of the problem.

We shall conclude this section with the three following remarks:

1. In a plane wave analysis of finite elements, the solution is not in the space of approximation. The schemes are simply regarded as finite difference schemes.
2. This kind of plane wave analysis which introduces different classes of the solution is a numerical version of the computation of Bloch waves in a crystal [74].
3. The Taylor expansions of the approximated velocities derived from (11.27a) and (11.34a) are of the same order as the dispersion relations, i.e. of fourth-order for P_2 and of sixth-order for P_3 . This reveals a superconvergence phenomenon since classical error estimates would provide third-order for P_2 and fourth-order for P_3 . It seems that we should obtain $2r$ -order accurate velocity for P_r .

11.3 Stability Analysis

11.3.1 The Leapfrog Scheme

For the following classical leapfrog approximation in time and a P_r approximation in space:

$$M_{1,r} \frac{\mathbf{U}^{n+1} - 2\mathbf{U}^n + \mathbf{U}^{n-1}}{\Delta t^2} - c^2 K_{1,r} \mathbf{U}^n = 0, \quad (11.46)$$

the dispersion relations are obviously the solutions of the eigenvalue problem:

$$\hat{N}_{1,r} \mathbf{U}_\alpha = \lambda \mathbf{U}_\alpha, \quad (11.47)$$

where

$$\lambda = \frac{4}{\Delta t^2} \sin^2 \frac{\omega_h \Delta t}{2}. \quad (11.48)$$

Hence, the dispersion relations are the same as that of the semi-discrete problem in space, in which we replace ω_h^2 by the λ defined in (11.48).

P_2 Approximation. For instance, after setting $w = \cos^2 kh/2$, we obtain, for a P_2 approximation in space, the two following dispersion relations:

$$\frac{4}{\Delta t^2} \sin^2 \frac{\omega_{h,1} \Delta t}{2} = \frac{c^2}{h^2} (10 + 2w + 2\sqrt{1 + 34w + w^2}), \quad (11.49a)$$

$$\frac{4}{\Delta t^2} \sin^2 \frac{\omega_{h,2} \Delta t}{2} = \frac{c^2}{h^2} (10 + 2w - 2\sqrt{1 + 34w + w^2}). \quad (11.49b)$$

As for the semi-discrete problem in space, $\omega_{h,1}$ and $\omega_{h,2}$ are the pulsations of a physical wave and a parasitic wave. Of course, *both* waves must not be exponentially increasing in order to keep the scheme stable. So, if we set, as for the finite difference methods, $\alpha = c\Delta t/h$, the stability condition is

$$\alpha \leq \alpha_M = \min_{1 \leq j \leq 2} \left[\frac{2}{\sup_{w \in [-1,1]} f_j(w)} \right], \quad (11.50)$$

with

$$f_1(w) = \sqrt{10 + 2w + 2\sqrt{1 + 34w + w^2}}, \quad (11.51a)$$

$$f_2(w) = \sqrt{10 + 2w - 2\sqrt{1 + 34w + w^2}}. \quad (11.51b)$$

Obviously, $\sup_{w \in [0,1]} f_1(w) = f_1(1) = 2\sqrt{6}$ and $\sup_{w \in [0,1]} f_2(w) = f_2(0) = 2\sqrt{2}$, so that

$$\alpha_M = \frac{\sqrt{6}}{6} \simeq 0.408\,248\,29. \quad (11.52)$$

P_3 and Higher-Order Approximations. In the same way, the stability condition for P_3 is given by

$$\alpha \leq \alpha_M = \max_{1 \leq j \leq 3} \left[\frac{2}{\sup_{w \in [-1,1]} f_j(w)} \right], \quad (11.53)$$

where $w = \cos kh$ and $f_1 = \alpha\omega_{h,1}$, $f_2 = \alpha\omega_{h,2}$, $f_3 = \alpha\omega_{h,3}$ with $\omega_{h,1}$, $\omega_{h,2}$, $\omega_{h,3}$ given by (11.33a)–(11.33c).

However, in this case, the maximum of these functions cannot be obtained explicitly because of their complexity. So, we are going to derive it from the characteristic polynomial of this approximation.

By setting $kh = 2\pi K$, $\lambda_j = h^2\omega_{h,j}^2/c^2$ and using the fact that each eigenvalue satisfies the characteristic polynomial, we can derive from (11.32) the following relations:

$$\begin{aligned} P(\lambda_j) &= \lambda_j^3 + 2(\cos 2\pi K - 46)\lambda_j^2 + 120(\cos 2\pi K + 14)\lambda_j \\ &\quad + 3600(\cos 2\pi K - 1) = 0, \quad \forall j = 1..3. \end{aligned} \quad (11.54)$$

The maximum of λ_j is reached at the value of K so that

$$\frac{d\lambda_j(K)}{dK} = 0. \quad (11.55)$$

So, by differentiating P versus K and taking (11.55) into account, we obtain:

$$\frac{dP(\lambda_j)}{dK} = \frac{dP(\lambda_j)}{d\lambda_j} \frac{d\lambda_j}{dK} = 4 \sin 2\pi K Q(\lambda_j) = 0, \quad (11.56)$$

where

$$Q(\lambda_j) = \lambda_j^2 + 60\lambda_j + 1800. \quad (11.57)$$

Since Q has no real root, this equation is satisfied for $\sin 2\pi K = 0$, which provides two classes of solutions $K = \ell$ or $K = 1/2 + \ell$, $\ell \in \mathbb{Z}$, for which λ_j can reach its maximum. The values of λ_j for these two classes can be obtained by inserting these solutions into P . So,

– For $K = \ell$, we have:

$$\lambda_j^3(0) - 90\lambda_j^2(0) + 1800\lambda_j(0) = 0, \quad (11.58)$$

whose solutions are 0, 30 and 60.

– For $K = 1/2 + \ell$, we obtain:

$$\lambda_j^3(1/2) - 94\lambda_j^2(1/2) + 1560\lambda_j(1/2) - 7200 = 0, \quad (11.59)$$

whose solutions are $6(7 - \sqrt{29})$, 10 and $6(7 + \sqrt{29})$.

Obviously, $6(7 + \sqrt{29})$ is the maximum of all these solutions and the stability condition is

$$\alpha \leq \alpha_M = \frac{2}{\sqrt{6(7 + \sqrt{29})}} = \frac{\sqrt{30}}{30} \sqrt{7 - \sqrt{29}} \simeq 0.232\,008\,27. \quad (11.60)$$

Such a process can also be applied, with more difficulty, to higher-order approximations.

In Table 11.2, we give the stability conditions for the leapfrog scheme used with a Q_r approximation, $r = 1$ to 5 in 1D. These CFL conditions must be divided by \sqrt{d} in dimension d and by $\sqrt{2}$ for the fourth-order symmetric scheme. These CFL conditions in 2D and 3D are those for the quadrilateral and hexahedral elements which will be constructed in the following chapters.

Table 11.2. The stability conditions for Q_r approximations in 1D with a leapfrog scheme, when $r = 1$ to 5

r	1	2	3	4	5
$\alpha_M \simeq$	1	0.4082	0.2320	0.1476	0.1010

Remarks:

1. The stability condition for P_3 seems to be about twice as restrictive as that of P_2 and much more restrictive than the stability condition obtained for FDM. In fact, this is not true because the $\alpha = c\Delta t/h$ used for FEM takes into account the length h of the element but, actually, in terms of points, a P_r mesh of N_e elements contains $rN_e + 1$ points, i.e. r points per element, and, therefore, the stability conditions versus the number of points is 0.816 496 58 for P_2 and 0.696 024 83 for P_3 .
2. Of course, one could obtain an approximate value, as accurate as we wish, of the maximum of $\omega_{h,j}$ by using its Taylor expansion which is always possible to obtain. However, this will never lead to the exact value, which is indispensable for a theorem.

11.3.2 The Modified Equation Approach

The study of the stability for the modified equation approach is based on the two following properties:

1. The symbol of the n th power D^n of a matricial operator D is equal to the n th power of the symbol of D .
2. If λ is the eigenvalue of a matrix M then, for any polynomial P , $P(\lambda)$ is the eigenvalue of $P(M)$.

By using the first property, one can see that the dispersion relations of the modified equation (11.18) are the eigenvalues of the following problem:

$$4 \sin^2 \frac{\omega_h \Delta t}{2} \mathbf{U}_\alpha = \left(\Delta t^2 \hat{N}_{1,r} - \frac{\Delta t^4}{12} \hat{N}_{1,r}^2 \right) \mathbf{U}_\alpha. \quad (11.61)$$

If $\lambda_j(k)$ is an eigenvalue of $\hat{N}_{1,r}$, we set $\lambda'_j = h^2 \lambda_j / c^2$ and, by using the second property, one can write the different dispersion relations of the problem in the following form:

$$4 \sin^2 \frac{\omega_h \Delta t}{2} = \alpha^2 \lambda'_j - \frac{\alpha^4}{12} \lambda'^2_j, \quad (11.62)$$

with $\alpha = c \Delta t / h$.

So, the stability condition is given by the double inequality

$$4 \geq \alpha^2 \lambda'_j - \frac{\alpha^4}{12} \lambda'^2_j \geq 0. \quad (11.63)$$

A simple computation shows that the left-hand inequality is always satisfied and the right-hand one holds for $\alpha \leq 2\sqrt{3/\lambda'_j}$. So, the stability condition is

$$\alpha \leq \max_{1 \leq j \leq r} \left[\frac{2\sqrt{3}}{\sup_k \sqrt{\lambda'_j}} \right]. \quad (11.64)$$

In particular, we obtain for P_2 :

$$\alpha_M = \frac{2\sqrt{3}}{\sqrt{24}} = \frac{\sqrt{2}}{2} \simeq 0.707\,106\,78 \quad (11.65)$$

and, for P_3 :

$$\alpha \leq \alpha_M = \frac{2\sqrt{3}}{\sqrt{6(7 + \sqrt{29})}} = \frac{\sqrt{10}}{10} \sqrt{7 - \sqrt{29}} \simeq 0.401\,850\,11. \quad (11.66)$$

As we said previously, the stability condition is almost twice as large as that of the leapfrog scheme, which balances the additional cost induced by the modified equation approach.

In the same way, the stability condition given by a sixth-order modified equation is given by

$$4 \geq \alpha^2 \lambda'_j - \frac{\alpha^4}{12} \lambda'^2_j + \frac{\alpha^6}{360} \lambda'^3_j \geq 0. \quad (11.67)$$

The right-hand equation is always true and the left-hand one is satisfied when $\alpha \leq \sqrt{2(5^{3/2} - 25^{1/3} + 5)/\lambda'_j}$. So, the stability condition can be written as

$$\alpha \leq \max_{1 \leq j \leq r} \left[\frac{\sqrt{2(5^{3/2} - 25^{1/3} + 5)}}{\sup_k \sqrt{\lambda'_j}} \right]. \quad (11.68)$$

In particular, we obtain for P_3 :

$$\alpha_M = \frac{\sqrt{2(5^{3/2} - 25^{1/3} + 5)}}{\sqrt{6(7 + \sqrt{29})}} \simeq 0.319\,209\,92. \quad (11.69)$$

For this approximation in time, one must compute three discrete operators instead of one for the leapfrog scheme and, obviously, the additional cost is not balanced by the stability condition which is less than 1.4 times greater than that of the leapfrog scheme. This is also the case of higher-order approximations in time obtained by using the modified equation.

11.3.3 Symmetric Schemes

We shall only examine the fourth-order symmetric scheme given in (4.48), since, as we saw previously, higher-order versions of such schemes have an excessively restrictive stability condition which makes them inefficient.

The FEM version of (4.48),

$$\begin{aligned} & \frac{1}{\Delta t^2} \left(-\frac{\theta+2}{6} \mathbf{U}^{n+2} + \frac{1+2\theta}{3} \mathbf{U}^{n+1} - \theta \mathbf{U}^n \right. \\ & \left. + \frac{1+2\theta}{3} \mathbf{U}^{n-1} - \frac{\theta+2}{6} \mathbf{U}^{n-2} \right) = \\ & c^2 N_{1,r} \left(\frac{5+2\theta}{12} \mathbf{U}^{n+1} + \frac{1-2\theta}{6} \mathbf{U}^n + \frac{5+2\theta}{12} \mathbf{U}^{n-1} \right), \end{aligned} \quad (11.70)$$

leads, for a plane wave solution, to the following eigenvalue problem:

$$\widehat{N}_{1,r} \mathbf{U}_\alpha = \lambda \mathbf{U}_\alpha, \quad (11.71)$$

where

$$\lambda = \frac{-8(2+\theta)a^2 + 12a}{\Delta t^2(3 - (2\theta + 5)a)}, \quad (11.72)$$

with $a = \sin^2(\omega \Delta t / 2)$.

So, as in Sect. 6.5.1, for each eigenvalue λ_j of (11.71), we must find the conditions such that the equation in a

$$-8(2+\theta)a^2 + 12a = (3 - (2\theta + 5)a)\lambda_j, \quad (11.73)$$

has all its roots between 0 and 1.

In Sect. 6.5.1, this problem has already been solved for $\theta = 0$ and led to $0 \leq \lambda_j \leq 2$. So, in our case, the stability condition can be written as

$$\alpha \leq \alpha_M = \max_{1 \leq j \leq r} \left[\frac{\sqrt{2}}{\sup_k \sqrt{\lambda'_j}} \right]. \quad (11.74)$$

For P_2 , we obtain:

$$\alpha \leq \alpha_M = \frac{\sqrt{2}}{\sqrt{24}} = \frac{\sqrt{3}}{6} \simeq 0.288\,675\,13 \quad (11.75)$$

and, for P_3 :

$$\alpha \leq \alpha_M = \frac{\sqrt{2}}{\sqrt{6(7 + \sqrt{29})}} = \frac{\sqrt{15(7 - \sqrt{29})}}{30} \simeq 0.164\,054\,62. \quad (11.76)$$

Here also, the stability conditions are $\sqrt{2}$ times more restrictive than that of the leapfrog scheme.

11.4 Dispersion Analysis

11.4.1 Taylor Expansions

The Taylor expansion of the numerical dispersion of the physical wave, defined in (7.4), reads, for the P_2 approximation in space:

$$q_h = \frac{\omega_{h,1}}{ck} = 1 - \frac{k^4 h^4}{2880} - \frac{k^6 h^6}{96\,768} + O(k^8 h^8) \quad (11.77)$$

and, for the P_3 approximation in space:

$$q_h = \frac{\omega_{h,2}}{ck} = 1 - \frac{k^6 h^6}{604\,800} - \frac{k^{10} h^{10}}{855\,360\,000} + O(k^{12} h^{12}). \quad (11.78)$$

Now, if we take into account the approximation in time, we obtain, for the leapfrog scheme:

– In P_2 :

$$\begin{aligned} q_h &= \frac{2}{ck\Delta t} \arcsin \left(\frac{c\Delta t}{2} \omega_{h,1} \right) \\ &= 1 + \frac{1}{24} c^2 k^2 \Delta t^2 - \frac{k^4 h^4}{2880} + \frac{3}{640} c^4 k^4 \Delta t^4 \\ &\quad - \frac{1}{23\,040} c^2 k^2 h^4 \Delta t^2 - \frac{k^6 h^6}{96\,768} + \dots \end{aligned} \quad (11.79)$$

– In P_3 :

$$q_h = \frac{2}{ck\Delta t} \arcsin \left(\frac{c\Delta t}{2} \omega_{h,2} \right)$$

$$\begin{aligned}
&= 1 + \frac{1}{24} c^2 k^2 \Delta t^2 + \frac{3}{640} c^4 k^4 \Delta t^4 - \frac{k^6 h^6}{604800} \\
&\quad + \frac{5}{7168} c^6 k^6 \Delta t^6 + \dots
\end{aligned} \tag{11.80}$$

One can see that we obtain exactly the same terms in Δt as for FDM.

The Taylor expansions for the symmetric schemes are derived from the physical eigenvalue in exactly the same way as in the FDM case and, here also, the terms in Δt are the same as those obtained for FDM. For instance, we obtain for P_2 :

$$q_h = 1 - \frac{k^4 h^4}{2880} - \frac{7}{1440} c^4 k^4 \Delta t^4 - \frac{k^6 h^6}{96768} - \frac{41}{12096} c^6 k^6 \Delta t^6 + \dots \tag{11.81}$$

The modified equation approach behaves differently from the case of FDM since its construction is not the same. In order to obtain the dispersion relation in this case, one must replace, in (11.79) and (11.80), $\omega_{h,j}$ by $\sqrt{\omega_{h,j}^2 - (\Delta t^2/12)\omega_{h,j}^4}$ for the fourth-order scheme in time. Then, we obtain:

– In P_2 :

$$q_h = 1 - \frac{k^4 h^4}{2880} - \frac{1}{720} c^4 k^4 \Delta t^4 - \frac{k^6 h^6}{96768} - \frac{5}{24192} c^6 k^6 \Delta t^6 + \dots \tag{11.82}$$

– In P_3 :

$$q_h = 1 - \frac{1}{720} c^4 k^4 \Delta t^4 - \frac{k^6 h^6}{604800} - \frac{5}{24192} c^6 k^6 \Delta t^6 + \dots \tag{11.83}$$

Unlike the FDM, we obtain here a true fourth-order approximation in time.

Remark

The P_3 approximation is much better than the P_2 one since, besides its higher-order character, the next term of its Taylor expansion is in $O(h^{10})$ instead of $O(h^8)$. This superiority will be illustrated in the dispersion curves.

11.4.2 Dispersion Curves

In Figs. 11.4–11.8, we give different dispersion curves for P_2 and P_3 approximations. The curves for P_1 are the same as for second-order FDM given in Sects. 7.6.1 and 7.6.2. In all the curves, K is the inverse of the number of points per wavelength, which is equal to rN_e , where r is the order of the approximation and N_e , the number of elements per wavelength. This convention is more convenient to compare the different orders of approximation.

In Figs. 11.4–11.6, we give the dispersion curves for P_2 and P_3 with a leapfrog scheme, the fourth-order modified equation approach and the fourth-order symmetric scheme. Of course, the superiority of P_3 over P_2 , confirmed in Fig. 11.7, is always obvious.

However, unlike the FDM, the accuracy for the modified equation approach decreases when α increases. This suggests the use of a value of α slightly less than α_M (about 25%) [115].

On the other hand, one can see that the symmetric scheme could do much better than the modified equation approach. This assertion, confirmed in Fig. 11.8, derives from the fact that we have to use a much smaller α for the symmetric schemes. So, this scheme, which has a coefficient of Δt^4 greater than that of the modified equation approach, provides a better accuracy for its maximum value of α . In terms of efficiency, the modified equation approach requires the computation of two discrete operators whereas the symmetric scheme needs only one. If we compare the curves obtained for $2\alpha_M$ for the modified equation approach with α_M for the symmetric scheme, where α_M is the stability condition of the symmetric scheme, the curves given in Fig. 11.8 show that we obtain a better accuracy for the same CPU time for the symmetric scheme. Unfortunately, besides the fact that this scheme has the drawback of requiring the storage of two more values of \mathbf{U} , the presence of parasitic waves pointed out in its plane wave analysis introduces some ripples in the solutions, as shown in Fig. 11.9. In this figure, we represent the propagation of an initial solution given by

$$(a(x - b)^2 - 1)e^{-a(x-b)^2/2} \quad (11.84)$$

where $a = 5.751183$ and $b = 1.35$, on the interval $[0,12]$ with periodical boundary conditions by using the homogeneous wave equation, at the instant $t = 48$. The approximation in space is made by a P_3 mass-lumped FEM, with 42 elements.

11.5 Some Results on the Amplitudes

In this section, we provide a way of analyzing the global error of the solution on a regular grid by using interpolation functions and Fourier analysis. In particular, this analysis will enable us to obtain valuable informations about the amplitudes of the parasitic waves. The computations will be carried out in the case of a P_2 approximation (which is not obvious to obtain) and its extension to P_3 will be given without proof. In order to simplify the notation, we shall assume that $c = 1$ in this section.

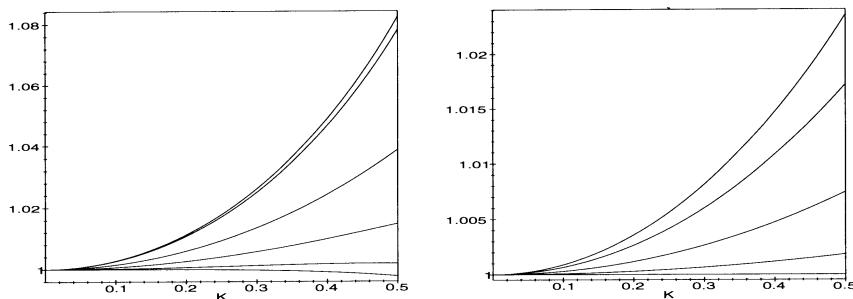


Fig. 11.4. Dispersion curves for a leapfrog scheme in time with P_2 (left) and P_3 (right) approximations in space. The curves are given from $\alpha = 0$ (lower curve) to $\alpha = \alpha_M$ (upper curve) in steps of 0.2 for each approximation

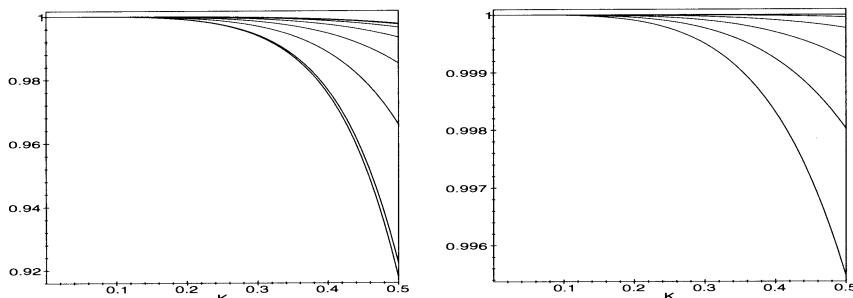


Fig. 11.5. Dispersion curves for a fourth-order approximation in time by the modified equation approach with P_2 (left) and P_3 (right) approximations in space. The curves are given from $\alpha = 0$ (upper curve) to $\alpha = \alpha_M$ (lower curve) in steps of 0.2 for each approximation

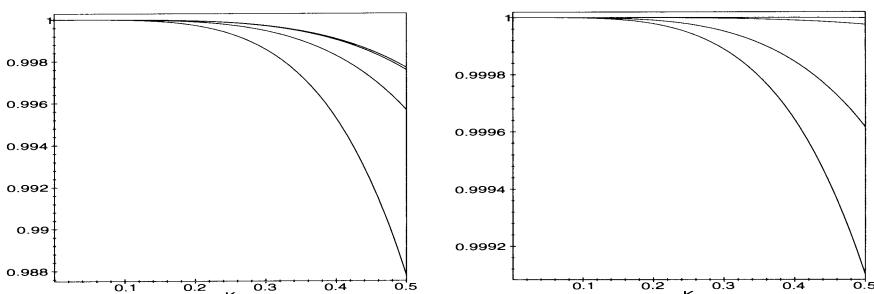


Fig. 11.6. Dispersion curves for a fourth-order approximation in time by the symmetric scheme with P_2 (left) and P_3 (right) approximations in space. The curves are given from $\alpha = 0$ (upper curve) to $\alpha = \alpha_M$ (lower curve) in steps of 0.2 for each approximation

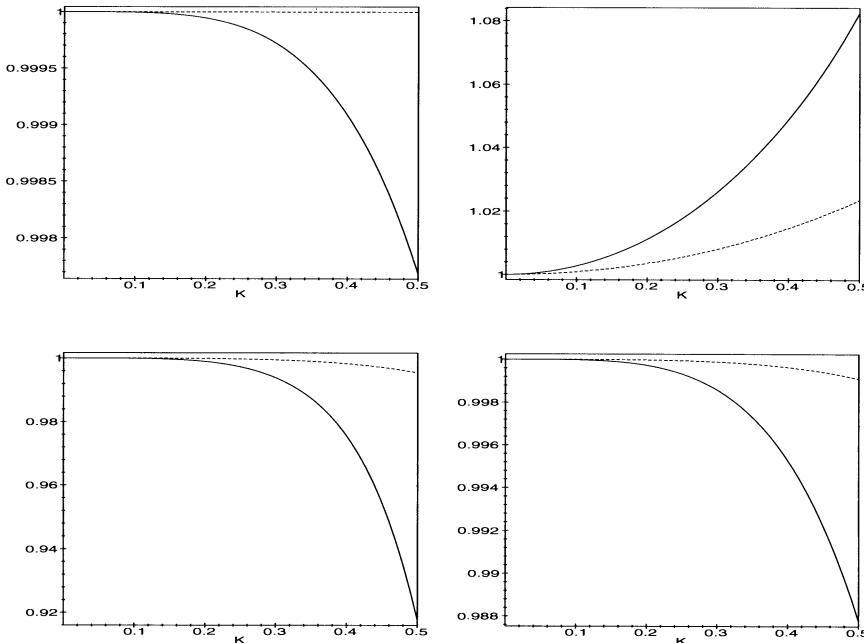


Fig. 11.7. Comparison of P_2 (continuous line) and P_3 (dashed line) approximations in space for a semi-discretization in space (above left), a leapfrog scheme in time (above right), a fourth-order approximation in time by the modified equation approach (below left) and by the symmetric scheme (below right). For each approximation in time, $\alpha = \alpha_M$

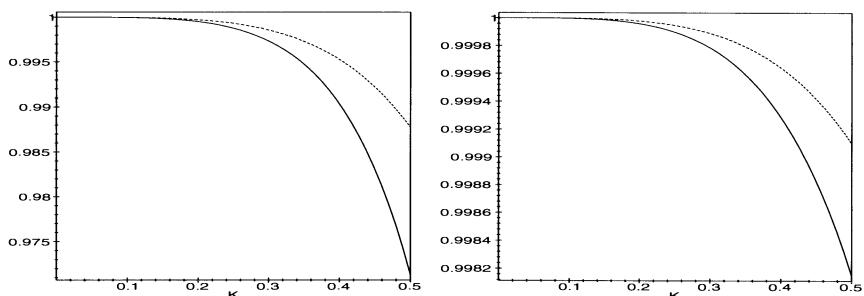


Fig. 11.8. Comparison of the fourth-order approximations in time by the modified equation approach (continuous line) and by the symmetric scheme (dashed line) for P_2 (left) and P_3 (right) approximations in space. Here, $\alpha = \alpha_M$ for the symmetric scheme and α is taken for the modified equation approach such that one obtains the same CPU time as that of the symmetric scheme

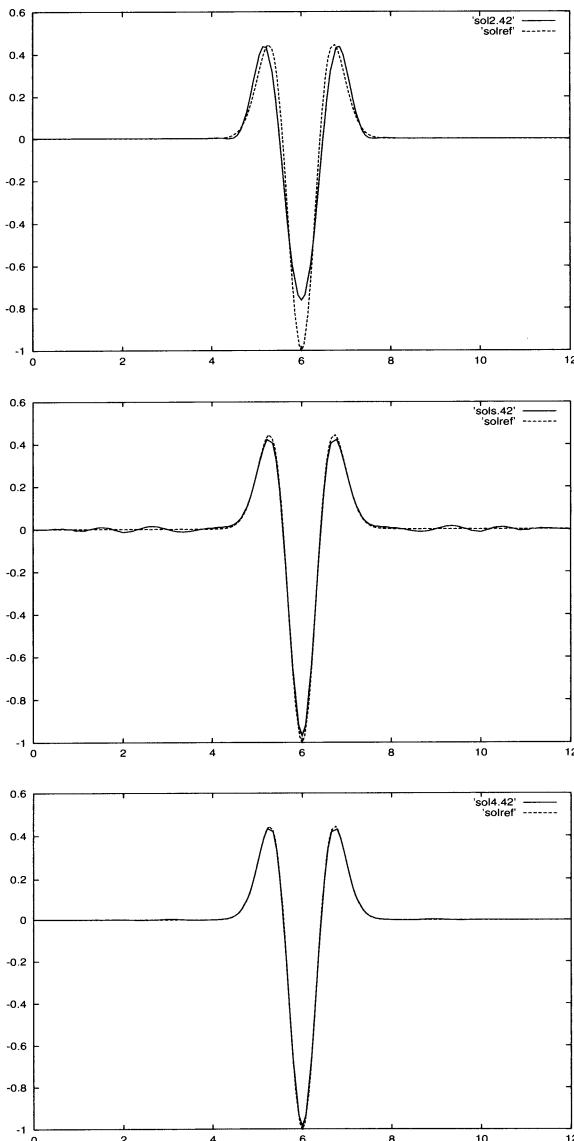


Fig. 11.9. The propagated initial solution given in (11.84) at $t = 48$ by using a P_3 FEM approximation with 42 elements. The *upper* solution is obtained by a leapfrog approximation, the *middle* one, by the symmetric scheme and the *lower* one, by the modified equation approach. All the solutions, compared to a reference one, have the same CPU time, obtained for the maximum CFL of the symmetric scheme

In a first step, we replace the equations (11.21a) and (11.21b) by the “continuified” equations

$$\begin{aligned} \frac{d^2\tilde{u}_{h,0}}{dt^2}(x,t) &= -\frac{1}{h^2}(14\tilde{u}_{h,0}(x,t) - 8(\tilde{u}_{h,1}(x+\frac{h}{2},t) \\ &\quad + \tilde{u}_{h,1}(x-\frac{h}{2},t)) + \tilde{u}_{h,0}(x+h,t) \end{aligned} \quad (11.85a)$$

$$\begin{aligned} \frac{d^2\tilde{u}_{h,1}}{dt^2}(x,t) &= \frac{4}{h^2}(\tilde{u}_{h,0}(x+\frac{h}{2},t) - 2\tilde{u}_{h,1}(x,t) \\ &\quad + \tilde{u}_{h,0}(x-\frac{h}{2},t)) \quad \forall p \in \mathbb{Z}, \end{aligned} \quad (11.85b)$$

where $\tilde{u}_{h,0}$ and $\tilde{u}_{h,1}$ are interpolation functions such that:

$$\tilde{u}_{h,0}(x_p, t) = u_h(x_p, t), \quad (11.86a)$$

$$\tilde{u}_{h,1}(x_{p,1}, t) = u_h(x_{p,1}, t). \quad (11.86b)$$

On the basis of this interpolation, we evaluate $\tilde{u}_{h,0} - u$ and $\tilde{u}_{h,1} - u$ separately. Both differences will provide errors at the points x_p and $x_{p,1}$.

The computation of these errors, uses the following lemma:

Lemma 4. *The eigenvalues λ_1, λ_2 and the eigenvectors $\mathbf{W}_1, \mathbf{W}_2$ of problem (11.24) have the following Taylor expansions:*

$$\lambda_1 = k^2 \left(1 - \frac{k^4 h^4}{1440} - \frac{k^6 h^6}{48384} + O(k^8 h^8) \right), \quad (11.87a)$$

$$\mathbf{W}_1 = \mathbf{X}_0 + a \mathbf{Y}_0 k^4 h^4 + b \mathbf{Y}_0 k^6 h^6 + O(k^8 h^8), \quad (11.87b)$$

$$\lambda_2 = k^2 \left(\frac{24}{k^2 h^2} - 2 + \frac{k^2 h^2}{12} - \frac{k^4 h^4}{480} + O(k^6 h^6) \right), \quad (11.88a)$$

$$\mathbf{W}_2 = \mathbf{Y}_0 - a \mathbf{X}_0 k^4 h^4 - 2b \mathbf{X}_0 k^6 h^6 + O(k^8 h^8), \quad (11.88b)$$

where $\mathbf{X}_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $\mathbf{Y}_0 = \frac{1}{\sqrt{2}} \begin{pmatrix} -2 \\ 1 \end{pmatrix}$, $a = \frac{\sqrt{2}}{1152}$, $b = \frac{\sqrt{2}}{13824}$.

Moreover, the eigenvectors $\mathbf{W}_1, \mathbf{W}_2$ are such that $((\mathbf{W}_\ell, \mathbf{W}_m)) = \delta_{\ell,m}$, where $\delta_{\ell,m}$ is the Kronecker symbol, $((\mathbf{W}_1, \mathbf{W}_2)) = (\widehat{M}_{1,2} \mathbf{W}_1, \mathbf{W}_2)$ and $(,)$ is the Hermitian product of \mathbb{C}^2 . \diamondsuit

The Taylor expansions of the eigenvalues were already computed. The computation of the eigenvectors, which is more technical, will be given at the end of this chapter.

Now, the application of the Fourier transform in space \mathcal{F}_x to the problem (11.85a) and (11.85b) leads to the problem

$$\widehat{M}_{1,2} \frac{d^2 \widehat{\mathbf{U}}_{h,0}}{dt^2} = \widehat{K}_{1,2} \widehat{\mathbf{U}}_{h,0}, \quad (11.89)$$

where $\widehat{M}_{1,2}$ and $\widehat{K}_{1,2}$ are defined as in (11.24) and $\widehat{\mathbf{U}}_{h,0} = (\widehat{u}_{h,0}, \widehat{u}_{h,1})^T$, with

$$\widehat{u}_{h,0}(k, t) = \mathcal{F}_x(\widetilde{u}_{h,0}(x, t)), \quad (11.90a)$$

$$\widehat{u}_{h,1}(k, t) = \mathcal{F}_x(\widetilde{u}_{h,1}(x, t)). \quad (11.90b)$$

To (11.89), we add the initial values

$$\widehat{u}_{h,0}(k, 0) = \widehat{u}_{h,1}(k, 0) = \widehat{u}_0(k), \quad (11.91a)$$

$$\frac{\partial \widehat{u}_{h,0}}{\partial t}(k, 0) = \frac{\partial \widehat{u}_{h,1}}{\partial t}(k, 0) = \widehat{u}_1(k). \quad (11.91b)$$

In order to solve (11.89), one must diagonalize the matrix $\widehat{N}_{1,2} = \widehat{M}_{1,2}^{-1} \widehat{K}_{1,2}$. For this purpose, we shall use the eigenvalues and eigenvectors defined in lemma 4.

So, on the basis of these quantities, an elementary computation shows that the solution of (11.89) can be written as

$$\widehat{\mathbf{U}}_{h,0} = \widehat{U}_{h,1} \mathbf{W}_1 + \widehat{U}_{h,2} \mathbf{W}_2, \quad (11.92)$$

where

$$\widehat{U}_{h,\ell} = \widehat{u}_0(k) \alpha_{h,\ell}(k) \cos(\omega_{h,\ell}(k) t) + \widehat{u}_1(k) \alpha_{h,\ell}(k) \frac{\sin(\omega_{h,\ell}(k) t)}{\omega_{h,\ell}(k)}, \quad (11.93)$$

with $\alpha_{h,\ell}(k) = ((\mathbf{1}, \mathbf{W}_\ell(k)))$, $\mathbf{1} = (1, 1)^T$ and $\omega_{h,\ell}(k) = \sqrt{\lambda_\ell(k)}$, $\forall \ell = 1, 2$.

Following the definitions of the eigenvectors, $\widehat{U}_{h,1} \mathbf{W}_1$ corresponds to the physical part of the solution

$$\widehat{U}_{h,1} \mathbf{W}_1 = \begin{pmatrix} \widehat{u}_{h,0}^{np} \\ \widehat{u}_{h,1}^{np} \end{pmatrix} \quad (11.94)$$

and $\widehat{U}_{h,2} \mathbf{W}_2$ to its parasitic part

$$\widehat{U}_{h,2} \mathbf{W}_2 = \begin{pmatrix} \widehat{u}_{h,0}^p \\ \widehat{u}_{h,1}^p \end{pmatrix}. \quad (11.95)$$

Now, we derive some error estimates from this solution.

11.5.1 Error Estimates on the Physical Part of the Solution

The estimation of $\|\widehat{u}_{h,0}^{np}(.,t) - \widehat{u}(.,t)\|$ and $\|\widehat{u}_{h,1}^{np} - \widehat{u}(.,t)\|$, is based on the Plancherel theorem which provides the following identities:

$$\|\widehat{u}_{h,0}^{np}(.,t) - \widehat{u}(.,t)\| = \|\widetilde{u}_{h,0}^{np}(.,t) - u(.,t)\|, \quad (11.96a)$$

$$\|\widehat{u}_{h,1}^{np} - \widehat{u}(.,t)\| = \|\widetilde{u}_{h,1}^{np} - u(.,t)\|, \quad (11.96b)$$

where $\|\cdot\|$ is the L^2 norm on \mathbb{R} .

Let us set:

$$\begin{pmatrix} \widehat{e}_{h,0}^{np}(k,t) \\ \widehat{e}_{h,1}^{np}(k,t) \end{pmatrix} = \begin{pmatrix} \widehat{u}_{h,0}^{np} - \widehat{u} \\ \widehat{u}_{h,1}^{np} - \widehat{u} \end{pmatrix}. \quad (11.97)$$

Since

$$\widehat{u}(k,t) = \widehat{u}_0(k) \cos(\omega(k)t) + \widehat{u}_1(k) \frac{\sin(\omega(k)t)}{\omega(k)}, \quad (11.98)$$

we have, $\forall j = 0, 1$:

$$\begin{aligned} \widehat{e}_{h,j}^{np}(k,t) &= \widehat{u}_0(k) [\alpha_{h,1}(k) \cos(\omega_{h,1}(k)t) - \cos(\omega(k)t)] (\mathbf{W}_1)_{j+1} \\ &\quad + \widehat{u}_1(k) \left[\alpha_{h,1}(k) \frac{\sin(\omega_{h,1}(k)t)}{\omega_{h,1}(k)} - \frac{\sin(\omega(k)t)}{\omega(k)} \right] (\mathbf{W}_1)_{j+1}, \end{aligned} \quad (11.99)$$

where $(\mathbf{W}_1)_{j+1}$, $j = 0, 1$ is the $(j+1)$ th component of \mathbf{W}_1 .

Using the triangular inequality and the fact that $\omega = k$, we obtain:

$$\begin{aligned} \|\widehat{e}_{h,j}^{np}(k,t)\| &\leq \\ &\left[\int_{\mathbb{R}} \left(\widehat{u}_0(k) \cos(\omega_{h,1}(k)t) + \widehat{u}_1(k) \frac{\sin(\omega_{h,1}(k)t)}{\omega_{h,1}(k)} \right)^2 \right. \\ &\quad \left. (\alpha_{h,1}(k) (\mathbf{W}_1)_{j+1} - 1)^2 dk \right]^{\frac{1}{2}} \end{aligned} \quad (11.100)$$

$$\begin{aligned} &+ \left[\int_{\mathbb{R}} (\cos(\omega_{h,1}(k)t) - \cos(kt))^2 dk \right]^{\frac{1}{2}} \\ &+ \left[\int_{\mathbb{R}} \left(\frac{\sin(\omega_{h,1}(k)t)}{\omega_{h,1}(k)} - \frac{\sin(kt)}{k} \right)^2 dk \right]^{\frac{1}{2}}. \end{aligned}$$

Now, by using the mean value theorem, we obtain:

$$(\cos(\omega_{h,1}(k)t) - \cos(kt))^2 \leq t^2 (\omega_{h,1}(k) - k)^2, \quad (11.101a)$$

$$|\sin(\omega_{h,1}(k)t) - \sin(kt)| \leq |\omega_{h,1}(k) - k|t, \quad (11.101b)$$

$$|\sin(\omega_{h,1}(k)t)| \leq |\omega_{h,1}(k)|t. \quad (11.101c)$$

So, by noticing that

$$\begin{aligned} \frac{\sin(\omega_{h,1}(k)t)}{\omega_{h,1}(k)} - \frac{\sin(kt)}{k} &= \sin(\omega_{h,1}(k)t) \left(\frac{k - \omega_{h,1}(k)}{k\omega_{h,1}(k)} \right) \\ &\quad + \frac{1}{k} (\sin(\omega_{h,1}(k)t) - \sin(kt)) \end{aligned} \quad (11.102)$$

and by combining (11.101a) and (11.102) with (11.87a), we obtain:

$$|\cos(\omega_{h,1}(k)t) - \cos(kt)| \leq Ck^5h^4t, \quad (11.103a)$$

$$\left| \frac{\sin(\omega_{h,1}(k)t)}{\omega_{h,1}(k)} - \frac{\sin(kt)}{k} \right| \leq Ck^5h^4t. \quad (11.103b)$$

On the other hand, we have $\forall j = 1, 2$:

$$\left| \widehat{u}_0(k) \cos(\omega_{h,j}(k)t) + \widehat{u}_1(k) \frac{\sin(\omega_{h,j}(k)t)}{\omega_{h,j}(k)} \right| \leq |\widehat{u}_0(k)| + |\widehat{u}_1(k)|t, \quad (11.104a)$$

$$|(\alpha_{h,1}(k)(W_1)_{j+1} - 1)_j| \leq Ck^4h^4. \quad (11.104b)$$

Finally, by inserting (11.103a) and (11.104b) into (11.100), we obtain:

$$\begin{aligned} \|\widehat{e}_{h,j}^{np}(k,t)\| &\leq \\ Ch^4 &\left[\left(\int_{\mathbb{R}} k^8 |\widehat{u}_0(k)|^2 dk \right)^{\frac{1}{2}} + t \left(\int_{\mathbb{R}} k^8 |\widehat{u}_1(k)|^2 dk \right)^{\frac{1}{2}} \right. \\ &\quad \left. + \left(\int_{\mathbb{R}} k^{10} |\widehat{u}_0(k)|^2 dk \right)^{\frac{1}{2}} + t \left(\int_{\mathbb{R}} k^8 |\widehat{u}_1(k)|^2 dk \right)^{\frac{1}{2}} \right], \end{aligned} \quad (11.105)$$

which provides, $\forall j = 1, 2$, after applying the Plancherel theorem in the other way and assuming that $u_0 \in C^5(\mathbb{R})$ and that $u_1 \in C^4(\mathbb{R})$:

$$\|e_{h,j}^{np}(x,t)\| \leq Ch^4 \left\{ \left\| \frac{d^4 u_0}{dx^4} \right\| + t \left\| \frac{d^5 u_0}{dx^5} \right\| + 2t \left\| \frac{d^4 u_1}{dx^4} \right\| \right\}. \quad (11.106)$$

11.5.2 Error Estimates on the Parasitic Part of the Solution

As for the physical part of the solution, we set, for the parasitic part:

$$\begin{pmatrix} \widehat{e}_{h,0}^p(k,t) \\ \widehat{e}_{h,1}^p(k,t) \end{pmatrix} = \begin{pmatrix} \widehat{u}_{h,0}^p \\ \widehat{u}_{h,1}^p \end{pmatrix} = \widehat{U}_{h,2} \mathbf{W}_2. \quad (11.107)$$

With these notations, we have, $\forall j = 1, 2$

$$\begin{aligned} \|\widehat{e}_{h,j}^{np}(k,t)\| &= \\ &\left[\int_{\mathbb{R}} \left(\widehat{u}_0(k) \cos(\omega_{h,2}(k)t) + \widehat{u}_1(k) \frac{\sin(\omega_{h,2}(k)t)}{\omega_{h,2}(k)} \right)^2 \right. \\ &\quad \left. (\alpha_{h,2}(k) (\mathbf{W}_2)_{j+1})^2 dk \right]^{\frac{1}{2}}. \end{aligned} \quad (11.108)$$

Since $\alpha_{h,2}(k) \mathbf{W}_2 = Ck^4 h^4 \mathbf{1} = O(h^6)$, we obtain, $\forall j = 1, 2$, by using (11.104b) and applying the Plancherel theorem:

$$\|e_{h,j}^p(x,t)\| \leq Ch^4 \left\{ \left\| \frac{d^4 u_0}{dx^4} \right\| + t \left\| \frac{d^4 u_1}{dx^4} \right\| \right\}. \quad (11.109)$$

This last result shows that the parasitic waves decrease in $O(h^4)$, which confirms their negligible effect. This result can be extended to the 2D case [115]. However, for non-regular meshes, the parasitic waves can be more troublesome, as we shall show in the next chapter.

A similar study was made for P_3 and provided an error in h^6 at the ends of the elements and in h^5 at the interior points for both physical and parasitic waves [115].

Remarks

1. The error obtained by this technique includes the error of interpolation derived from the interpolated solution. Therefore, one can simply say that the true error is at least equal to what we have obtained.
2. Although more natural, the use of the discrete Fourier transform on the discrete problem does not enable us to obtain error estimates.

11.6 Reflection-Transmission Analysis

The reflection-transmission analysis follows the same principles as for finite differences. Here, we give guidelines for formulating the equations satisfied by

the transmitted and reflected waves for the P_2 approximation. Then, we shall simply give the order of the error in this case and in the case of higher-order approximations.

11.6.1 FEM Approximation of the Heterogeneous Wave Equation

The variational formulation of the 1D heterogeneous wave equation

$$\eta(x) \frac{\partial^2 u}{\partial t^2} - \frac{\partial}{\partial x} \left(\gamma(x) \frac{\partial u}{\partial x} \right) = 0, \quad (11.110)$$

is obviously

$$\frac{d^2}{dt^2} \int_{\mathbb{R}} \eta u v dx + \int_{\mathbb{R}} \gamma \frac{\partial u}{\partial x} \frac{\partial v}{\partial x} dx = 0 \quad \forall v \in H^1(\mathbb{R}). \quad (11.111)$$

In this case, it is necessary to compute the mass integral and the stiffness integral by using a Gauss-Lobatto quadrature rule, because of the presence of η and γ . This process leads, for P_2 , to the following semi-discrete in space equations on a regular mesh:

$$\begin{aligned} \eta_p \frac{d^2 u_p}{dt^2}(t) &= \frac{1}{h^2} \left[-\frac{1}{2} (3\gamma_{p-1} - 4\gamma_{p-1,1} + 3\gamma_p) u_{p-1}(t) \right. \\ &\quad + 2(\gamma_{p-1} + 3\gamma_p) u_{p-1,1}(t) \\ &\quad - \frac{1}{2} (\gamma_{p-1} + 4\gamma_{p-1,1} + 18\gamma_p + 4\gamma_{p,1} + \gamma_{p+1}) u_p(t) \quad (11.112a) \\ &\quad + 2(\gamma_p + 3\gamma_{p+1}) u_{p,1}(t) \\ &\quad \left. - \frac{1}{2} (3\gamma_p - 4\gamma_{p,1} + 3\gamma_{p+1}) u_{p+1}(t) \right], \quad \forall p \in \mathbb{Z}, \end{aligned}$$

$$\begin{aligned} \eta_{p,1} \frac{d^2 u_{p,1}}{dt^2}(t) &= \frac{1}{h^2} [(3\gamma_p + \gamma_{p+1}) u_p(t) - 4(\gamma_p + \gamma_{p+1}) u_{p,1}(t) \quad (11.112b) \\ &\quad + (\gamma_p + 3\gamma_{p+1}) u_{p+1}(t)], \quad \forall p \in \mathbb{Z}, \end{aligned}$$

where $\eta_p = \eta(x_p)$, $\eta_{p,1} = \eta(x_{p,1})$, $\gamma_p = \gamma(x_p)$ and $\gamma_{p,1} = \gamma(x_{p,1})$.

11.6.2 Taylor Expansion of the Wavenumber

As in Sect. 10.1, we look for an expression of k in terms of ωh in the first step of our study. This expression can be derived from the characteristic polynomial of (11.24) which can be written as

$$4 \left(24 + \frac{\omega^2 h^2}{c^2} \right) \sin^2 \frac{kh}{2} - \frac{\omega^2 h^2}{c^2} \left(24 - \frac{\omega^2 h^2}{c^2} \right) = 0, \quad (11.113)$$

from which one can easily derive:

$$k = \pm \frac{2}{h} \arcsin \left(\frac{\omega h}{2c} \sqrt{\left(24 - \frac{\omega^2 h^2}{c^2} \right) \left(24 + \frac{\omega^2 h^2}{c^2} \right)^{-1}} \right). \quad (11.114)$$

The Taylor expansion of (11.114) is

$$k = \pm \frac{\omega}{c} \left(1 + \frac{\omega^4 h^4}{2880 c^4} + \frac{\omega^6 h^6}{96768 c^6} + O(h^8) \right). \quad (11.115)$$

Unlike fourth-order FDM, the equation giving k versus ωh has two solutions of opposite sign although the inverse equation has four solutions. This implies, in particular, that we have no parasitic wave coming from the interface as we had in the case of FDM.

Now, let us consider a two-layer infinite medium such that $\eta = \eta_1$ and $\gamma = \gamma_1$ for $x < 0$ and $\eta = \eta_2$ and $\gamma = \gamma_2$ for $x > 0$. In each layer, we write

$$k_j = \pm \frac{\omega}{c_j} \left(1 + \frac{\omega^4 h^4}{2880 c_j^4} + \frac{\omega^6 h^6}{96768 c_j^6} + O(h^8) \right), \quad j = 1, 2. \quad (11.116)$$

11.6.3 Interface Between Two Elements

Our first study will be carried out for the case of an interface located between two elements. In this case, the solution can be written as

$$u_p = \begin{cases} \alpha_1 (e^{i(\omega t - pk_1 h)} + R_h e^{i(\omega t + pk_1 h)}) & \text{for } p \leq 0, \\ \alpha_2 T_h e^{i(\omega t - pk_2 h)} & \text{for } p \geq 0, \end{cases} \quad (11.117a)$$

$$u_{p,1} = \begin{cases} \beta_1 (e^{i(\omega t - (p+\frac{1}{2})k_1 h)} + R_h e^{i(\omega t + (p+\frac{1}{2})k_1 h)}) & \text{for } p \leq -1, \\ \beta_2 T_h e^{i(\omega t - (p+\frac{1}{2})k_2 h)} & \text{for } p \geq 0, \end{cases} \quad (11.117b)$$

where $(\alpha_j, \beta_j)^T$, $j = 1, 2$ is the eigenvector corresponding to the physical solution of the discrete system given in (11.87b) (which is the correct one for $\omega h / c < 2$), in which k is replaced by k_j .

The reflection-transmission coefficients R_h and T_h are determined by

- the condition of continuity at $x = 0$:

$$\alpha_1 (1 + R_h) = \alpha_2 T_h, \quad (11.118)$$

– the equation at this same point:

$$\begin{aligned} \omega^2 \frac{\eta_1 + \eta_2}{2} u_0 &= \frac{1}{h^2} (7(\gamma_1 + \gamma_2)u_0 - 8(\gamma_2 u_{0,1} + \gamma_1 u_{-1,1}) \\ &\quad + \gamma_2 u_1 + \gamma_1 u_{-1}), \end{aligned} \quad (11.119)$$

in which we plug (11.117a) and (11.117b).

An elementary computation (aided by Maple) leads to the following result:

$$R_h = \frac{\sigma_1 - \sigma_2}{\sigma_1 + \sigma_2} + \frac{1}{288} \frac{\sigma_1 \sigma_2}{(\sigma_1 + \sigma_2)^2} \left[\frac{1}{c_1^4} - \frac{1}{c_2^4} \right] \omega^4 h^4 + O(h^6), \quad (11.120a)$$

$$T_h = \frac{2\sigma_1}{\sigma_1 + \sigma_2} - \frac{1}{288} \frac{\sigma_1^2}{(\sigma_1 + \sigma_2)^2} \left[\frac{1}{c_1^4} - \frac{1}{c_2^4} \right] \omega^4 h^4 + O(h^6), \quad (11.120b)$$

where

$$\sigma_j = \sqrt{\eta_j \gamma_j}, \quad j = 1, 2. \quad (11.121)$$

Equations (11.120a) and (11.120b) show that the error committed on the reflection-transmission coefficients is, in this case, of the order of the dispersion.

11.6.4 Interface at an Interior Point

When the interface is located inside the element whose ends are 0 and h , three equations (instead of one) mix the two layers together: those at $x = 0$, at $x = h/2$ and at $x = h$. In order to have as many equations as unknowns, we consider the value $u_{0,1}$ as an unknown that we add to R and T .

The three equations are

$$\begin{aligned} \omega^2 \eta_1 u_0 &= \frac{1}{h^2} \left(\frac{25\gamma_1 + 3\gamma_2}{2} u_0 - (6\gamma_1 + 2\gamma_2)u_{0,1} \right. \\ &\quad \left. - 8\gamma_1 u_{-1,1} + \frac{\gamma_1 + \gamma_2}{2} u_1 + \gamma_1 u_{-1} \right), \end{aligned} \quad (11.122a)$$

$$\begin{aligned} \omega^2 \frac{\eta_1 + \eta_2}{2} u_{0,1} &= -\frac{1}{h^2} ((3\gamma_1 + \gamma_2)u_0 + (3\gamma_2 + \gamma_1)u_1 \\ &\quad - 4(\gamma_1 + \gamma_2)u_{0,1}), \end{aligned} \quad (11.122b)$$

$$\begin{aligned} \omega^2 \eta_2 u_1 &= \frac{1}{h^2} \left(\frac{\gamma_1 + \gamma_2}{2} u_0 - (6\gamma_2 + 2\gamma_1)u_{0,1} \right. \\ &\quad \left. - 8\gamma_2 u_{1,1} + \frac{25\gamma_2 + 3\gamma_1}{2} u_1 + \gamma_2 u_2 \right). \end{aligned} \quad (11.122c)$$

By inserting the plane wave solution

$$u_p = \begin{cases} \alpha_1(e^{i(\omega t - pk_1 h)} + R_h e^{i(\omega t + pk_1 h)}) & \text{for } p \leq 0, \\ \alpha_2 T_h e^{i(\omega t - pk_2 h)} & \text{for } p \geq 1, \end{cases} \quad (11.123a)$$

$$u_{p,1} = \begin{cases} \beta_1(e^{i(\omega t - (p+\frac{1}{2})k_1 h)} + R_h e^{i(\omega t + (p+\frac{1}{2})k_1 h)}) & \text{for } p \leq -1, \\ \beta_2 T_h e^{i(\omega t - (p+\frac{1}{2})k_2 h)} & \text{for } p \geq 1, \end{cases} \quad (11.123b)$$

into (11.122a)–(11.122c), we obtain the following orders for R_h and T_h :

$$R_h = \frac{\sigma_1 - \sigma_2}{\sigma_1 + \sigma_2} + O(h), \quad (11.124a)$$

$$T_h = \frac{2\sigma_1}{\sigma_1 + \sigma_2} + O(h). \quad (11.124b)$$

This result shows that, if one locates the interface of discontinuity at an interior point of the element, one loses the accuracy provided by the use of a higher-order method.

Remark

The condition of continuity is contained in equations (11.122a) and (11.122b) since the solution is not expressed as an explicit plane wave in the interval $]0, h[$.

11.6.5 Extension to Higher-Order Approximations

We carried out the computations up to a P_5 approximation. For all the orders, the characteristic polynomial provided only two solutions for k in terms of ωh , which means that we never have parasitic waves generated by a discontinuity. So, for an interface between two elements, the equations remain the same i.e., the condition of continuity at $x = 0$ and the equation at the same point. We obtained the following results [55]:

– P_3 elements:

$$R_h = \frac{\sigma_1 - \sigma_2}{\sigma_1 + \sigma_2} + \frac{1}{1800} \frac{\sigma_1(\sigma_2 - \sigma_1)}{c_1^4(\sigma_1 + \sigma_2)^2} \omega^4 h^4 + O(h^6), \quad (11.125a)$$

$$T_h = \frac{2\sigma_1}{\sigma_1 + \sigma_2} + \frac{1}{1800} \frac{\sigma_1(\sigma_2 - \sigma_1)}{c_1^4(\sigma_1 + \sigma_2)^2} \omega^4 h^4 + O(h^6). \quad (11.125b)$$

– P_4 elements:

$$R_h = \frac{\sigma_1 - \sigma_2}{\sigma_1 + \sigma_2} + \frac{1}{470\,400} \frac{\sigma_1(\sigma_2 - \sigma_1)}{c_1^6(\sigma_1 + \sigma_2)^2} \omega^6 h^6 + O(h^8), \quad (11.126a)$$

$$T_h = \frac{2\sigma_1}{\sigma_1 + \sigma_2} + \frac{1}{470\,400} \frac{\sigma_1(\sigma_2 - \sigma_1)}{c_1^6(\sigma_1 + \sigma_2)^2} \omega^6 h^6 + O(h^8). \quad (11.126b)$$

– P_5 elements:

$$R_h = \frac{\sigma_1 - \sigma_2}{\sigma_1 + \sigma_2} - \frac{1}{1\,587\,600} \frac{\sigma_1(\sigma_2 - \sigma_1)}{c_1^6(\sigma_1 + \sigma_2)^2} \omega^6 h^6 + O(h^8), \quad (11.127a)$$

$$T_h = \frac{2\sigma_1}{\sigma_1 + \sigma_2} - \frac{1}{1\,587\,600} \frac{\sigma_1(\sigma_2 - \sigma_1)}{c_1^6(\sigma_1 + \sigma_2)^2} \omega^6 h^6 + O(h^8). \quad (11.127b)$$

These results show that the order increases by two orders of approximation. It is interesting to compare the order of approximation of the reflection transmission with the L^2 error estimates of these methods obtained without mass-lumping (which should be the same with mass-lumping) and the order of the numerical dispersion. These different orders, given in Table 11.3, show that the reflection-transmission is roughly of the same order as the global error estimates whereas the numerical dispersion always has a superconvergence phenomenon.

Table 11.3. Comparison of the errors for different orders of approximation

	L^2 error estimates	Reflection-transmission	Numerical dispersion
P_2	h^3	h^4	h^4
P_3	h^4	h^4	h^6
P_4	h^5	h^6	h^8
P_5	h^6	h^6	h^{10}

When the interface is inside an element, we obtain $r + 1$ equations and we add the values of u_h at the $r - 1$ interior points as unknowns. For all the orders of approximation, we obtain an error in $O(h)$. This result shows that

one must carefully treat the discontinuities when one uses finite elements. As we shall see in the next chapter, these results seem to hold in 2D and 3D. So, for multidimensional meshes, the mesh must follow the interfaces of discontinuities in order to maintain a good accuracy. In particular, for very discontinuous media (as in seismic simulations, for instance), the accuracy of the method can be completely destroyed if the mesh does not follow the interfaces.

Remark

Unlike the finite difference method, we obtain a high-accurate treatment of discontinuities without any artifact. This is one of the advantages of the finite element methods over finite difference methods.

11.7 Taylor Expansions of the Eigenvectors

As we saw in Sects. 11.5 and 11.6, some computations require the knowledge of the Taylor expansion of the eigenvectors of the problems obtained by plane wave analysis. In this section, we show how to compute such a Taylor expansion for a P_2 approximation. We shall only treat the case of the physical eigenvalue, the other case being computed in the same way.

The first step of this computation consists in writing the Taylor expansion of the matrix introduced in (11.24), i.e.

$$\widehat{N}_{1,2} = \frac{N_{-2}}{K^2} + N_0 + N_2 K^2 + N_4 K^4 + N_6 K^6 + O(K^8), \quad (11.128)$$

where $K = kh$ and

$$N_{-2} = 8 \begin{pmatrix} 2 & -2 \\ -1 & 1 \end{pmatrix}, \quad N_0 = \begin{pmatrix} -1 & 2 \\ -1 & 0 \end{pmatrix}, \quad N_2 = \frac{1}{48} \begin{pmatrix} 4 & -2 \\ -1 & 0 \end{pmatrix},$$

$$N_4 = \frac{1}{5760} \begin{pmatrix} -16 & 2 \\ 1 & 0 \end{pmatrix}, \quad N_6 = \frac{1}{1290240} \begin{pmatrix} 64 & -2 \\ -1 & 0 \end{pmatrix}.$$

In a second step, we write down the formal series of the eigenvector

$$\mathbf{W}_1 = \sum_{j=0}^{+\infty} S_{2j} K^{2j}. \quad (11.129)$$

Then, we solve the eigenvalue problem

$$\widehat{N}_{1,2} \mathbf{W}_1 = \tilde{\lambda}_1 \mathbf{W}_1, \quad (11.130)$$

where $\tilde{\lambda}_1 = \lambda_1/k^2$ with λ_1 given in (11.87a), by identifying the different terms of the Taylor expansions.

For any vector \mathbf{S}_{2j} , we set $\mathbf{S}_{2j} = (x^{(2j)}, y^{(2j)})^T$. The first equation is derived from the term in K^{-2} which provides

$$N_{-2} \mathbf{S}_0 = 0, \quad (11.131)$$

whose solution is $x^{(0)} = y^{(0)} = \alpha$.

To determine α , we use the fact that the eigenvector is sought such that

$$|\mathbf{W}_1| = 1, \quad (11.132)$$

where $|\mathbf{W}_1|^2 = ((\mathbf{W}_1, \mathbf{W}_1)) = (\widehat{M}_{1,2} \mathbf{W}_1, \mathbf{W}_1)$, (\cdot, \cdot) being the Hermitian product of \mathbb{C}^2 .

Equation (11.132) implies that $|\mathbf{S}_0| = 1$ and then, $\alpha = 1$. So,

$$\mathbf{S}_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \quad (11.133)$$

The constant term leads to the equation

$$N_{-2} \mathbf{S}_2 + N_0 \mathbf{S}_0 = \mathbf{S}_0, \quad (11.134)$$

which is equivalent to

$$N_{-2} \mathbf{S}_2 = 0, \quad (11.135)$$

whose solution is $\mathbf{S}_2 = \alpha \mathbf{S}_0$. On the other hand, we obtain from (11.132) the relation $\alpha |\mathbf{S}_0|^2 = 0$, so $\alpha = 0$.

From the term in K^2 , we obtain

$$N_{-2} \mathbf{S}_4 + N_0 \mathbf{S}_2 + N_2 \mathbf{S}_0 = 0. \quad (11.136)$$

Since $\mathbf{S}_2 = 0$, this equation leads to the system

$$768 \left(x^{(4)} - y^{(4)} \right) = 2, \quad (11.137a)$$

$$384 \left(y^{(4)} - x^{(4)} \right) = -1, \quad (11.137b)$$

whose general solution is $x^{(4)} = y^{(4)} + 1/384$. Here also, $(x^{(4)}, y^{(4)})$ are determined by using (11.132) which provides $2((\mathbf{S}_0, \mathbf{S}_4)) = 0$ and from which we obtain

$$\mathbf{S}_4 = \frac{1}{1152} \begin{pmatrix} 2 \\ -1 \end{pmatrix}. \quad (11.138)$$

In the same way, the term in K^4 leads to the equation

$$N_{-2} \mathbf{S}_6 + N_0 \mathbf{S}_4 + N_4 \mathbf{S}_0 = \mathbf{S}_4 - \frac{1}{1440} \mathbf{S}_0, \quad (11.139)$$

which, combined with the fact that (11.132) implies that $2((\mathbf{S}_0, \mathbf{S}_6)) = 0$, provides finally:

$$\mathbf{S}_6 = \frac{1}{13824} \begin{pmatrix} 2 \\ -1 \end{pmatrix}. \quad (11.140)$$

12. Spectral Elements

12.1 Construction of Quadrilateral and Hexahedral Finite Elements

The purpose of this chapter is to construct mass-lumped FEM in higher dimensions and to extend the results of the plane wave analysis carried out in 1D to 2D and 3D. We shall call the mass-lumped FEM the *spectral element method*, following the terminology of [83, 84]. In this first section, we shall deal with the most natural extensions of the 1D finite elements, i.e. quadrilateral elements in 2D and hexahedral elements in 3D, which can be straightforwardly derived from the 1D case by the means of Cartesian and tensor products. In order to construct spectral elements on non-regular meshes, we must first introduce the reference elements, i.e. elements defined on the unit square or the unit cube.

12.1.1 Reference Spectral Elements

Let us consider the 1D P_r finite element on the interval $[0, 1]$ defined in Sect. 11.1.1 and the set¹:

$$\Xi_1 = \left\{ \hat{\xi}_p, p = 1..r+1 \right\}, \quad (12.1)$$

where $\hat{\xi}_p$ are the Gauss-Lobatto quadrature points on this interval. The degrees of freedom of this finite element are the values of the functions of P_r at the points of Ξ_1 .

The basis functions for this element are the Lagrange interpolation functions $\hat{\varphi}_\ell$ defined in (11.11) which satisfy the relation:

$$\hat{\varphi}_\ell(\hat{\xi}_p) = \delta_{\ell p}, \quad (12.2)$$

where $\delta_{\ell p}$ is the Kronecker symbol.

¹ The hat symbol on a variable will indicate that we are on a unit element $[0, 1]^d$, $d = 1..3$.

From this element, we can derive an element defined on the unit square $\hat{K} = [0, 1]^2$ in the following way:

We first define the set of polynomials on \mathbb{R}^2 :

$$Q_r = \left\{ v(\hat{\mathbf{x}}) = \sum_{\ell=0}^r \sum_{m=0}^r a_{\ell,m} \hat{x}_1^\ell \hat{x}_2^m, a_{\ell,m} \in \mathbb{R} \right\}, \quad (12.3)$$

where $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2)$.

Then, we define the set of points of $[0, 1]^2$

$$\Xi_2 = \left\{ \hat{\xi}_{p,q} = (\hat{\xi}_p, \hat{\xi}_q), p = 1..r+1, q = 1..r+1 \right\}, \quad (12.4)$$

which is actually the Cartesian square of Ξ_1 .

In a third step, we define the degrees of freedom on the unit square by the values of the functions of Q_r at these points. Obviously, the basis functions for this finite element are the functions $\hat{\varphi}_{\ell,m}$ defined by

$$\hat{\varphi}_{\ell,m} = \hat{\varphi}_\ell \hat{\varphi}_m. \quad (12.5)$$

As in 1D, we have:

$$\hat{\varphi}_{\ell,m}(\hat{\xi}_{p,q}) = \delta_{\ell p} \delta_{m q}. \quad (12.6)$$

After the construction of the reference finite element, we must construct a 2D Gauss-Lobatto quadrature formula. This construction is derived from the following:

Let f be an integrable function of $\hat{\mathbf{x}}$. By using Fubini, one can write:

$$\begin{aligned} \int_{\hat{K}} f(\hat{\mathbf{x}}) d\hat{\mathbf{x}} &= \int_0^1 \int_0^1 f(\hat{\mathbf{x}}) d\hat{\mathbf{x}} = \int_0^1 \left(\int_0^1 f(x_1, x_2) dx_2 \right) dx_1 \\ &\simeq \int_0^1 \left(\sum_{q=1}^{r+1} \hat{\omega}_q f(x_1, \hat{\xi}_q) \right) dx_1 \\ &\simeq \sum_{p=1}^{r+1} \hat{\omega}_p \left(\sum_{q=1}^{r+1} \hat{\omega}_q f(\hat{\xi}_p, \hat{\xi}_q) \right) \\ &= \sum_{p=1}^{r+1} \sum_{q=1}^{r+1} \hat{\omega}_p \hat{\omega}_q f(\hat{\xi}_p, \hat{\xi}_q) = \sum_{p=1}^{r+1} \sum_{q=1}^{r+1} \hat{\omega}_{p,q} f(\hat{\xi}_{p,q}), \end{aligned} \quad (12.7)$$

where, of course, $\hat{\omega}_{p,q} = \hat{\omega}_p \hat{\omega}_q$.

On the other hand, since our 1D Gauss-Lobatto rule is exact for P_{2r-1} , we can write for any monomial $\hat{x}_1^\ell \hat{x}_2^m$ so that $0 \leq \ell \leq 2r-1$ and $0 \leq m \leq 2r-1$:

$$\begin{aligned} \int_{\hat{K}} \hat{x}_1^\ell \hat{x}_2^m d\hat{\mathbf{x}} &= \int_0^1 \hat{x}_1^\ell dx_1 \times \int_0^1 \hat{x}_2^m dx_2 \\ &= \sum_{p=1}^{r+1} \hat{\omega}_p \hat{\xi}_p^\ell \sum_{q=1}^{r+1} \hat{\omega}_q \hat{\xi}_q^m = \sum_{p=1}^{r+1} \sum_{q=1}^{r+1} \hat{\omega}_{p,q} \hat{\xi}_p^\ell \hat{\xi}_q^m. \end{aligned} \quad (12.8)$$

So, we can say that the 2D quadrature rule whose weights are $\hat{\omega}_{p,q}$ and whose point are $\hat{\xi}_{p,q}$, $1 \leq p \leq r+1$ and $1 \leq q \leq r+1$, is the 2D Gauss-Lobatto rule exact for Q_{2r-1} .

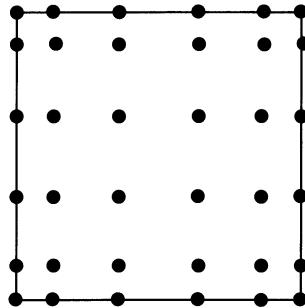


Fig. 12.1. The 36 Gauss-Lobatto points for Q_5 spectral elements in 2D

In 3D, we define, in the same way, on the unit cube $\hat{K} = [0, 1]^3$ ($\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, \hat{x}_3)$):

$$Q_r = \left\{ v(\hat{\mathbf{x}}) = \sum_{\ell=0}^r \sum_{m=0}^r \sum_{n=0}^r a_{\ell,m,n} \hat{x}_1^\ell \hat{x}_2^m \hat{x}_3^n, \ a_{\ell,m,n} \in \mathbb{R} \right\}, \quad (12.9)$$

$$\Xi_3 = \left\{ \hat{\xi}_{p,q,s} = (\hat{\xi}_p, \hat{\xi}_q, \hat{\xi}_s), p = 1..r+1, q = 1..r+1, s = 1..r+1 \right\}, \quad (12.10)$$

the basis functions

$$\hat{\varphi}_{\ell,m,n} = \hat{\varphi}_\ell \hat{\varphi}_m \hat{\varphi}_n \quad (12.11)$$

and the Gauss-Lobatto quadrature rule whose weights are $\hat{\omega}_{p,q,s} = \hat{\omega}_p \hat{\omega}_q \hat{\omega}_s$, whose points are $\hat{\xi}_{p,q,s}$, $1 \leq p \leq r+1$, $1 \leq q \leq r+1$, $1 \leq s \leq r+1$ and which is exact for Q_{2r-1} .

As in the 2D case, we have

$$\hat{\varphi}_{\ell,m,n}(\hat{\xi}_{p,q,s}) = \delta_{\ell p} \delta_{mq} \delta_{ns}. \quad (12.12)$$

12.1.2 Extension to Quadrilateral Meshes

The next step of our construction will be the generalization of these concepts to a quadrilateral mesh. We shall construct the method for the heterogeneous acoustics equation.

Let Ω be an open set of \mathbb{R}^2 . We want to solve the model problem:

$$\left\{ \begin{array}{l} \text{Find } u : \Omega \times [0, T] \rightarrow \mathbb{R} \text{ such that:} \\ \eta(\mathbf{x}) \frac{\partial^2 u}{\partial t^2}(\mathbf{x}, t) - \nabla \cdot (\gamma(\mathbf{x}) \nabla u(\mathbf{x}, t)) = f(\mathbf{x}, t), \text{ in } \Omega \times [0, T], \\ u(\mathbf{x}, 0) = u_0(\mathbf{x}), \quad \frac{\partial u}{\partial t}(\mathbf{x}, 0) = u_1(\mathbf{x}) \text{ in } \Omega, \quad u(\mathbf{x}, t) = 0 \text{ on } \partial\Omega. \end{array} \right. \quad (12.13)$$

by using a finite element method with mass-lumping on a mesh composed of quadrilaterals of any shape such as, for instance, that given in Fig. 12.2.

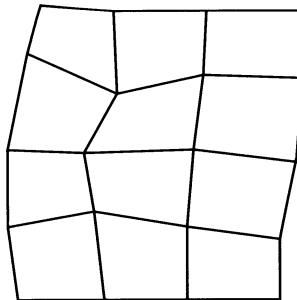


Fig. 12.2. A quadrilateral mesh in 2D

The variational formulation of (12.13) can be written as

$$\left\{ \begin{array}{l} \text{Find } u(., t) \in H_0^1(\Omega), \quad t \in [0, T] \quad \text{such that:} \\ \frac{d^2}{dt^2} \int_{\Omega} \eta u v \, d\mathbf{x} + \int_{\Omega} \gamma \nabla u \cdot \nabla v \, d\mathbf{x} = \int_{\Omega} f v \, d\mathbf{x}, \quad \forall v \in H_0^1(\Omega), \\ u(\mathbf{x}, 0) = u_0(\mathbf{x}), \quad \frac{\partial u}{\partial t}(\mathbf{x}, 0) = u_1(\mathbf{x}) \text{ in } \Omega. \end{array} \right. \quad (12.14)$$

Now, let

$$\mathcal{M}_h = \bigcup_{j=1}^{N_e} K_j \quad (12.15)$$

be a mesh of Ω composed of quadrilaterals (even with curved edges) K_j and $\mathbf{F}_j = (F_{j1}, F_{j2})$ the mapping such that $\mathbf{F}_j(\hat{K}) = K_j$. On this mesh, we define the following finite-dimensional subspace of $H_0^1(\Omega)$ ²:

$$U_h^r(\Omega) = \left\{ v_h \in H_0^1(\Omega) \text{ such that } v_h|_{K_j} \circ \mathbf{F}_j \in Q_r \right\}. \quad (12.16)$$

For quadrilaterals with straight edges, the mapping \mathbf{F}_j (Fig. 12.3) can be easily derived from the Q_1 basis functions on \hat{K} :

$$\hat{\varphi}_{1,1} = (1 - \hat{x}_1)(1 - \hat{x}_2), \quad (12.17a)$$

$$\hat{\varphi}_{2,1} = \hat{x}_1(1 - \hat{x}_2), \quad (12.17b)$$

$$\hat{\varphi}_{2,2} = \hat{x}_1\hat{x}_2, \quad (12.17c)$$

$$\hat{\varphi}_{1,2} = (1 - \hat{x}_1)\hat{x}_2. \quad (12.17d)$$

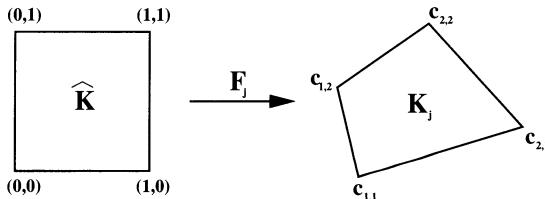


Fig. 12.3. The mapping \mathbf{F}_j

Let $\mathbf{c}_{1,1}, \mathbf{c}_{2,1}, \mathbf{c}_{2,2}, \mathbf{c}_{1,2}$, be the vertices of the quadrilateral K_j . The mapping \mathbf{F}_j can be written as

$$\mathbf{F}_j = \sum_{\ell=1}^2 \sum_{m=1}^2 \mathbf{c}_{\ell,m} \hat{\varphi}_{\ell,m}. \quad (12.18)$$

One can easily check that, although \mathbf{F}_j is a bilinear mapping, it is linear along the edges of \hat{K} , so that the four edges of \hat{K} are mapped onto the four straight edges of K_j . One could construct quadrilaterals with curved edges by using the basis functions of Q_r , $r > 1$, on \hat{K} or any other conform mapping. An element derived from the unit element by a non-linear mapping is called

² Which is also a subspace of C^0 .

an *isoparametric* element. The spectral element method holds for any isoparametric element, even once derived from \widehat{K} by a non-polynomial mapping.

On the basis of these definitions, we can now write the discrete variational formulation of (12.13):

$$\left\{ \begin{array}{l} \text{Find } u_h(., t) \in U_h^r(\Omega), t \in [0, T[\text{ such that:} \\ \frac{d^2}{dt^2} \int_{\Omega} \eta u_h v_h d\mathbf{x} + \int_{\Omega} \gamma \nabla u_h \cdot \nabla v_h d\mathbf{x} = \int_{\Omega} f v_h d\mathbf{x}, \\ \forall v_h \in U_h^r(\Omega), \\ u_h(\mathbf{x}, 0) = u_{0h}(\mathbf{x}), \quad \frac{\partial u_h}{\partial t}(\mathbf{x}, 0) = u_{1h}(\mathbf{x}) \text{ in } \Omega. \end{array} \right. \quad (12.19)$$

Of course, we have $U_h^r(\Omega) \subset C^0(\Omega)$.

The general framework being defined, we now show how to compute the different integrals involved in (12.19).

All the integrals of (12.19) will be computed by a change of variables.

Let us first define:

– The *Jacobian matrix*:

$$DF_j = \begin{pmatrix} \frac{\partial F_{j1}}{\partial \hat{x}_1} & \frac{\partial F_{j1}}{\partial \hat{x}_2} \\ \frac{\partial F_{j2}}{\partial \hat{x}_1} & \frac{\partial F_{j2}}{\partial \hat{x}_2} \end{pmatrix}. \quad (12.20)$$

– The *Jacobian*:

$$J_j = \det(DF_j). \quad (12.21)$$

The Mass Integral. With the above notation, we obtain, for the mass integral:

$$\begin{aligned} \int_{\Omega} \eta u_h v_h d\mathbf{x} &= \sum_{K_j \in \mathcal{M}_h} \int_{K_j} \eta u_h v_h d\mathbf{x} \\ &= \sum_{K_j \in \mathcal{M}_h} \int_{\widehat{K}} |J_j| \hat{\eta} u_h \circ \mathbf{F}_j v_h \circ \mathbf{F}_j d\hat{\mathbf{x}}, \end{aligned} \quad (12.22)$$

where $\hat{\eta} = \eta \circ \mathbf{F}_j$.

Now, since the functions v_h of $U_h^r(\Omega)$ are such that $v_h|_{K_j^\circ} \mathbf{F}_j \in Q_r$, one can say that, for each basis function φ_n of $U_h^r(\Omega)$, there exists a basis function $\hat{\varphi}_{\ell,m}$ on \hat{K} defined in (12.5) such that

$$\varphi_n|_{K_j^\circ} \mathbf{F}_j = \hat{\varphi}_{\ell,m}. \quad (12.23)$$

Let us set $n = \check{n}_j(\ell, m)$ the correlation defined in (12.23). Then, the restriction to K_j of the approximated solution u_h can be written as

$$u_h|_{K_j^\circ} \mathbf{F}_j = \sum_{\ell=0}^r \sum_{m=0}^r u_{\check{n}_j(\ell, m)} \varphi_{\check{n}_j(\ell, m)}|_{K_j^\circ} \mathbf{F}_j = \sum_{\ell=0}^r \sum_{m=0}^r u_{\check{n}_j(\ell, m)} \hat{\varphi}_{\ell, m}. \quad (12.24)$$

So, by setting $v_h = \varphi_{n_0}$, where n_0 is a given integer, we obtain from (12.22) :

$$\begin{aligned} \int_{\Omega} \eta u_h v_h \, d\mathbf{x} &= \\ \sum_{K_j \in \mathcal{S}} \sum_{\ell=0}^r \sum_{m=0}^r u_{\check{n}_j(\ell, m)} \int_{\hat{K}} |J_j| \hat{\eta} \hat{\varphi}_{\ell, m} \varphi_{n_0}|_{K_j^\circ} \mathbf{F}_j \, d\hat{\mathbf{x}}, & \end{aligned} \quad (12.25)$$

where

$$\mathcal{S} = \text{supp } \varphi_{n_0} = \left\{ K_j \in \mathcal{M}_h \text{ such that } \varphi_{n_0}|_{K_j^\circ} \neq 0 \right\}. \quad (12.26)$$

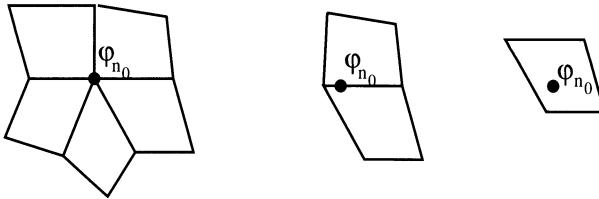


Fig. 12.4. The three possible configurations of \mathcal{S}

Now, let us set $\varphi_{n_0}|_{K_j^\circ} \mathbf{F}_j = \hat{\varphi}_{\ell_0, m_0}$ (i.e. $\check{n}_j(\ell_0, m_0) = n_0$). By applying the Gauss-Lobatto formula, one can write

$$\begin{aligned} \int_{\hat{K}} |J_j| \hat{\eta} \hat{\varphi}_{\ell, m} \hat{\varphi}_{\ell_0, m_0} \, d\hat{\mathbf{x}} &\simeq \\ \sum_{p=1}^{r+1} \sum_{q=1}^{r+1} \hat{\omega}_{p,q} |J_j(\hat{\xi}_{p,q})| \hat{\eta}(\hat{\xi}_{p,q}) \hat{\varphi}_{\ell, m}(\hat{\xi}_{p,q}) \hat{\varphi}_{\ell_0, m_0}(\hat{\xi}_{p,q}). & \end{aligned} \quad (12.27)$$

By using (12.6), we obtain:

$$\begin{aligned} \int_{\widehat{K}} |J_j| \hat{\eta} \hat{\varphi}_{\ell,m} \hat{\varphi}_{\ell_0,m_0} d\hat{\mathbf{x}} &\simeq \\ \sum_{p=1}^{r+1} \sum_{q=1}^{r+1} \hat{\omega}_{p,q} |J_j(\hat{\xi}_{p,q})| \hat{\eta}(\hat{\xi}_{p,q}) \delta_{\ell p} \delta_{mq} \delta_{\ell_0 p} \delta_{m_0 q}. \end{aligned} \quad (12.28)$$

Obviously, the right-hand side of (12.28) is not equal to zero if and only if $p = \ell = \ell_0$ and $q = m = m_0 =$, so that

$$\int_{\widehat{K}} |J_j| \hat{\eta} \hat{\varphi}_{\ell,m} \hat{\varphi}_{\ell_0,m_0} d\hat{\mathbf{x}} \simeq \hat{\omega}_{\ell_0,m_0} \frac{1}{J_j(\hat{\xi}_{\ell_0,m_0})} \hat{\eta}(\hat{\xi}_{\ell_0,m_0}). \quad (12.29)$$

The mass integral can then be written as

$$\int_{\Omega} \eta u_h v_h d\mathbf{x} \simeq \sum_{K_j \in \mathcal{S}} u_{\check{n}_{0j}(\ell_0, m_0)} \hat{\omega}_{\ell_0, m_0} |J_j(\hat{\xi}_{\ell_0, m_0})| \hat{\eta}(\hat{\xi}_{\ell_0, m_0}). \quad (12.30)$$

Now, since $\forall K_j \in \mathcal{S}$, $\check{n}_{0j}(\ell_0, m_0) = n_0$, we finally obtain:

$$\int_{\Omega} \eta u_h v_h d\mathbf{x} \simeq u_{n_0} \sum_{K_j \in \mathcal{S}} \hat{\omega}_{\ell_0, m_0} |J_j(\hat{\xi}_{\ell_0, m_0})| \hat{\eta}(\hat{\xi}_{\ell_0, m_0}). \quad (12.31)$$

For $v_h = \varphi_{n_0}$, the integrals on \widehat{K} in (12.25) are (parts of) the coefficients of the line of the mass matrix corresponding to φ_{n_0} . So, relation (12.31) shows that only one term of this line is different from zero and, thus, the mass matrix is diagonal.

The Stiffness Integral. The computation of the stiffness matrix is based on the following relation:

$$\nabla \varphi \circ \mathbf{F}_j = D\mathbf{F}_j^{*-1} \widehat{\nabla} \hat{\varphi}, \quad (12.32)$$

where φ is a function of $U_h^r(\Omega)$ such that $\varphi \circ \mathbf{F}_j = \hat{\varphi}$, $\nabla = (\partial/\partial x_1, \partial/\partial x_2)^T$, $\widehat{\nabla} = (\partial/\partial \hat{x}_1, \partial/\partial \hat{x}_2)^T$ and $D\mathbf{F}_j^{*-1}$ is the inverse of the transposed Jacobian matrix of \mathbf{F}_j .

By using (12.24) and (12.32) and setting, as for the mass integral $v_h = \varphi_{n_0}$, one can write the stiffness-matrix in the following form:

$$\begin{aligned} \int_{\Omega} \gamma \nabla u_h \cdot \nabla v_h \, dx &= \sum_{K_j \in \mathcal{M}_h} \int_{K_j} \gamma \nabla u_h \cdot \nabla v_h \, dx \\ &= \sum_{K_j \in \mathcal{S}} \sum_{\ell=0}^r \sum_{m=0}^r u_{\tilde{n}_j(\ell,m)} \int_{\hat{K}} |J_j| \hat{\gamma} D\hat{F}_j^{-1} D\hat{F}_j^{*-1} \hat{\nabla} \hat{\varphi}_{\ell,m} \cdot \hat{\nabla} \hat{\varphi}_{\ell_0,m_0} \, d\hat{x}, \end{aligned} \quad (12.33)$$

where $\hat{\gamma} = \gamma \circ \mathbf{F}_j$.

Then, the Gauss-Lobatto quadrature rule provides:

$$\begin{aligned} &\int_{\hat{K}} |J_j| \hat{\gamma} D\hat{F}_j^{-1} D\hat{F}_j^{*-1} \hat{\nabla} \hat{\varphi}_{\ell,m} \cdot \hat{\nabla} \hat{\varphi}_{\ell_0,m_0} \, d\hat{x} \simeq \\ &\sum_{p=1}^{r+1} \sum_{q=1}^{r+1} \hat{\omega}_{p,q} \hat{\gamma}(\hat{\xi}_{p,q}) |J_j(\hat{\xi}_{p,q})| \times \\ &D\hat{F}_j^{-1}(\hat{\xi}_{p,q}) D\hat{F}_j^{*-1}(\hat{\xi}_{p,q}) \hat{\nabla} \hat{\varphi}_{\ell,m}(\hat{\xi}_{p,q}) \cdot \hat{\nabla} \hat{\varphi}_{\ell_0,m_0}(\hat{\xi}_{p,q}). \end{aligned} \quad (12.34)$$

The stencils of the discrete operator obtained for the mass integral are rather large. In fact, since each quadrilateral has $(r+1)^2$ degrees of freedom, the stencil contains $(r+1)^2$ points for an interior degree of freedom, $(r+1)(2r+1)$ points for a degree of freedom on an edge and generally at least $(2r+1)^2$ points for a degree of freedom at a vertex of quadrilaterals (cf Fig. 12.4). One can, however, save both storage and computation time by applying a judicious algorithm described in [84]. We shall not give this algorithm here since it will appear in a mathematical form in Chap. 13.

In the case of regular meshes with a space-step h , the matrix $D\hat{F}_j^{-1} D\hat{F}_j^{*-1}$ is equal to the identity matrix multiplied by $1/h^2$ and J_j is equal to h^2 . So, the integral defined in (12.34) becomes:

$$\sum_{p=1}^{r+1} \sum_{q=1}^{r+1} \frac{\hat{\omega}_{p,q}}{h^2} \hat{\gamma}(\hat{\xi}_{p,q}) \hat{\nabla} \hat{\varphi}_{\ell,m}(\hat{\xi}_{p,q}) \cdot \hat{\nabla} \hat{\varphi}_{\ell_0,m_0}(\hat{\xi}_{p,q}). \quad (12.35)$$

Now, we have $\hat{\varphi}_{\ell,m}(\hat{\mathbf{x}}) = \hat{\varphi}_{\ell}(\hat{x}_1)\hat{\varphi}_m(\hat{x}_2)$ and $\hat{\varphi}_{\ell_0,m_0}(\hat{\mathbf{x}}) = \hat{\varphi}_{\ell_0}(\hat{x}_1)\hat{\varphi}_{m_0}(\hat{x}_2)$, such that

$$\begin{aligned} \hat{\nabla} \hat{\varphi}_{\ell,m}(\hat{\mathbf{x}}) \cdot \hat{\nabla} \hat{\varphi}_{\ell_0,m_0}(\hat{\mathbf{x}}) &= \frac{d\hat{\varphi}_{\ell}}{d\hat{x}_1}(\hat{x}_1) \frac{d\hat{\varphi}_{\ell_0}}{d\hat{x}_1}(\hat{x}_1) \hat{\varphi}_m(\hat{x}_2) \hat{\varphi}_{m_0}(\hat{x}_2) \\ &\quad + \frac{d\hat{\varphi}_m}{d\hat{x}_2}(\hat{x}_2) \frac{d\hat{\varphi}_{m_0}}{d\hat{x}_2}(\hat{x}_2) \hat{\varphi}_{\ell}(\hat{x}_1) \hat{\varphi}_{\ell_0}(\hat{x}_1). \end{aligned} \quad (12.36)$$

Obviously, one must have $\ell = \ell_0$ or $m = m_0$ to obtain a non-zero result. This condition is satisfied if and only if the straight line defined by the two degrees of freedom corresponding to $\hat{\varphi}_\ell$ and $\hat{\varphi}_{\ell_0}$ is parallel to the x_1 axis or the straight line defined by the two degrees of freedom corresponding to $\hat{\varphi}_m$ and $\hat{\varphi}_{m_0}$ is parallel to the x_2 axis. In other words, all the diagonal interactions of the basis functions are equal to zero. In Fig. 12.5, we represent the three kinds of stencils for Q_2 spectral elements. These stencils must be compared to that of the fourth-order approximation in finite difference given in Fig. 4.1.

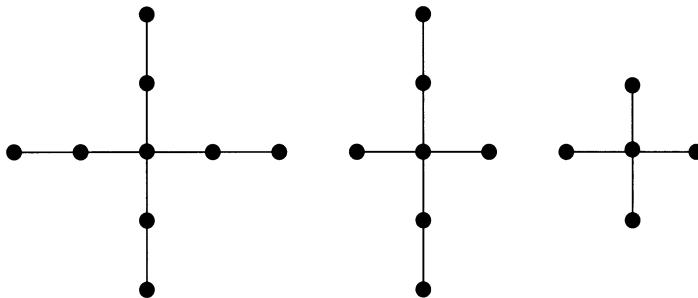


Fig. 12.5. The three stencils for Q_2 corresponding to a vertex (*left*) a point on an edge (*center*) and an interior point (*right*)

12.1.3 Extension to Hexahedral Meshes

Formulations (12.13) and (12.19) remain true for $\Omega \subset \mathbb{R}^3$ and $U_h^r(\Omega)$ is defined in the same way. The mapping \mathbf{F}_j is also derived from Q_1 -elements basis functions but, in this case, we have the following eight basis functions:

$$\hat{\varphi}_{1,1,1} = (1 - \hat{x}_1)(1 - \hat{x}_2)(1 - \hat{x}_3), \quad (12.37a)$$

$$\hat{\varphi}_{2,1,1} = \hat{x}_1(1 - \hat{x}_2)(1 - \hat{x}_3), \quad (12.37b)$$

$$\hat{\varphi}_{2,2,1} = \hat{x}_1\hat{x}_2(1 - \hat{x}_3), \quad (12.37c)$$

$$\hat{\varphi}_{1,2,1} = (1 - \hat{x}_1)\hat{x}_2(1 - \hat{x}_3), \quad (12.37d)$$

$$\hat{\varphi}_{1,1,2} = (1 - \hat{x}_1)(1 - \hat{x}_2)\hat{x}_3, \quad (12.37e)$$

$$\hat{\varphi}_{2,1,2} = \hat{x}_1(1 - \hat{x}_2)\hat{x}_3, \quad (12.37f)$$

$$\hat{\varphi}_{2,2,2} = \hat{x}_1\hat{x}_2\hat{x}_3, \quad (12.37g)$$

$$\hat{\varphi}_{1,2,2} = (1 - \hat{x}_1)\hat{x}_2\hat{x}_3. \quad (12.37h)$$

Now, if $\mathbf{c}_{1,1,1}, \mathbf{c}_{2,1,1}, \mathbf{c}_{2,2,1}, \mathbf{c}_{1,2,1}, \mathbf{c}_{1,1,2}, \mathbf{c}_{2,1,2}, \mathbf{c}_{2,2,2}, \mathbf{c}_{1,2,2}$, are the vertices of a hexahedron K_j , the mapping \mathbf{F}_j can be written in the same way as in the 2D case:

$$\mathbf{F}_j = \sum_{\ell=1}^2 \sum_{m=1}^2 \sum_{n=1}^2 \mathbf{c}_{\ell,m,n} \hat{\varphi}_{\ell,m,n}. \quad (12.38)$$

In this case, an edge of the unit cube is mapped to the segment whose ends are the vertices of the hexahedra corresponding to the ends of the edge but a face of the unit cube is not mapped onto a plane. It is mapped onto a surface defined by two parameters. For instance, the edge defined by $\hat{x}_1 = 0$, $\hat{x}_2 = 0$ and $0 \leq \hat{x}_3 \leq 1$ is mapped onto the curve defined by

$$x_p = c_{1,1,1}^{(p)}(1 - \hat{x}_3) + c_{1,1,2}^{(p)}\hat{x}_3, \quad p = 1..3, \quad (12.39)$$

where $\mathbf{c}_{\ell,m,n} = (c_{\ell,m,n}^{(1)}, c_{\ell,m,n}^{(2)}, c_{\ell,m,n}^{(3)})$.

Obviously, for $\hat{x}_3 \in [0, 1]$, (12.39) describes the segment whose ends are $\mathbf{c}_{1,1,1}$ and $\mathbf{c}_{1,1,2}$.

On the other hand, the face defined by $\hat{x}_1 = 0$, $0 \leq \hat{x}_2 \leq 1$ and $0 \leq \hat{x}_3 \leq 1$ is mapped onto the surface defined by

$$\begin{aligned} x_p &= c_{1,1,1}^{(p)}(1 - \hat{x}_2)(1 - \hat{x}_3) + c_{1,2,1}^{(p)}\hat{x}_2(1 - \hat{x}_3) \\ &\quad + c_{1,1,2}^{(p)}(1 - \hat{x}_2)\hat{x}_3 + c_{1,2,1}^{(p)}\hat{x}_2\hat{x}_3, \quad p = 1..3, \end{aligned} \quad (12.40)$$

which is not a plane when the four vertices of the face are not coplanar.

On the basis of this mapping, whose Jacobian matrix is

$$DF_j = \begin{pmatrix} \frac{\partial F_{j1}}{\partial \hat{x}_1} & \frac{\partial F_{j1}}{\partial \hat{x}_2} & \frac{\partial F_{j1}}{\partial \hat{x}_3} \\ \frac{\partial F_{j2}}{\partial \hat{x}_1} & \frac{\partial F_{j2}}{\partial \hat{x}_2} & \frac{\partial F_{j2}}{\partial \hat{x}_3} \\ \frac{\partial F_{j3}}{\partial \hat{x}_1} & \frac{\partial F_{j3}}{\partial \hat{x}_2} & \frac{\partial F_{j3}}{\partial \hat{x}_3} \end{pmatrix}, \quad (12.41)$$

the mass integral and the stiffness integral are computed in exactly the same way as in 2D and the mass matrix is diagonal here also. In practice, one obtains four kinds of stencils corresponding to a vertex, a point on an edge, a point on a face and an interior point.

The only point to describe in detail is the form of the stencils for regular meshes. Here, $\hat{\varphi}_{\ell,m,n}(\hat{\mathbf{x}}) = \hat{\varphi}_\ell(\hat{x}_1)\hat{\varphi}_m(\hat{x}_2)\hat{\varphi}_n(\hat{x}_3)$ and $\hat{\varphi}_{\ell_0,m_0,n_0}(\hat{\mathbf{x}}) = \hat{\varphi}_{\ell_0}(\hat{x}_1)\hat{\varphi}_{m_0}(\hat{x}_2)\hat{\varphi}_{n_0}(\hat{x}_3)$. So, the product of the gradients is of the form:

$$\begin{aligned}
& \widehat{\nabla} \hat{\varphi}_{\ell,m,n}(\hat{\mathbf{x}}) \cdot \widehat{\nabla} \hat{\varphi}_{\ell_0,m_0,n_0}(\hat{\mathbf{x}}) = \\
& \frac{d\hat{\varphi}_\ell}{d\hat{x}_1}(\hat{x}_1) \frac{d\hat{\varphi}_{\ell_0}}{d\hat{x}_1}(\hat{x}_1) \hat{\varphi}_m(\hat{x}_2) \hat{\varphi}_{m_0}(\hat{x}_2) \hat{\varphi}_n(\hat{x}_3) \hat{\varphi}_{n_0}(\hat{x}_3) \\
& + \frac{d\hat{\varphi}_m}{d\hat{x}_2}(\hat{x}_2) \frac{d\hat{\varphi}_{m_0}}{d\hat{x}_2}(\hat{x}_2) \hat{\varphi}_\ell(\hat{x}_1) \hat{\varphi}_{\ell_0}(\hat{x}_1) \hat{\varphi}_n(\hat{x}_3) \hat{\varphi}_{n_0}(\hat{x}_3), \\
& + \frac{d\hat{\varphi}_n}{d\hat{x}_2}(\hat{x}_3) \frac{d\hat{\varphi}_{n_0}}{d\hat{x}_2}(\hat{x}_3) \hat{\varphi}_\ell(\hat{x}_1) \hat{\varphi}_{\ell_0}(\hat{x}_1) \hat{\varphi}_m(\hat{x}_2) \hat{\varphi}_{m_0}(\hat{x}_2).
\end{aligned} \tag{12.42}$$

In this case, (12.42) is not equal to zero if and only if two of the indexes ℓ, m, n are equal to two of the indexes ℓ_0, m_0, n_0 . This only occurs when the degrees of freedom corresponding to the functions $\hat{\varphi}_{\ell,m,n}$ and $\hat{\varphi}_{\ell_0,m_0,n_0}$ are located on a line parallel to one of the axes. This means that here also, all the diagonal (even coplanar) interactions are equal to zero.

For 2D and 3D, the criterion given by [24] is still valid for rectangular elements and, therefore, the accuracy is maintained. The order of the method on a non-regular mesh remains an open problem which will be partially treated in the next section.

Of course, spectral elements can be applied to the 2D and 3D elastodynamics system without any change. For this system, one can see that both the geometry and the anisotropy are contained in the stiffness matrix of the system. We shall give a more detailed description of this system in Chap. 13.

12.2 Plane Wave Analysis of Regular Meshes

12.2.1 Decomposition of the Discrete Equations

The construction of 2D and 3D elements based on the 1D elements will enable us to derive, for regular meshes, the eigenvalues and eigenvectors in higher dimensions from those obtained in 1D in the plane wave analysis. This result is based on the following lemma³:

Lemma 5. *Let $N_{d,r}$ be the matrix defined in 11.16 and $N_{d,r}(p,q)$ the term located at the p th line and q th column of this matrix ($1 \leq p \leq \nu^d$ and $1 \leq q \leq \nu^d$, $d = 1..3$). Then, for a regular mesh composed of squares, we have*

³ The results of this section were given by N. Tordjman in her thesis [115] for Q_2 and Q_3 in 2D. Their generalization to any order and to the 3D case is a part of S. Fauqueux's thesis [55].

$$\begin{aligned} N_{2,r}((\ell_2 - 1)\nu + \ell_1, (m_2 - 1)\nu + m_1) = \\ N_{1,r}(\ell_1, m_1)\delta_{\ell_2, m_2} + N_{1,r}(\ell_2, m_2)\delta_{\ell_1, m_1}, \end{aligned} \quad (12.43)$$

$$\begin{aligned} N_{3,r}(((\ell_3 - 1)\nu + \ell_2 - 1)\nu + \ell_1, ((m_3 - 1)\nu + m_2 - 1)\nu + m_1) = \\ N_{1,r}(\ell_1, m_1)\delta_{\ell_2, m_2}\delta_{\ell_3, m_3} + N_{1,r}(\ell_2, m_2)\delta_{\ell_1, m_1}\delta_{\ell_3, m_3} \\ + N_{1,r}(\ell_3, m_3)\delta_{\ell_1, m_1}\delta_{\ell_2, m_2}, \end{aligned} \quad (12.44)$$

where $\delta_{\ell, m}$ is the Kronecker symbol.

Proof. We shall give the proof in details in the 2D case and then extend it to the 3D case. This proof is based on the fact that, any basis function $\varphi_{p,q}$ has a rectangular support $[\alpha_1, \alpha_2] \times [\beta_1, \beta_2]$ and we have:

$$\varphi_{p,q}(x_1, x_2) = \varphi_p(x_1)\varphi_q(x_2), \quad (12.45)$$

where φ_p and φ_q are two 1D basis functions so that $[\alpha_1, \alpha_2]$ is the support of φ_p and $[\beta_1, \beta_2]$ is that of φ_q .

On the other hand, the intersection of the supports of two 2D basis functions is always a rectangle that we shall denote $\mathcal{S} = [a_1, a_2] \times [b_1, b_2]$ in the following. With this notation, the terms of the mass matrix can be written as

$$\int_{\mathcal{S}} \varphi_{\ell_1, \ell_2} \varphi_{m_1, m_2} d\mathbf{x}. \quad (12.46)$$

Now, by using (12.45), we obtain:

$$\begin{aligned} \int_{\mathcal{S}} \varphi_{\ell_1, \ell_2} \varphi_{m_1, m_2} d\mathbf{x} = \\ \int_{a_1}^{a_2} \int_{b_1}^{b_2} \varphi_{\ell_1}(x_1)\varphi_{\ell_2}(x_2)\varphi_{m_1}(x_1)\varphi_{m_2}(x_2) dx_1 dx_2 = \\ \int_{a_1}^{a_2} \varphi_{\ell_1}(x_1) dx_1 \int_{b_1}^{b_2} \varphi_{\ell_2}(x_2)\varphi_{m_2}(x_2) dx_2. \end{aligned} \quad (12.47)$$

So, if $D_{d,r}(p, q)$ denotes the term located at the p th line and q th column of the diagonal mass matrix in dimension d obtained by computing the above integrals by a Gauss-Lobatto rule, we obtain the following relation:

$$\begin{aligned} D_{2,r}((\ell_2 - 1)\nu + \ell_1, (m_2 - 1)\nu + m_1) = \\ D_{1,r}(\ell_1, m_1) \times D_{1,r}(\ell_2, m_2). \end{aligned} \quad (12.48)$$

In the same way, the terms of the stiffness matrix are given by the integrals:

$$\int_S \nabla \varphi_{\ell_1, \ell_2} \cdot \nabla \varphi_{m_1, m_2} d\mathbf{x}. \quad (12.49)$$

Since

$$\nabla \varphi_{\ell_1, \ell_2}(x_1, x_2) = \begin{pmatrix} \frac{d\varphi_{\ell_1}}{dx_1}(x_1) \varphi_{\ell_2}(x_2) \\ \varphi_{\ell_1}(x_1) \frac{d\varphi_{\ell_2}}{dx_2}(x_2) \end{pmatrix}, \quad (12.50)$$

we obtain:

$$\begin{aligned} & \int_S \nabla \varphi_{\ell_1, \ell_2} \cdot \nabla \varphi_{m_1, m_2} d\mathbf{x} = \\ & \int_S \left[\frac{d\varphi_{\ell_1}}{dx_1}(x_1) \frac{d\varphi_{m_1}}{dx_1}(x_1) \varphi_{\ell_2}(x_2) \varphi_{m_2}(x_2) \right. \\ & \quad \left. + \varphi_{\ell_1}(x_1) \varphi_{m_1}(x_1) \frac{d\varphi_{\ell_2}}{dx_2}(x_2) \frac{d\varphi_{m_2}}{dx_2}(x_2) \right] dx_1 dx_2 = \\ & \int_{a_1}^{a_2} \frac{d\varphi_{\ell_1}}{dx_1}(x_1) \frac{d\varphi_{m_1}}{dx_1}(x_1) dx_1 \int_{b_1}^{b_2} \varphi_{\ell_2}(x_2) \varphi_{m_2}(x_2) dx_2 \\ & \quad + \int_{a_1}^{a_2} \varphi_{\ell_1}(x_1) \varphi_{m_1}(x_1) dx_1 \int_{b_1}^{b_2} \frac{d\varphi_{\ell_2}}{dx_2}(x_2) \frac{d\varphi_{m_2}}{dx_2}(x_2) dx_2. \end{aligned} \quad (12.51)$$

So, if we define $K_{d,r}(p, q)$ for the stiffness matrix as for the mass matrix, we obtain:

$$\begin{aligned} & K_{2,r}((\ell_2 - 1)\nu + \ell_1, (m_2 - 1)\nu + m_1) = \\ & K_{1,r}(\ell_1, m_1) D_{1,r}(\ell_2, m_2) + K_{1,r}(\ell_2, m_2) D_{1,r}(\ell_1, m_1). \end{aligned} \quad (12.52)$$

Since the mass matrices are all diagonal, we can write:

$$\begin{aligned} & N_{2,r}((\ell_2 - 1)\nu + \ell_1, (m_2 - 1)\nu + m_1) = \\ & D_{2,r}^{-1}((\ell_2 - 1)\nu + \ell_1, (\ell_2 - 1)\nu + \ell_1) \\ & \times K_{2,r}((\ell_2 - 1)\nu + \ell_1, (m_2 - 1)\nu + m_1) = \\ & \frac{1}{D_{1,r}(\ell_1, \ell_1)} \frac{1}{D_{1,r}(\ell_2, \ell_2)} [K_{1,r}(\ell_1, m_1) D_{1,r}(\ell_2, m_2) \\ & \quad + K_{1,r}(\ell_2, m_2) D_{1,r}(\ell_1, m_1)] = \\ & K_{1,r}(\ell_1, m_1) \frac{D_{1,r}(\ell_1, m_1)}{D_{1,r}(\ell_1, \ell_1)} + K_{1,r}(\ell_2, m_2) \frac{D_{1,r}(\ell_2, m_2)}{D_{1,r}(\ell_2, \ell_2)}. \end{aligned} \quad (12.53)$$

So, by taking into account the fact that $D_{1,r}$ is diagonal, we obtain (12.43).

In order to obtain (12.44), one proves, by a similar process, that

$$\begin{aligned} D_{3,r}(((\ell_3 - 1)\nu + \ell_2 - 1)\nu + \ell_1), ((m_3 - 1)\nu + m_2 - 1)\nu + m_1) = \\ D_{1,r}(\ell_1, m_1)D_{1,r}(\ell_2, m_2)D_{1,r}(\ell_3, m_3) \end{aligned} \quad (12.54)$$

and

$$\begin{aligned} K_{3,r}(((\ell_3 - 1)\nu + \ell_2 - 1)\nu + \ell_1), ((m_3 - 1)\nu + m_2 - 1)\nu + m_1) = \\ K_{1,r}(\ell_1, m_1)D_{1,r}(\ell_2, m_2)D_{1,r}(\ell_3, m_3) \\ + K_{1,r}(\ell_2, m_2)D_{1,r}(\ell_1, m_1)D_{1,r}(\ell_3, m_3) \\ + K_{1,r}(\ell_3, m_3)D_{1,r}(\ell_1, m_1)D_{1,r}(\ell_2, m_2). \end{aligned} \quad (12.55)$$

12.2.2 Decomposition of the Eigenvalues and Eigenvectors

Now, we search for a plane wave solution of the discrete wave equation on the regular mesh:

$$\frac{d^2}{dt^2} \mathbf{U} + N_{d,r} \mathbf{U} = 0. \quad (12.56)$$

For this purpose, we suppose that our regular mesh is unbounded in all directions and we replace, in the following, the one-dimensional indexes by d -dimensional indexes. For instance, $(\ell_2 - 1)\nu + \ell_1$ will be replaced by $(\ell_1, \ell_2) \in \mathbb{Z}^2$ and $((\ell_3 - 1)\nu + \ell_2 - 1)\nu + \ell_1$ by $(\ell_1, \ell_2, \ell_3) \in \mathbb{Z}^3$. Here, we have r^d degrees of freedom invariable by translation which define r^d classes. Each of these classes can be considered as the Cartesian product of r classes in 1D, as represented in Fig. 12.6.

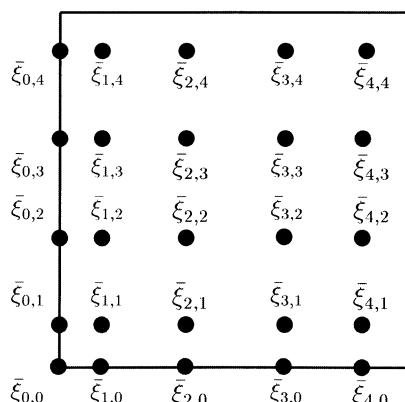


Fig. 12.6. The r^d classes of degrees of freedom in 2D (here, $r = 5$ and $d = 2$). The corresponding 1D classes are those located on the edges

With this notation, the dispersion relation is given by the following eigenvalue problem:

$$\widehat{N}_{2,r} \mathbf{U}_\alpha = \omega_h^2 \mathbf{U}_\alpha, \quad (12.57)$$

where, if $\mathcal{I}_d = \{0, \dots, r-1\}^d$, $\mathbf{U}_\alpha = (\bar{\alpha}_{\mathbf{p}})_{\mathbf{p} \in \mathcal{I}_d}$ with $\mathbf{p} = p$ in 1D, $\mathbf{p} = (p_1, p_2)$ in 2D and $\mathbf{p} = (p_1, p_2, p_3)$ in 3D.

In the following, we shall suppose that $d = 2$. The plane wave solution of (12.56) is defined as

$$\mathbf{U} = \left(\alpha_{p,q} e^{i(\omega_h t - k_1 \tilde{x}_p - k_2 \tilde{y}_q)} \right)_{(p,q) \in \mathbb{Z}^2}. \quad (12.58)$$

So, for $(p, q) = (\ell_1, \ell_2)$, after inserting (12.58) into (12.56), we obtain:

$$\begin{aligned} \omega_h^2 \alpha_{\ell_1, \ell_2} e^{i(\omega_h t - k_1 \tilde{x}_{\ell_1} - k_2 \tilde{y}_{\ell_2})} &= \\ \sum_{(m_1, m_2) \in \mathbb{Z}^2} N_{2,r}((\ell_1, \ell_2), (m_1, m_2)) \alpha_{m_1, m_2} e^{i(\omega_h t - k_1 \tilde{x}_{m_1} - k_2 \tilde{y}_{m_2})}. \end{aligned} \quad (12.59)$$

By taking into account the fact that we have r^2 degrees of freedom invariable by translation, one can write

$$\alpha_{\ell_1, \ell_2} = \alpha_{rp_1 + q_1, rp_2 + q_2} = \bar{\alpha}_{q_1, q_2}, \quad (12.60a)$$

$$\alpha_{m_1, m_2} = \alpha_{rp'_1 + q'_1, rp'_2 + q'_2} = \bar{\alpha}_{q'_1, q'_2}, \quad (12.60b)$$

where $1 \leq q_j \leq r$, $1 \leq q'_j \leq r$, $p_j \in \mathbb{Z}$, $p'_j \in \mathbb{Z}$, $j = 1, 2$. With this notation, we obtain from (12.59), after simplification:

$$\begin{aligned} \omega_h^2 \bar{\alpha}_{q_1, q_2} &= \sum_{(q'_1, q'_2) \in \mathcal{I}_2} \bar{\alpha}_{q'_1, q'_2} \sum_{(p'_1, p'_2) \in \mathbb{Z}^2} N_{2,r}((\ell_1, \ell_2), (rp'_1 + q'_1, rp'_2 + q'_2)) \\ &\quad \times e^{i(k_1(\tilde{x}_{rp'_1 + q'_1} - \tilde{x}_{\ell_1}) - k_2(\tilde{y}_{rp'_2 + q'_2} - \tilde{y}_{\ell_2}))} \\ &= \sum_{(q'_1, q'_2) \in \mathcal{I}_2} \widehat{N}_{2,r}[k_1, k_2]((q_1, q_2), (q'_1, q'_2)) \bar{\alpha}_{q'_1, q'_2}. \end{aligned} \quad (12.61)$$

By taking (12.43) into account, (12.59) can be rewritten as

$$\begin{aligned}
& \omega_h^2 \alpha_{\ell_1, \ell_2} e^{i(\omega_h t - k_1 \tilde{x}_{\ell_1} - k_2 \tilde{y}_{\ell_2})} = \\
& \sum_{(m_1, m_2) \in \mathbb{Z}^2} N_{1,r}(\ell_1, m_1) \delta_{\ell_2, m_2} \alpha_{m_1, m_2} e^{i(\omega_h t - k_1 \tilde{x}_{m_1} - k_2 \tilde{y}_{m_2})} \\
& + \sum_{(m_1, m_2) \in \mathbb{Z}^2} N_{1,r}(\ell_2, m_2) \delta_{\ell_1, m_1} \alpha_{m_1, m_2} e^{i(\omega_h t - k_1 \tilde{x}_{m_1} - k_2 \tilde{y}_{m_2})} = \\
& \sum_{m_1 \in \mathbb{Z}} N_{1,r}(\ell_1, m_1) \alpha_{m_1, \ell_2} e^{i(\omega_h t - k_1 \tilde{x}_{m_1} - k_2 \tilde{y}_{\ell_2})} \\
& + \sum_{m_2 \in \mathbb{Z}} N_{1,r}(\ell_2, m_2) \alpha_{\ell_1, m_2} e^{i(\omega_h t - k_1 \tilde{x}_{\ell_1} - k_2 \tilde{y}_{m_2})}, \tag{12.62}
\end{aligned}$$

which provides, after simplification:

$$\begin{aligned}
\omega_h^2 \alpha_{\ell_1, \ell_2} &= \sum_{m_1 \in \mathbb{Z}} N_{1,r}(\ell_1, m_1) \alpha_{m_1, \ell_2} e^{ik_1(\tilde{x}_{m_1} - \tilde{x}_{\ell_1})} \\
&+ \sum_{m_2 \in \mathbb{Z}} N_{1,r}(\ell_2, m_2) \alpha_{\ell_1, m_2} e^{ik_2(\tilde{y}_{m_2} - \tilde{y}_{\ell_2})}. \tag{12.63}
\end{aligned}$$

So, by writing

$$\alpha_{m_1, \ell_2} = \alpha_{rp'_1 + q'_1, rp_2 + q_2} = \bar{\alpha}_{q'_1, q_2}, \tag{12.64a}$$

$$\alpha_{\ell_1, m_2} = \alpha_{rp_1 + q_1, rp'_2 + q'_2} = \bar{\alpha}_{q_1, q'_2}, \tag{12.64b}$$

we obtain from (12.63):

$$\begin{aligned}
\omega_h^2 \bar{\alpha}_{q_1, q_2} &= \sum_{q'_1=0}^{r-1} \bar{\alpha}_{q'_1, q_2} \sum_{p'_1 \in \mathbb{Z}} N_{1,r}(\ell_1, rp'_1 + q'_1) e^{ik_1(\tilde{x}_{rp_1 + q_1} - \tilde{x}_{\ell_1})} \\
&+ \sum_{q'_2=0}^{r-1} \bar{\alpha}_{q_1, q'_2} \sum_{p'_2 \in \mathbb{Z}} N_{1,r}(\ell_2, rp'_2 + q'_2) e^{ik_2(\tilde{y}_{rp'_2 + q'_2} - \tilde{y}_{\ell_1})}, \tag{12.65}
\end{aligned}$$

which can be written as

$$\omega_h^2 \bar{\alpha}_{q_1, q_2} = \sum_{q'_1=0}^{r-1} \hat{N}_{1,r}[k_1](q'_1, q_2) \bar{\alpha}_{q'_1, q_2} + \sum_{q'_2=0}^{r-1} \hat{N}_{1,r}[k_2](q_1, q'_2) \bar{\alpha}_{q_1, q'_2}. \tag{12.66}$$

By comparing (12.61) and (12.66), we obtain the following relation between $\hat{N}_{2,r}[k_1, k_2]$, $\hat{N}_{1,r}[k_1]$ and $\hat{N}_{1,r}[k_2]$:

$$\begin{aligned}\widehat{N}_{2,r}[k_1, k_2]((q_1, q_2), (q'_1, q'_2)) &= \widehat{N}_{1,r}[k_1](q_1, q'_1)\delta_{q_2, q'_2} \\ &\quad + \widehat{N}_{1,r}[k_2](q_2, q'_2)\delta_{q_1, q'_1}.\end{aligned}\tag{12.67}$$

Let us recall the tensor product of two vectors \mathbf{X} and \mathbf{Y} of \mathbb{R}^r :

$$\mathbf{X} \otimes \mathbf{Y} = (X_1 Y_1, X_2 Y_1, \dots, X_r Y_1, X_1 Y_2, \dots, X_r Y_2, \dots, X_1 Y_r, \dots, X_r Y_r). \tag{12.68}$$

Now, let $\mathbf{U}_\lambda[k_1] = (U_1, \dots, U_r)^T$ and $\mathbf{U}_{\lambda'}[k_2] = (U'_1, \dots, U'_r)^T$ be two eigenvectors corresponding to two eigenvalues $\lambda[k_1]$ and $\lambda'[k_2]$ of $\widehat{N}_{1,r}[k_1]$ and $\widehat{N}_{1,r}[k_2]$. We have:

$$\begin{aligned}&\left(\widehat{N}_{2,r}[k_1, k_2] (\mathbf{U}_\lambda[k_1] \otimes \mathbf{U}_{\lambda'}[k_2]) \right)_{q_1, q_2} = \\ &\sum_{(q'_1, q'_2) \in \mathcal{I}_2} \widehat{N}_{2,r}[k_1, k_2]((q_1, q_2), (q'_1, q'_2)) U_{q'_1} U'_{q'_2} = \\ &\sum_{(q'_1, q'_2) \in \mathcal{I}_2} \widehat{N}_{1,r}[k_1](q_1, q'_1) \delta_{q_2, q'_2} U_{q'_1} U'_{q'_2} \\ &+ \sum_{(q'_1, q'_2) \in \mathcal{I}_2} \widehat{N}_{1,r}[k_2](q_2, q'_2) \delta_{q_1, q'_1} U_{q'_1} U'_{q'_2} = \\ &\sum_{q'_1 \in \mathcal{I}_1} \widehat{N}_{1,r}[k_1](q_1, q'_1) U_{q'_1} U'_{q_2} + \sum_{q'_2 \in \mathcal{I}_1} \widehat{N}_{1,r}[k_2](q_2, q'_2) U_{q_1} U'_{q'_2} = \\ &\lambda[k_1] U_{q_1} U'_{q_2} + \lambda'[k_2] U_{q_1} U'_{q_2} = \\ &(\lambda[k_1] + \lambda'[k_2]) U_{q_1} U'_{q_2}, \\ &\forall (q_1, q_2) \in \mathcal{I}_2.\end{aligned}\tag{12.69}$$

This shows that for any eigenvalue $\lambda[k_1]$ of $\widehat{N}_{1,r}[k_1]$ and for any eigenvalue $\lambda'[k_2]$ of $\widehat{N}_{1,r}[k_2]$, $\mathbf{U}_\lambda[k_1] \otimes \mathbf{U}_{\lambda'}[k_2]$ is the eigenvector of $\widehat{N}_{2,r}[k_1, k_2]$ associated to the eigenvalue $\lambda[k_1] + \lambda'[k_2]$. This important result can be summarized in

Theorem 4. Let $\{\lambda_j\}_{j=1}^r$ be the eigenvalues and $\{\mathbf{W}_j\}_{j=1}^r$ the eigenvectors of $\widehat{N}_{r,1}$, $\{T\lambda_j\}_{j=1}^{r^2}$ the eigenvalues and $\{\mathbf{TW}_j\}_{j=1}^{r^2}$ the eigenvectors of $\widehat{N}_{r,2}$ and ℓ an integer between 1 and r^2 such that $\ell = rp + q$, $0 \leq p \leq r - 1$, $1 \leq q \leq r$. Then, we have:

$$\begin{cases} T\lambda_\ell = \lambda_{p+1} + \lambda_q \\ T\mathbf{W}_\ell = \mathbf{W}_{p+1} \otimes \mathbf{W}_q \end{cases}$$

where $\mathbf{W}_{q+1} \otimes \mathbf{W}_r$ is defined as in (12.68).

In 3D, a similar process leads to the relation:

$$\begin{aligned} \widehat{N}_{2,r}[k_1, k_2, k_3]((q_1, q_2, q_3), (q'_1, q'_2, q'_3)) &= \\ \widehat{N}_{1,r}[k_1](q_1, q'_1)\delta_{q_2, q'_2}\delta_{q_3, q'_3} + \widehat{N}_{1,r}[k_2](q_2, q'_2)\delta_{q_1, q'_1}\delta_{q_3, q'_3} &\quad (12.70) \\ \widehat{N}_{1,r}[k_3](q_3, q'_3)\delta_{q_1, q'_1}\delta_{q_2, q'_2}. \end{aligned}$$

By denoting

$$\mathbf{X} \otimes \mathbf{Y} \otimes \mathbf{Z} = (X_1 Y_1 Z_1, \dots, X_r Y_1 Z_1, \dots, X_1 Y_r Z_r, \dots, X_r Y_r Z_r). \quad (12.71)$$

one shows, in the same way, that if $\mathbf{U}_\lambda[k_1]$ and $\mathbf{U}_{\lambda'}[k_2]$ and $\mathbf{U}_{\lambda''}[k_3]$ are three eigenvectors corresponding to three eigenvalues $\lambda[k_1]$, $\lambda'[k_2]$ and $\lambda''[k_3]$ of $\widehat{N}_{1,r}[k_1]$, $\widehat{N}_{1,r}[k_2]$ and $\widehat{N}_{1,r}[k_3]$, then, $\mathbf{U}_\lambda[k_1] \otimes \mathbf{U}_{\lambda'}[k_2] \otimes \mathbf{U}_{\lambda''}[k_3]$ is an eigenvector of $\widehat{N}_{2,r}[k_1, k_2, k_3]$ corresponding to the eigenvalue $\lambda[k_1] + \lambda'[k_2] + \lambda''[k_3]$.

These results imply, in particular, that the stability condition in dimension d is equal to the stability condition in 1D divided by \sqrt{d} .

On the other hand, the expressions of the eigenvectors and the eigenvalues in 2D enable us to obtain the errors committed on the amplitudes on a regular mesh [115].

Remarks

1. The infinite sums involved in (12.61) and (12.65) are actually finite.
2. In 2D, relation (12.67) can be written as

$$\widehat{N}_{2,r}[k_1, k_2] = \widehat{N}_{1,r}[k_1] \otimes Id_r + Id_r \otimes \widehat{N}_{1,r}[k_2], \quad (12.72)$$

where Id_r is the identity matrix of order r and, for any $r \times r$ matrix N :

$$N \otimes Id_r = \begin{pmatrix} N & 0 & \dots & 0 & 0 \\ 0 & N & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & N & 0 \\ 0 & 0 & \dots & 0 & N \end{pmatrix}.$$

$$Id_r \otimes N =$$

$$\left(\begin{array}{ccccccccc} N_{1,1} & \dots & 0 & N_{1,2} & \dots & 0 & & N_{1,r} & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \dots & \vdots & \ddots & \vdots \\ 0 & \dots & N_{1,1} & 0 & \dots & N_{1,2} & & 0 & \dots & N_{1,r} \\ \\ N_{2,1} & \dots & 0 & N_{2,2} & \dots & 0 & & N_{2,r} & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \dots & \vdots & \ddots & \vdots \\ 0 & \dots & N_{2,1} & 0 & \dots & N_{2,2} & & 0 & \dots & N_{2,r} \\ \\ \vdots & & & \vdots & & & \ddots & & & \vdots \\ \\ N_{r,1} & \dots & 0 & N_{r,2} & \dots & 0 & & N_{r,r} & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \dots & \vdots & \ddots & \vdots \\ 0 & \dots & N_{r,1} & 0 & \dots & N_{r,2} & & 0 & \dots & N_{r,r} \end{array} \right),$$

This notation provides a more straightforward proof of Theorem 4 by using the fact that $N \otimes Id_r(\mathbf{X} \otimes \mathbf{Y}) = (N\mathbf{X}) \otimes \mathbf{Y}$ and $Id_r \otimes N(\mathbf{X} \otimes \mathbf{Y}) = \mathbf{X} \otimes (N\mathbf{Y})$ but it is not easy to generalize these formulas to the 3D case.

12.3 Some Analysis of Non-Regular Meshes in 2D

In this section, we analyze, first by plane waves then numerically, the effect of the distortion of an element on the accuracy and the stability of the method. In order to control this distortion, the analysis will be made on periodic meshes.

12.3.1 Dispersion Analysis

A dispersion analysis is based on plane wave analysis which must be performed on a periodic infinite mesh. For this purpose, we define a periodic mesh of \mathbb{R}^2 composed of square cells of size $2h$ divided into four quadrilaterals. If $\mathcal{P} = \{P_i\}_{i=1..4}$ are the vertices of the square, $\mathcal{C} = \{C_i\}_{i=1..4}$, the midpoints of its edges and A an interior point, each quadrilateral has two vertices in \mathcal{C} , one vertex in \mathcal{P} and A as fourth vertex (Fig. 12.7).

For a Q_3 approximation, this periodic structure contains 36 classes of equations instead of 9 for a regular mesh (3 in 1D). As for a regular mesh in 1D, we substitute in these 36 equations a 36-dimensional vector valued plane

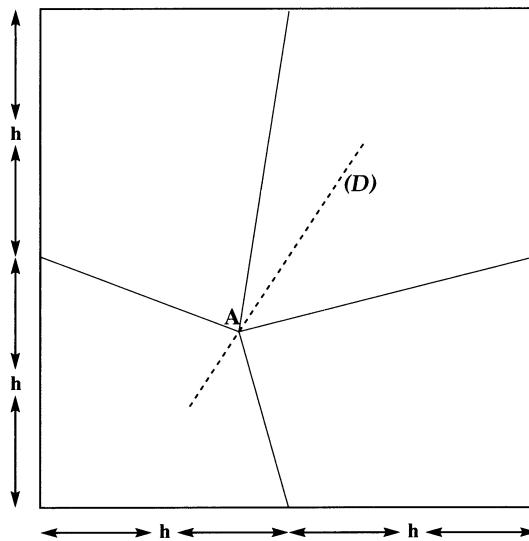


Fig. 12.7. The square cell. In the analysis and the experiments, $A = (ah, bh)$ is moved along the straight line (D) of equation $b = (3a - 1)/2$ ($0.6 \leq a \leq 1.4$)

wave solution which leads to a 36-dimensional eigenvalue problem. The matrix of this problem is constructed with the help of Maple and its eigenvalues are then computed (by a double precision FORTRAN program) numerically. Of course, we obtain one physical eigenvalue and 35 parasitic ones.

12.3.2 Numerical Study of the Stability

Next, we carry out a numerical study of the behavior of the schemes for distorted meshes. For this purpose, we construct a mesh based on the pattern given in Fig. 12.7 on a $]0, 12[\times]0, 12[$ bounded domain with homogeneous Dirichlet conditions on the boundary (Fig. 12.9). On this domain, we test Q_3 and Q_5 approximations. The source is modeled by the following right-hand side:

$$f(\mathbf{x}, t) = g_1(x_1, x_2)g_2(t), \quad (12.73)$$

where

$$g_1(x_1, x_2) = \sqrt{\frac{S(R)}{\pi}} e^{-S(R)[(x_1 - 6)^2 + (x_2 - 6)^2]}$$

$$g_2(t) = 2\gamma(2\gamma(t - \lambda_1)^2 - 1)e^{-\gamma(t - \lambda_1)^2}, \quad \gamma = \left(\frac{\pi}{\lambda_2}\right)^2, \quad \lambda_1 = 1.35, \quad \lambda_2 = 1.31.$$

and $S(R)$ is taken so that $g_1(R \cos \theta, R \sin \theta) = 10^{-6}$, $\forall \theta \in [0, 2\pi[$ (actually $S(R) = \log(10^{-6})/R^2$).

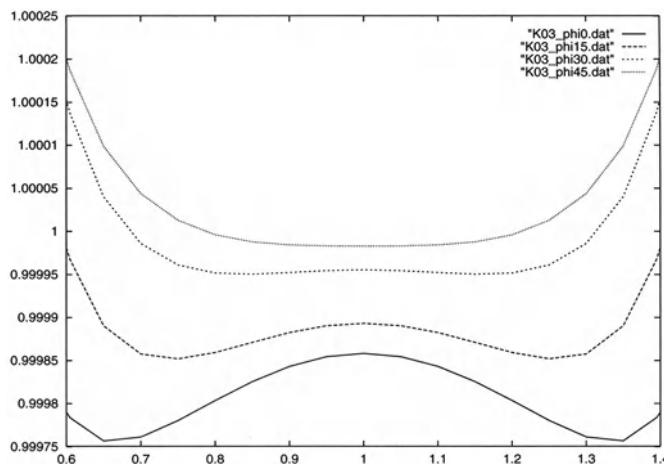


Fig. 12.8. Dispersion curves (c_h/c) versus a for some angles of propagation (0° (lower curve), 15° , 30° , 45° (upper curve)) and 3 elements per wavelength. $a = 1$ corresponds to a regular orthogonal mesh $a = 0.6$ or $a = 1.4$ correspond to degenerations into triangles

In Fig. 12.8, we give the dispersion curves, i.e. the ratio between the numerical velocity c_h of the physical wave and the exact velocity c . One can see that the loss of accuracy remains reasonable, even for significant distortions.

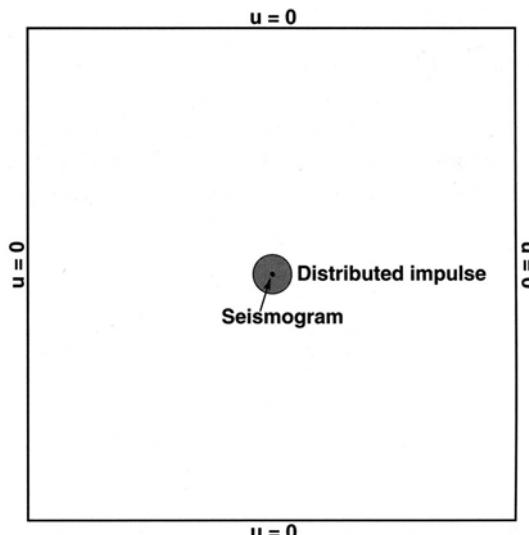


Fig. 12.9. The physical domain

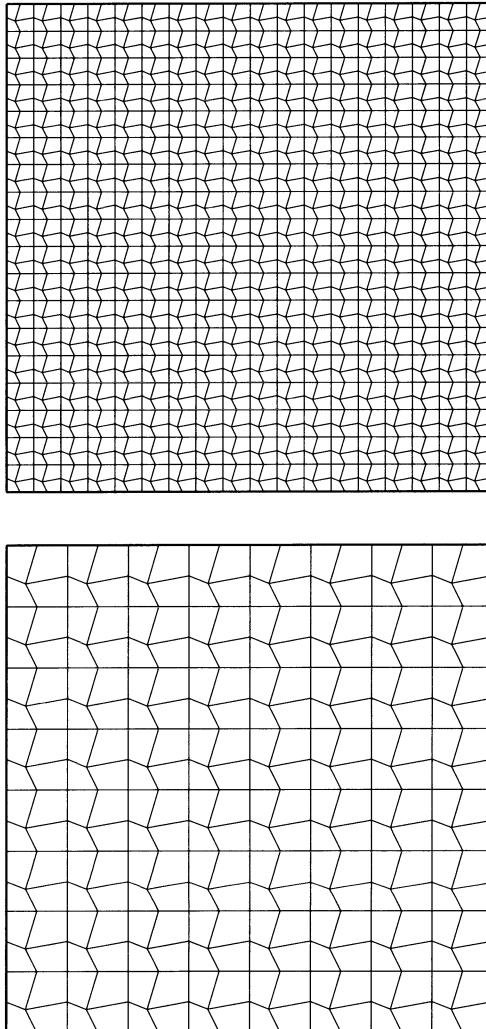


Fig. 12.10. Periodic meshes with $a = 0.7$ for Q_5 with 16×16 elements (*below*) and Q_3 with 36×36 elements (*above*)

In this section, we study the behavior of periodic meshes (Fig. 12.10) constructed by using cells defined as in Fig. 12.7. We move the point A along the straight line (D) of equation $b = (3a - 1)/2$ ($0.6 \leq a \leq 1$) and we draw the curves giving the stability condition (CFL) (i.e. the ratio between the time-step Δt and h/r where r represents the number of points per element which is, as stated in Sect. 11.4.2, more convenient to compare different orders of FEM and FDM) versus the parameter a (Fig. 12.11). In particular, this study shows that the method remains stable for significant distortions when one divides the CFL by 2. The use of such a CFL is a reasonable choice

to maintain global accuracy of the scheme. For regular meshes, the CFL is about 0.49 in Q_3 versus 0.57 for a sixth-order approximation in finite difference and about 0.35 in Q_5 versus 0.54 for a tenth-order approximation in finite difference.

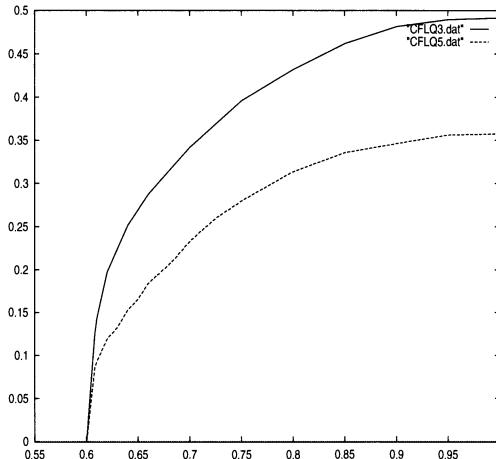


Fig. 12.11. CFL for a leapfrog scheme versus a for Q_3 and Q_5 approximations. The lower CFL curve is that of Q_5

12.3.3 Numerical Study of the Accuracy

The study of the accuracy of the solutions will be based on seismograms drawn between $t = 25$ and $t = 50$ (this interval gives a good idea of the behaviour of the solution for long times of integration) obtained for different meshes at the center of the domain (which is the most representative point for the solution) defined in Fig. 12.9.

Influence of the Approximation in Space. One of the features of the spectral element is to lead to “converged” solutions for a given number of points. However, as long as one does not reach this number of points, the solution contains some ripples due to the parasitic waves. This phenomenon is represented in Fig. 12.12 in which we give the seismograms of solutions computed with $R = 2$, by using a Q_5 approximation with 10×10 , 11×11 and 12×12 element regular meshes. In order to deal only with the approximation in space, we took a very small time-step ($\Delta t = 0.001$).

The comparison of these solutions with the solution obtained by using a 18×18 elements approximation with the same time-step (given as a reference

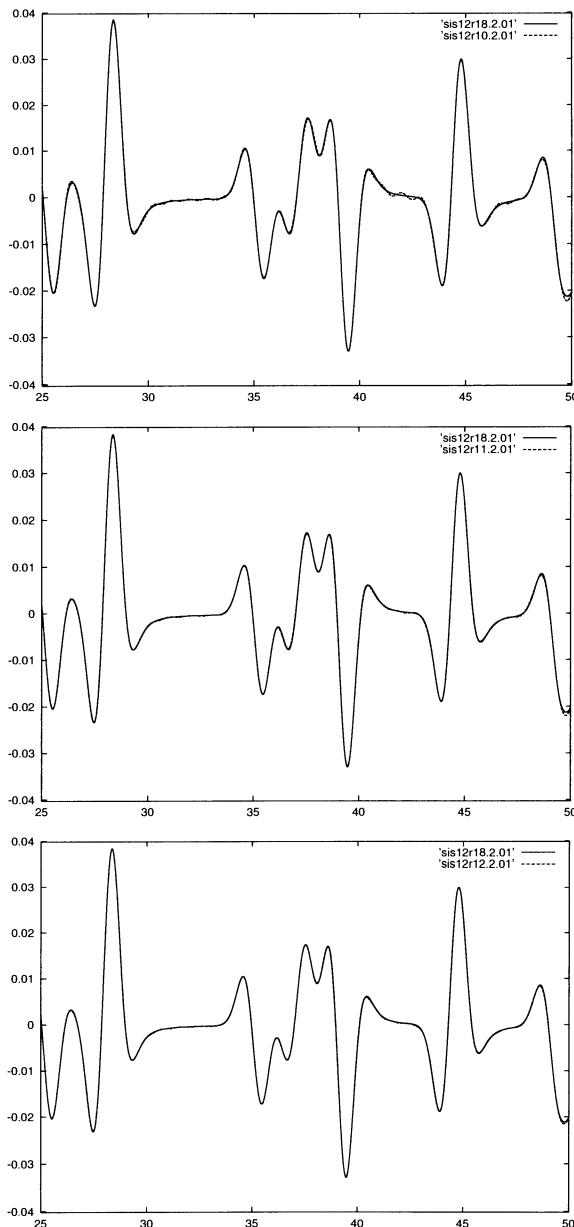


Fig. 12.12. Seismograms at the center of the domain on the time interval $[25, 50]$ for Q_5 with 10×10 elements (above), 11×11 elements (middle) and 12×12 elements (below) compared with a reference solution (dashed line). One can notice the progressive attenuation of the ripples

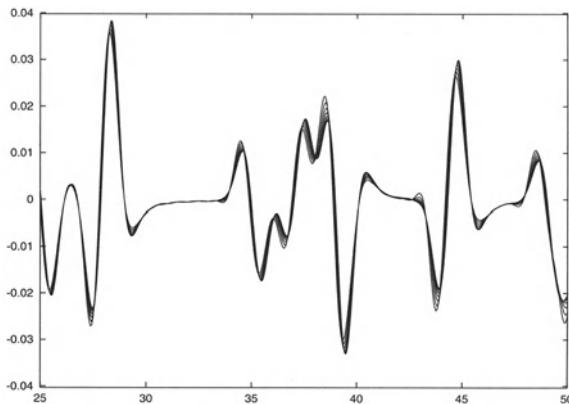


Fig. 12.13. Seismograms at the center of the domain on the time interval $[25, 50]$ for Q_5 with 12×12 elements obtained for time-steps varying from $\Delta t = 0.01$ to $\Delta t = 0.07$ in intervals of 0.01 and compared to the solution computed by using a time-step equal to 0.001

solution in Fig. 12.12) shows that the solution actually converges for a 12×12 element mesh.

Influence of the Approximation in Time. Now, for the converged solution defined in the previous subsection, we study the influence of the time-step on the accuracy. For this purpose, we solve the equation by using a 12×12 element mesh and a leapfrog scheme with different time-steps varying from $\Delta t = 0.01$ to $\Delta t = 0.07$ in intervals of 0.01 (the maximum time-step is $\Delta t = 0.0713$), and we compare in Fig. 12.13 the solutions obtained to that computed by using a time-step equal to 0.001. One can easily see the great influence of the time-step on the accuracy. In a second step, we compare the solution obtained for $\Delta t = 0.03$ to a solution computed with the same time-step for a converged solution in Q_5 (i.e. by using a 30×30 elements mesh (Fig. 12.14)). This study shows that the time-step has the same influence for any order of approximation. This result is not surprising since, as we saw in (11.79) and (11.80), the leading term in Δt remains the same in the dispersion relation for any order of approximation. From this study, we decide to take a time-step equal to 40% of the maximum time-step for Q_5 in the following.

Of course, the use of higher-order approximations in time could reduce this influence but, as we shall see in the last chapter, these approximation are very troublesome for modeling unbounded domains and, moreover, a leapfrog scheme is accurate enough for heterogeneous media.

Influence of the Support of the Source. Function g_1 is actually an approximation of the Dirac function which should be used as a point source. It is used in finite element approximations because a Dirac function would be

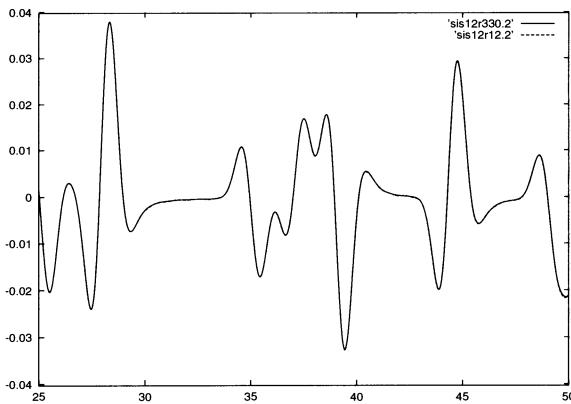


Fig. 12.14. Seismograms at the center of the domain on the time interval $[25, 50]$ for Q_5 with 12×12 elements and Q_3 with 30×30 elements. Both solutions were computed with $\Delta t = 0.03$

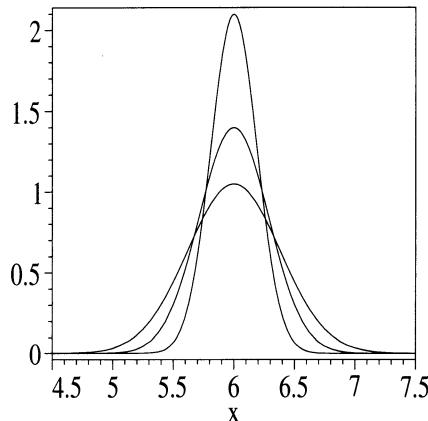


Fig. 12.15. Cross-sections of function g_1 when $R = 1$ (inner curve), $R = 1.5$ (intermediate curve) and $R = 2$ (outer curve)

too “tough” for the polynomial interpolation. This phenomenon can be illustrated by using different values of R . In this section, we study the influence of this parameter (i.e. of the spatial support of the source) on the accuracy of the solution. For this purpose, we compare a “minimal” converged solution in Q_5 to the solution obtained for the same CPU time⁴ in Q_3 and the “minimal” converged solution for the same approximation for $R = 1, 1.5$ and 2 (Figs. 12.17–12.19). Cross-sections (obtained for $x_2 = 6$) of function g_1 with the 3 values of R are represented in Fig. 12.15. In Fig. 12.16, we give the

⁴ For a DEC AlphaStation 500 MHz, 1 processor 21164 (500 MHz), 256 MB, 4.3 GB with an “optimized” option of the FORTRAN 90 compiler. The non-optimized option provides a slightly better performance for Q_3 (in 2D) [31].

3 reference seismograms for these values of R . One can easily see that the solution is reasonably affected by the change of support. It is interesting to note that the CPU times are exactly the same in Q_3 and Q_5 as soon as the numbers of degrees of freedom are the same. A Q_1 approximation for $R = 1$ uses a 300×300 element mesh and takes more than 100 s of CPU time to obtain an equivalent accuracy.

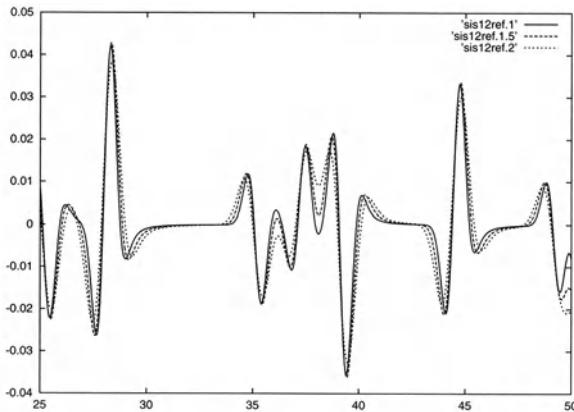


Fig. 12.16. Seismograms at the center of the domain on the time interval $]25, 50[$ for 3 converged Q_5 solutions when $R = 1$, $R = 1.5$ and $R = 2$

Influence of the Distortion of the Mesh. We now carry out a numerical study of the accuracy of the method for distorted meshes, as represented in Fig. 12.10. For this purpose, we compare two minimal Q_3 and Q_5 converged solutions on regular meshes for $R = 1.5$ and when $a = 1$, $a = 0.85$ and $a = 0.7$. This comparison of a 16×16 approximation in Q_5 with a 36×36 approximation in Q_3 on distorted meshes shows that the remarkable accuracy of Q_5 on a regular mesh is destroyed by the ripples which appear for large distortions (Fig. 12.20).

For $a = 0.7$, a 38×38 mesh in Q_3 and a 20×20 mesh in Q_5 are the minimal converged solutions. The CPU time for the Q_3 mesh is 24.7 s and for the Q_5 mesh it is 20.2 s. So, even for large distortions, Q_5 remains (slightly) better than Q_3 (Fig. 12.21). As we shall see in Chap. 13, this superiority should be significantly amplified in 3D.

12.3.4 A Two-Layer Experiment

In this section, we carry out a numerical study of a reflection-transmission process in 2D in order to check if the theoretical results found in 1D are confirmed in a 2D configuration.

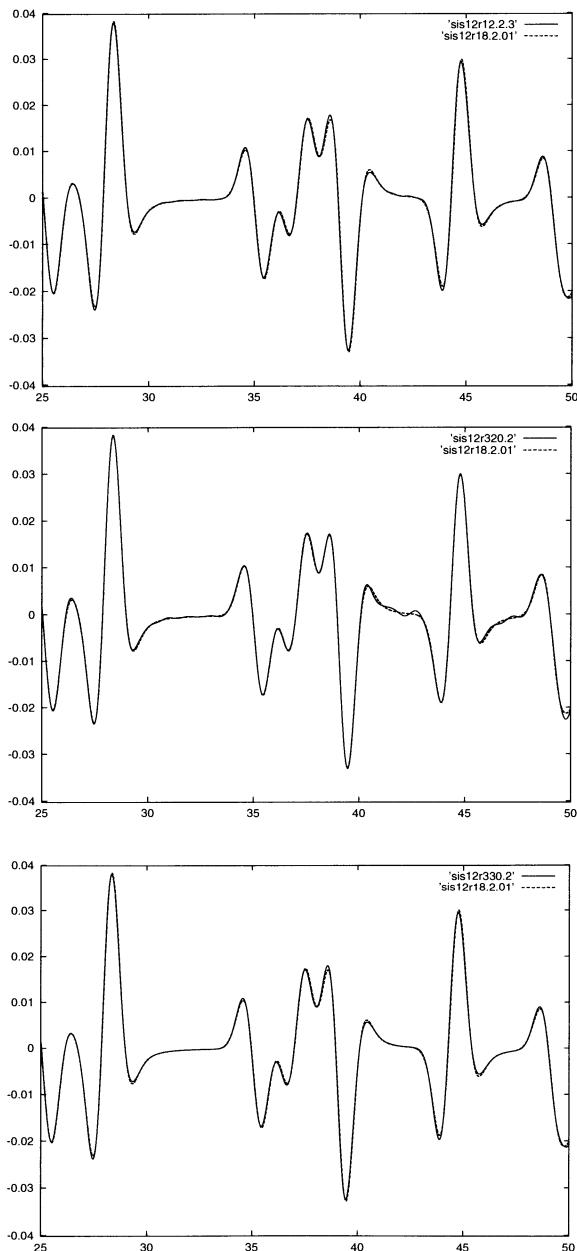


Fig. 12.17. Seismograms at the center of the domain on the time interval $]25, 50[$ for Q_5 with 12×12 elements (CPU: 6.2 s) (*above*), Q_3 with 20×20 elements (CPU: 6.2 s) (*middle*) and 30×30 elements (CPU: 12.2 s) (*below*). All the solutions were obtained with $\Delta t = 0.03$

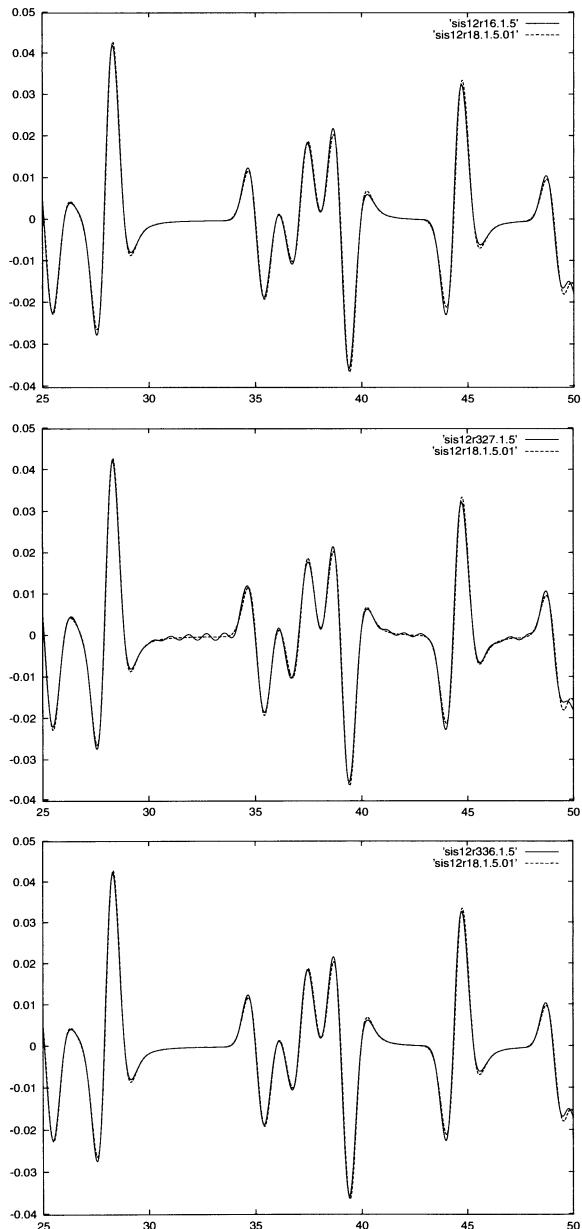


Fig. 12.18. Seismograms at the center of the domain on the time interval $]25, 50[$ for Q_5 with 16×16 elements (CPU: 11 s) (*above*), Q_3 with 27×27 elements (CPU: 11 s) (*middle*) and 36×36 elements (CPU: 19 s) (*below*). All the solutions were obtained with $\Delta t = 0.03$

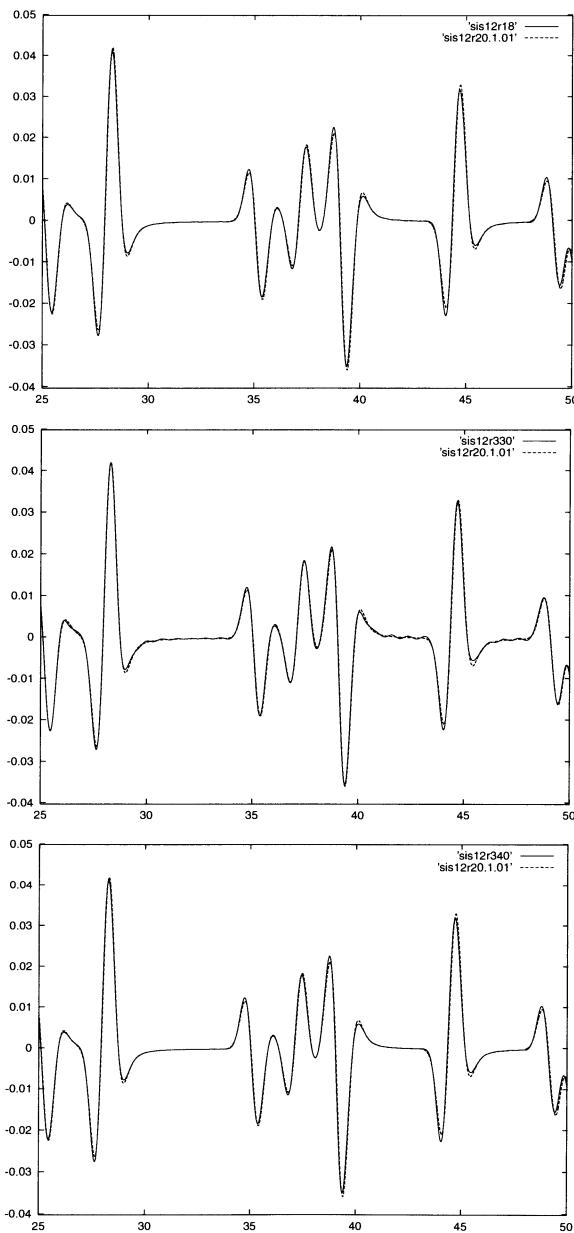


Fig. 12.19. Seismograms at the center of the domain on the time interval $[25, 50]$ for Q_5 with 18×18 elements (CPU: 15.8 s) (above), Q_3 with 30×30 elements (CPU: 15.8 s) (middle) and 40×40 elements (CPU: 23 s) (below). All the solutions were obtained with $\Delta t = 0.025$

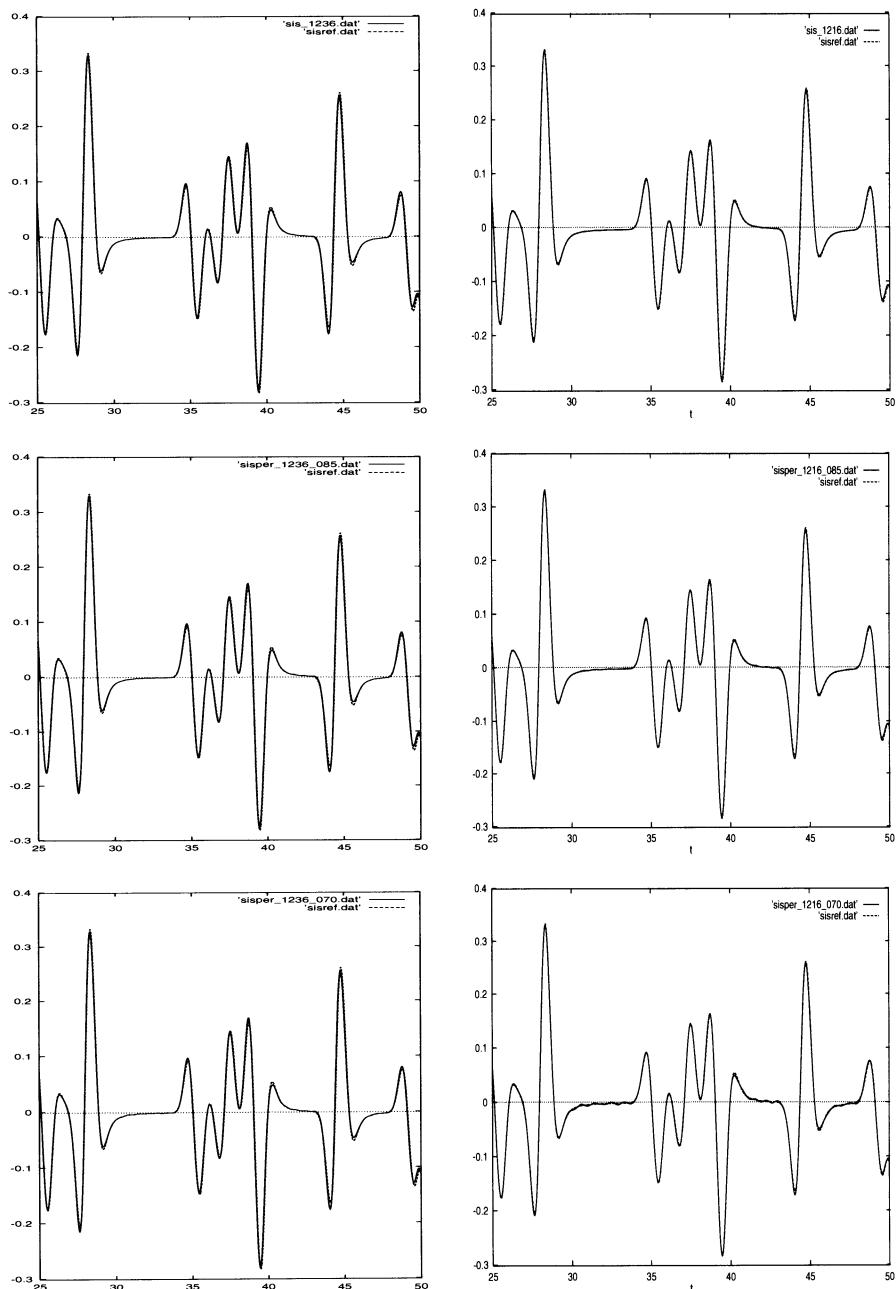


Fig. 12.20. Seismograms at the center of the domain on the time interval $[25, 50]$ for a Q_3 36×36 mesh (left) and a Q_5 16×16 mesh (right) when $a = 1$ (above), $a = 0.85$ (middle) and $a = 0.70$ (below) compared with the reference solution (dashed line). One can notice some ripples (more important in Q_5 than in Q_3) when $a = 0.70$.

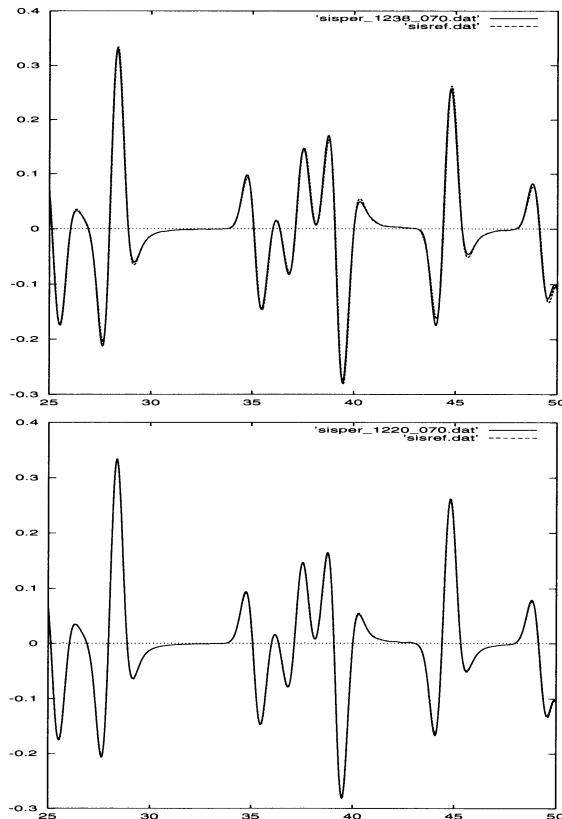


Fig. 12.21. Seismograms at the center of the domain on the time interval $]25, 50[$ for a Q_3 38×38 mesh (*above*) and a Q_5 20×20 mesh (*below*) when $a = 0.70$. The CPU time for Q_3 is 24.2 s and 20.2 s for Q_5

Let us consider a quasi-rectangular domain $\Omega \subset \mathbb{R}^2$ of dimensions 4200 m \times 3000 m whose upper boundary is a curve and such that $\Omega = \Omega_1 \cup \Omega_2$. Moreover, $\bar{\Omega}_1 \cap \bar{\Omega}_2 = D$, where D is a segment with a negative slope (Fig. 12.22). We solve the wave equation

$$\frac{1}{c^2(\mathbf{x})} \frac{\partial^2 u}{\partial t^2}(\mathbf{x}, t) - \Delta u(\mathbf{x}, t) = f(\mathbf{x}, t) \text{ in } \Omega. \quad (12.74)$$

We have a homogeneous Dirichlet condition on the upper boundary and the three other boundaries are open. We set $c = c_1$ in Ω_1 and $c = c_2$ in Ω_2 . In this domain, we propagate a pulse with a frequency of 17 Hz located at the middle of the upper boundary. The domain is meshed in two ways:

- By a regular mesh \mathcal{M}_1 (Fig. 12.22)
- By a mesh \mathcal{M}_2 which follows the interface between Ω_1 and Ω_2 (Fig. 12.22).

In the first experiment, we set $c_1 = c_2 = 3500$ m/s and we draw the seismograms (at the midpoint in the x direction and 100 m below the upper boundary) obtained for the two meshes on the same figure. One can notice the perfect fitting of the two curves (Fig. 12.23) which proves that the solution is not at all affected by the mesh (the number of elements is taken in order to obtain an accurate solution).

In a second experiment, we set $c_1 = 3500$ m/s in the higher part of the domain and $c_2 = 5500$ m/s in the lower part. The seismogram in Fig. 12.23 shows that the two curves do not agree as soon as the wave crosses the interface. This indicates the loss of accuracy predicted in Sect. 11.6 when the discretization in space does not follow the interface between two media. This loss of accuracy is confirmed by the use of refined regular and adapted meshes in which each quadrilateral is divided by 4. In Fig. 12.24, we see that the refinement of mesh provides the same solution for the adapted mesh whereas the solution provided by the refined regular mesh is different from that given by the coarse mesh but closer to that given by the adapted mesh.

12.4 Triangular and Tetrahedral Meshes

12.4.1 The Basic Problem

For triangles, the space of approximation can be directly defined as

$$W_h^r(\Omega) = \left\{ v_h \in H^1(\Omega) \text{ such that } v_h|_{K_j} \in P_r \right\}. \quad (12.75)$$

where

$$P_r = \left\{ v(\hat{\mathbf{x}}) = \sum_{(\ell,m) \in \mathcal{G}_r} a_{\ell,m} \hat{x}_1^\ell \hat{x}_2^m, a_{\ell,m} \in \mathbb{R} \right\}, \quad (12.76)$$

with $\mathcal{G}_r = \{(\ell, m) \in \mathbb{N}^2 \text{ such that } \ell + m \leq r\}$.

An immediate computation shows that $\text{card}(\mathcal{G}_r) = C_{r+2}^r = (r+1)(r+2)/2$, such that $\dim(P_r) = (r+1)(r+2)/2$. In particular, $\dim(P_2) = 6$. So, if

$$\mathcal{T} = \bigcup_{j=1}^{N_e} T_j \quad (12.77)$$

is a mesh composed of triangles, the degrees of freedom on a triangle T_j are the values of a function of $W_h^2(\Omega)$ at the three vertices and at the midpoints of the three edges of the triangle T_j . These points are the interpolation points of this function.

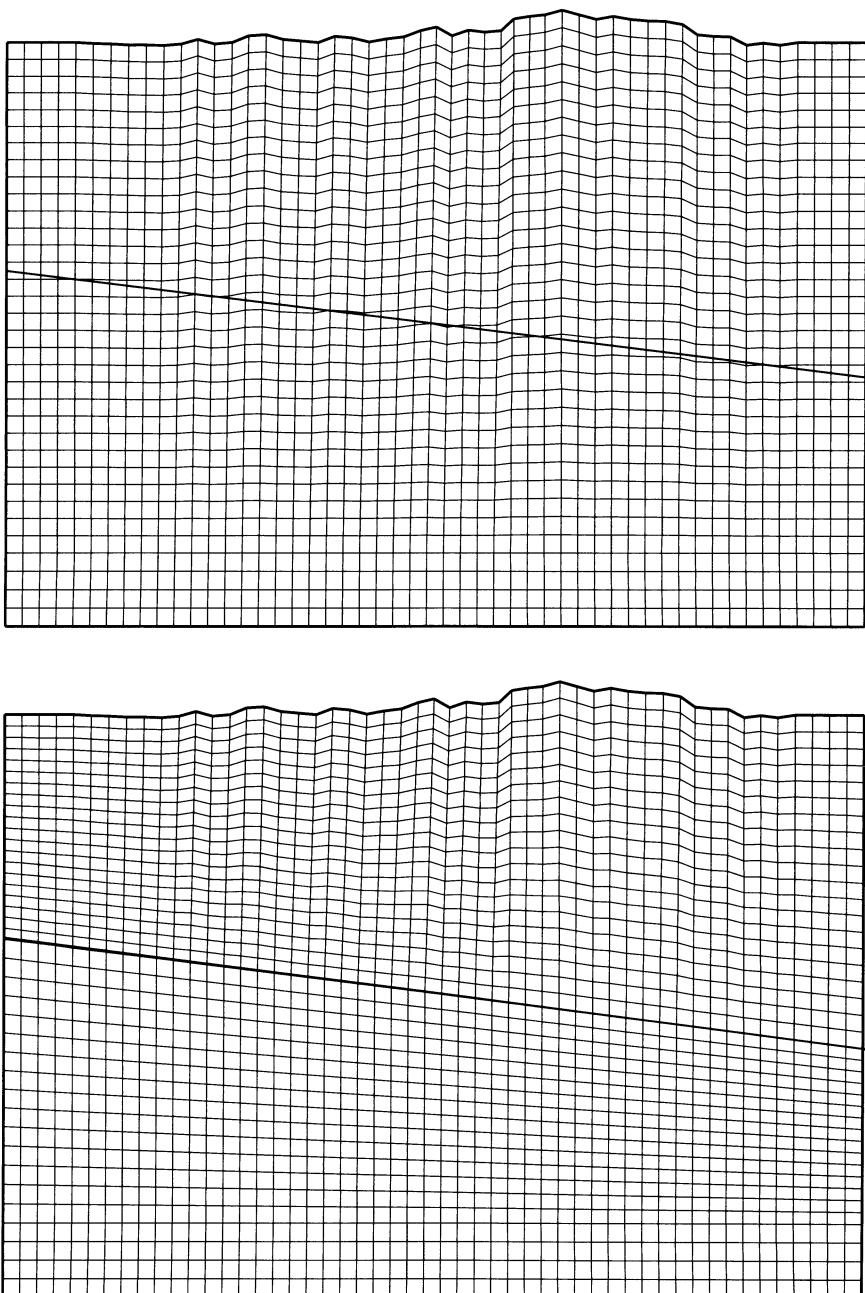


Fig. 12.22. The meshes \mathcal{M}_1 (above) and \mathcal{M}_2 (below). The interior *bold line* marks the interface

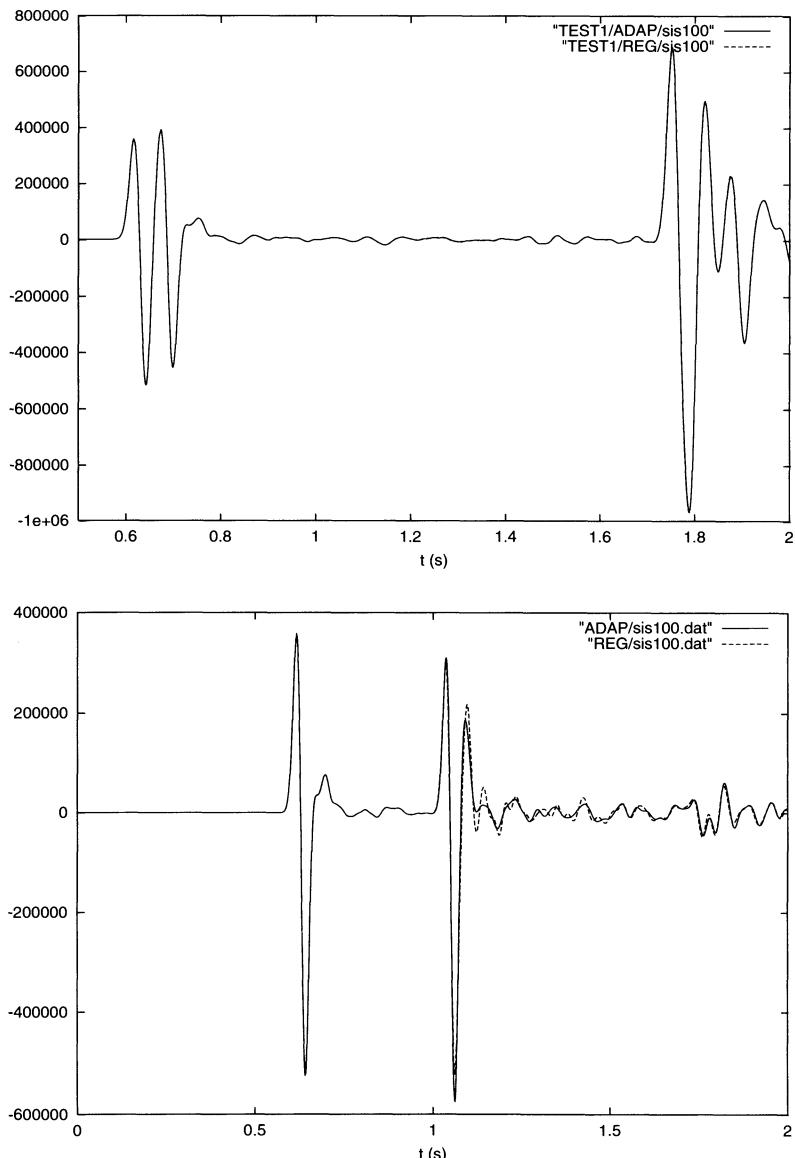


Fig. 12.23. The seismogram at the point of abscissa 2100 m, 100 m below the upper boundary obtained for the two meshes when $c_1 = c_2 = 3500$ m/s (*above*) and when $c_1 = 3500$ m/s in the higher part of the domain and $c_2 = 5500$ m/s in the lower part (*below*)

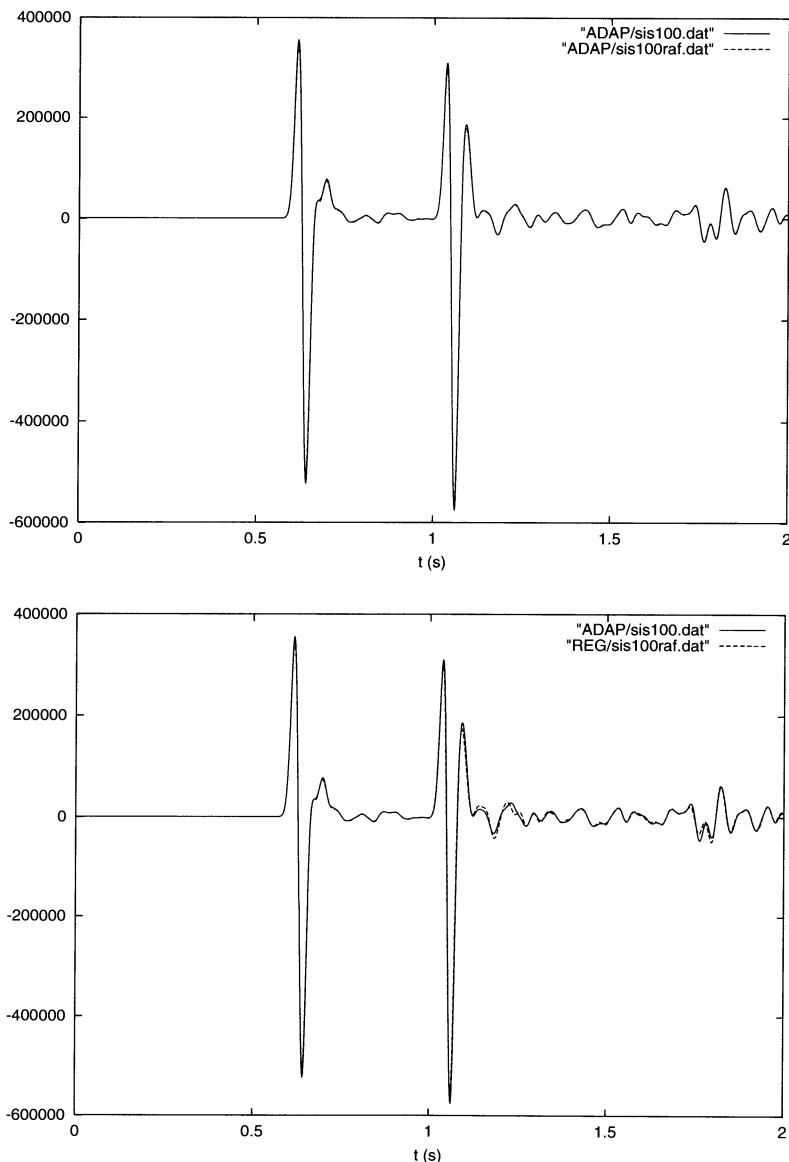


Fig. 12.24. The seismogram at the point of abscissa 2100 m, 100 m below the upper boundary obtained for the refined adapted mesh (*above*) and the refined regular mesh (*below*) when $c_1 = 3500$ m/s in the higher part of the domain and $c_2 = 5500$ m/s in the lower part. The two seismograms are compared to the seismogram obtained by the coarse adapted mesh

Now, our purpose is to obtain mass-lumping on this triangular mesh. Of course, as for tetrahedra, we must obtain a quadrature formula whose points coincide with the interpolation points and for which we are sure to keep the accuracy of the method. Following [24], this second condition is satisfied as soon as the quadrature rule is exact for $P_{2(r-1)}$. This means that, for a P_2 approximation, the quadrature rule must be exact for P_2 .

For symmetry reasons, the vertices and midpoints of the edges will be the quadrature points for P_2 (Fig. 12.25). So, we only have to determine the weights of the quadrature rule. Let us do this for the unit triangle \hat{T} whose vertices are $S_1(0, 0)$, $S_2(1, 0)$, $S_3(0, 1)$ and whose midpoints of the edges are $M_1(1/2, 0)$, $M_2(1/2, 1/2)$, $M_3(0, 1/2)$. All the vertices have the same weight ω_S and the midpoints are all affected of the weight ω_M . Since this quadrature rule must be exact for P_2 , these weights must satisfy the equations⁵:

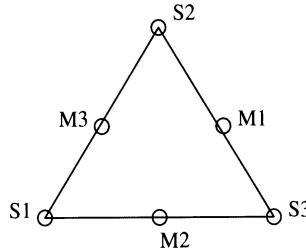


Fig. 12.25. The P_2 finite element

$$3\omega_S + 3\omega_M = \int_{\hat{T}} d\hat{\mathbf{x}} = \frac{1}{2}, \quad (12.78a)$$

$$\omega_S + \frac{\omega_M}{2} = \int_{\hat{T}} \hat{x}_1^2 d\hat{\mathbf{x}} = \frac{1}{12}. \quad (12.78b)$$

From (12.78a) and (12.78b), we obtain:

$$\omega_S = 0, \quad \omega_M = \frac{1}{6}. \quad (12.79)$$

Equation (12.79) would provide a diagonal matrix with zeroes on its diagonal, which is, of course, not invertible.

In P_3 , the quadrature points must be located at (Fig. 12.26)

- The vertices S_1, S_2, S_3 with the weight ω_S .

⁵ We choose the polynomials 1 and \hat{x}_1^2 because the equation derived from \hat{x}_1 is the same as that derived from 1.

- The center G of the triangle with the weight ω_G .
- Six points $M_{12}(\theta)$, $M_{21}(\theta)$, $M_{13}(\theta)$, $M_{31}(\theta)$, $M_{23}(\theta)$, $M_{32}(\theta)$ such that $M_{\ell m}(\theta)$ is the barycenter of S_ℓ and S_m with the weights θ and $1 - \theta$ (classically, $\theta = 1/3$). These points are affected of the weight ω_M .

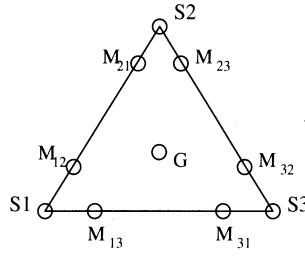


Fig. 12.26. The P_3 finite element

In this case, the quadrature rule must be exact for P_4 . So, we can write the four following relations:

$$\begin{aligned} \frac{1}{3}\omega_S \theta(\theta - 1) &= \int_{\hat{T}} (\hat{x}_1 - \theta)(\hat{x}_1 - 1 + \theta)\hat{x}_1(\hat{x}_2 - \frac{1}{3})d\hat{\mathbf{x}} \\ &= \frac{1}{360}(5\theta^2 - 5\theta + 1), \end{aligned} \quad (12.80a)$$

$$\begin{aligned} \frac{1}{81}\omega_G (1 - 3\theta)(3\theta - 2) &= \int_{\hat{T}} (\hat{x}_1 - \theta)(\hat{x}_1 - 1 + \theta)\hat{x}_1\hat{x}_2 d\hat{\mathbf{x}} \\ &= \frac{1}{120}(-5\theta^2 + 5\theta - 1), \end{aligned} \quad (12.80b)$$

$$\begin{aligned} \omega_M \theta(2\theta^3 - 4\theta^2 + 3\theta - 1) &= \int_{\hat{T}} \hat{x}_1(\hat{x}_1 - 1)(\hat{x}_2 - 1)(\hat{x}_2 - \frac{1}{3})d\hat{\mathbf{x}} \\ &= -\frac{1}{180}, \end{aligned} \quad (12.80c)$$

$$3\omega_S + \omega_G + 6\omega_M = \int_{\hat{T}} d\hat{\mathbf{x}} = \frac{1}{2}. \quad (12.80d)$$

ω_S , ω_G and ω_M come from (12.80a)–(12.80c). By inserting them into (12.80d), we obtain, after some computations:

$$36\theta^4 - 72\theta^3 + 30\theta^2 + 6\theta - 2 = 0. \quad (12.81)$$

Since θ and $1 - \theta$ play the same role in the quadrature rule, we can replace, in (12.81), θ by $1/2 - \zeta$, so that (12.81) becomes:

$$144\zeta^4 - 96\zeta^2 + 7 = 0. \quad (12.82)$$

By solving this equation, we obtain the following four values of θ :

$$\theta_1 = \frac{3 - \sqrt{3}}{6}, \theta_2 = \frac{3 + \sqrt{3}}{6}, \theta_3 = \frac{3 - \sqrt{21}}{6}, \theta_4 = \frac{3 + \sqrt{21}}{6}. \quad (12.83)$$

Since θ_3 and θ_4 are not in the interval $[0, 1]$, which would mean that the points $M_{\ell m}(\theta)$ would be outside of the triangle, only θ_1 and θ_2 are acceptable (symmetric) values. Now, by setting, $\theta = \theta_1$ for instance in (12.80a)–(12.80c), we obtain:

$$\omega_S = -\frac{1}{120}, \omega_G = -\frac{9}{40}, \omega_M = \frac{1}{20}. \quad (12.84)$$

The negative value of ω_S is very troublesome because an approximation constructed by using such a quadrature rule is unconditionally unstable. Similar computations show that we have the same problem with higher-order approximations.

Remarks

1. The fact that a negative weight provides unstable schemes arises from the fact that the eigenvalue problem giving the dispersion relations of the problem has, in this case, negative eigenvalues which lead to exponentially growing solutions.
2. The quadrature formula for P_3 was first derived by Hughes [71] who proposes a way to ensure the positivity of the weights which does not maintain the accuracy of the quadrature rule.

12.4.2 A New Family of Triangular Elements

In order to construct quadrature rules with positive weights, we define new spaces of triangular finite elements. The idea is to enrich the basic space of approximation with some functions which are not in the space [37, 115]. Basically, we want a space of approximation derived from a polynomial space \tilde{P}_r such that

$$P_r \subset \tilde{P}_r \subset P_{r'}, \quad r < r'. \quad (12.85)$$

To maintain the accuracy of the approximation, it is sufficient to have, in this case, a quadrature formula exact for $P_{r+r'-2}$ [24].

A triangular \tilde{P}_2 element method can be constructed as follows: let $(\lambda_1, \lambda_2, \lambda_3)$ be the three barycentric coordinates of a point based on the three vertices S_1, S_2, S_3 of a triangle T . We define

$$\tilde{P}_2 = P_2 \oplus [b], \quad (12.86)$$

where $[b]$ is the space generated by the “bubble function”:

$$b = \lambda_1 \lambda_2 \lambda_3, \quad (12.87)$$

which vanishes on the edges of T .

We have $P_2 \subset \tilde{P}_2 \subset P_3$ and $\dim(\tilde{P}_2) = 7$. The degrees of freedom corresponding to this space are located at the three vertices of the triangle, the midpoints of its edges and its center (Fig. 12.27).

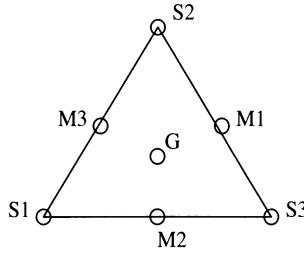


Fig. 12.27. The \tilde{P}_2 element

Now, we must find a quadrature rule, exact for P_3 , based on these points whose weights are (we hope) positive. We shall call ω_S , ω_M and ω_G the weights corresponding to the vertices, the midpoints and the center of the triangle. On the unit triangle \hat{T} , we can write:

$$3\omega_S + 3\omega_M + \omega_G = \int_{\hat{T}} d\hat{x} = \frac{1}{2}, \quad (12.88a)$$

$$\omega_S + \frac{\omega_M}{2} + \frac{\omega_G}{9} = \int_{\hat{T}} \hat{x}_1^2 d\hat{x} = \frac{1}{12}, \quad (12.88b)$$

$$\omega_S + \frac{\omega_M}{4} + \frac{\omega_G}{27} = \int_{\hat{T}} \hat{x}_1^3 d\hat{x} = \frac{1}{20}. \quad (12.88c)$$

From (12.88a)–(12.88c), we obtain:

$$\omega_S = \frac{1}{40}, \quad \omega_M = \frac{1}{15}, \quad \omega_G = \frac{9}{40}, \quad (12.89)$$

which are positive weights. This quadrature rule is known as the Simpson rule for triangles [108, 124].

The construction of the space \tilde{P}_3 is more difficult. We would like to have $P_3 \subset \tilde{P}_3 \subset P_4$ and to construct a quadrature formula exact for P_5 . The minimum (and also most natural) choice for the degrees of freedom is to

locate them at the vertices and on the edges as for P_3 and, on the other hand, to split the degree of freedom located at the center of the triangle into three degrees of freedom located on the medians, as indicated in Fig. 12.28. The notations are the same as those of P_3 . The only difference is that G is replaced by three points denoted by $\{G_1(\zeta), G_2(\zeta), G_3(\zeta)\}$. Here, we obtain two parameters: θ , defined as for P_3 and ζ such that, for any point O and $\ell = 1..3$, we have

$$OG_\ell = \zeta OS_\ell + (1 - \zeta)OM_\ell, \quad (12.90)$$

where M_ℓ is the midpoint of the edge opposite to S_ℓ .

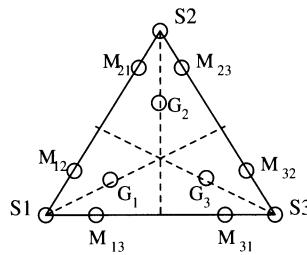


Fig. 12.28. The \tilde{P}_3 element

The most natural polynomial space corresponding to these degrees of freedom is

$$\tilde{P}_3 = P_3 + bP_1, \quad (12.91)$$

where b is the bubble function.

Now, we have to determine ω_S , ω_M (defined as for P_3), ω_G (corresponding to the three new points) θ and ζ so that the quadrature rule remains exact for P_5 , with the hope that its weights will be positive. Here also, the computations will be carried out for the unit triangle. Since the equations are too complicated, we shall simply indicate the different steps of the computations (aided by Maple).

We first compute the integral of $\hat{x}_1\hat{x}_2(1 - \hat{x}_1 - \hat{x}_2)(\hat{x}_1 - \zeta)(\hat{x}_1 - (1 - \zeta)/2)$ exactly and using the quadrature formula, which provides the value of ζ . We iterate this process by using the polynomials $\hat{x}_1\hat{x}_2(1 - \hat{x}_1 - \hat{x}_2)$, $\hat{x}_1(1 - \hat{x}_1)(\hat{x}_1 - \theta)(\hat{x}_1 - 1 + \theta)$, $(\hat{x}_1 - \theta)(\hat{x}_2 - \theta)(1 - \hat{x}_1 - \hat{x}_2 - \theta)$ and $(\hat{x}_1 - \zeta)(\hat{x}_2 - \zeta)(1 - \hat{x}_1 - \hat{x}_2 - \zeta)$ successively, which provides the values of ω_G , θ , ω_S and ω_M . We finally obtain:

$$\omega_G = \frac{7(14 - \sqrt{7})}{720} \simeq 0.110\,388\,53, \quad (12.92a)$$

$$\omega_M = \frac{7 + 4\sqrt{7}}{720} \simeq 0.024\,420\,84, \quad (12.92b)$$

$$\omega_S = \frac{8 - \sqrt{7}}{720} \simeq 0.007\,436\,46, \quad (12.92c)$$

$$\theta = \frac{21 - \sqrt{84\sqrt{7} - 147}}{42} \simeq 0.293\,469\,56, \quad (12.92d)$$

$$\zeta = \frac{7 + 2\sqrt{7}}{21} \simeq 0.585\,309\,65. \quad (12.92e)$$

A (numerical) study of the dispersion relations of the \tilde{P}_2 and \tilde{P}_3 elements provide errors in $O(h^4)$ and $O(h^6)$ respectively [30, 115].

Of course, the same process can be applied to higher-order approximations, but the computations are more complicated [23, 89].

From the stability point of view, a numerical study of the stability condition made for a regular mesh composed of squares divided into two triangles provides the following results:

– For \tilde{P}_2 :

- $\frac{c\Delta t}{h} \leq 0.218\,739$ for the leapfrog scheme,
- $\frac{c\Delta t}{h} \leq 0.378\,868$ for the modified equation approach.

– For \tilde{P}_3 :

- $\frac{c\Delta t}{h} \leq 0.124\,442$ for the leapfrog scheme,
- $\frac{c\Delta t}{h} \leq 0.215\,540$ for the modified equation approach.

Remarks

1. The above points and weights given for the unit triangle can be extended to any triangle. The coefficients defining the locations of the quadrature points remain the same and the weights must be multiplied by twice the area of the triangle.
2. Error estimates for these elements can be found in [10, 11, 37].
3. Another approach, based on the degeneration of quadrilateral spectral elements is given in [48].

12.4.3 Tetrahedral Elements

For tetrahedral elements, the basic problem remains the same and the solution is even more complicated.

Here,

$$P_r = \left\{ v(\hat{\mathbf{x}}) = \sum_{(\ell,m,n) \in \mathcal{G}_r} a_{\ell,m,n} \hat{x}_1^\ell \hat{x}_2^m \hat{x}_3^n, a_{\ell,m,n} \in \mathbb{R} \right\}, \quad (12.93)$$

with $\mathcal{G}_r = \{(\ell, m, n) \in \mathbb{N}^3 \text{ such that } \ell + m + n \leq r\}$ and $\text{card}(\mathcal{G}_r) = C_{r+3}^r = (r+1)(r+2)(r+3)/6$, such that $\dim(P_r) = (r+1)(r+2)(r+3)/6$.

In particular, $\dim(P_2) = 10$ and the degrees of freedom of the corresponding element are located at the vertices and at the middles of the edges (Fig. 12.29).

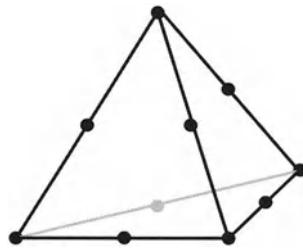


Fig. 12.29. The P_2 element in 3D

The weights ω_S and ω_M (defined as in the 2D case) of the quadrature rule corresponding to this element are

$$\omega_S = -\frac{1}{20}, \quad \omega_M = \frac{1}{5}, \quad (12.94)$$

which are obviously not acceptable.

The first idea would be, as in 2D, to search for a space \tilde{P}_2 such that $P_2 \subset \tilde{P}_2 \subset P_3$. Such a space would correspond to an element which would have its degrees of freedom located at

- the vertices of the tetrahedron
- the midpoints of its edges
- the centers of the faces

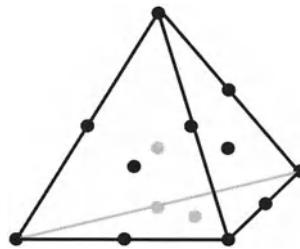


Fig. 12.30. The \tilde{P}_2 element such that $P_2 \subset \tilde{P}_2 \subset P_3$ in 3D

as indicated in Fig. 12.30. No interior point could be involved since its corresponding basis function would be in P_4 .

Of course, the quadrature formula must be exact for P_3 . So, if ω_M is the weight of the quadrature rule corresponding to the midpoint of an edge, we obtain, for the unit hexahedron \hat{H} :

$$\int_{\hat{H}} \hat{x}_1(1 - \hat{x}_1) \left(\hat{x}_1 - \frac{1}{3} \right) d\hat{\mathbf{x}} = 0 = \frac{3}{24} \omega_M. \quad (12.95)$$

In other words, we have $\omega_M = 0$, which is, of course, unacceptable.

For the P_3 element, the quadrature formula must be exact for P_4 . On this element, which has the same degrees of freedom as the previous one except for the degrees of freedom on the edges which are located at two symmetric points instead of the midpoint (Fig. 12.31), we obtain:

$$\int_{\hat{H}} \hat{x}_1(1 - \hat{x}_1) \left(\hat{x}_1 - \frac{1}{3} \right) d\hat{\mathbf{x}} = 0 = \theta(1 - \theta)\omega_M, \quad (12.96)$$

where ω_M is the weight corresponding to these degrees of freedom and θ and $1 - \theta$ are the coefficients that define these points as a barycenter of the two vertices of the edge. Equation (12.96) shows that we have two possibilities which are both unacceptable: either $\omega_M = 0$ or $\theta = 0$ or 1. In the first case, we have a weight equal to zero and, in the second case, the points on the edges are shifted to the vertices.

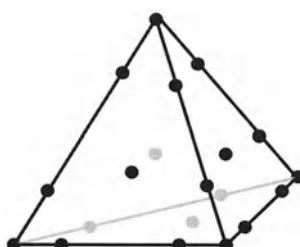


Fig. 12.31. The structure of the P_3 element in 3D

All the efforts to determine a space \tilde{P}_2 such that $P_2 \subset \tilde{P}_2 \subset P_3$ having been unsuccessful, we now construct a space \tilde{P}_2 such that $P_2 \subset \tilde{P}_2 \subset P_4$. Several possibilities can be developed but only one provides a quadrature rule with positive weights. The corresponding finite element has its degrees of freedom at the following points:

- the vertices of the tetrahedron,
- the midpoints of its edges,
- three points of its faces defined as in (12.90) with the same notations,
- the center of the tetrahedron.

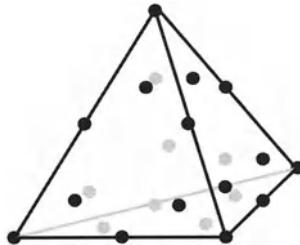


Fig. 12.32. The structure of the \tilde{P}_2 element such that $P_2 \subset \tilde{P}_2 \subset P_4$ in 3D

For this element, the space of approximation is

$$\tilde{P}_2 = P_2 + \sum_{j=1}^4 b_j P_1 + [b], \quad (12.97)$$

where b_j is the surface P_2 bubble function and b is the P_4 bubble function.

The weights and the value of ζ for the corresponding quadrature rule (exact for P_4) are, for the unit tetrahedron [23]:

$$\omega_S = \frac{13 - 3\sqrt{13}}{10080} \simeq 0.000\,216\,60, \quad (12.98a)$$

$$\omega_M = \frac{4 - \sqrt{13}}{315} \simeq 0.001\,252\,22, \quad (12.98b)$$

$$\omega_\zeta = \frac{29 + 17\sqrt{13}}{10080} \simeq 0.008\,957\,77, \quad (12.98c)$$

$$\omega_G = \frac{16}{315} \simeq 0.050\,793\,65, \quad (12.98d)$$

$$\zeta = \frac{7 - \sqrt{13}}{18} \simeq 0.188\,580\,48, \quad (12.98e)$$

where ω_S corresponds to the vertices, ω_M to the midpoints of the edges, ω_ζ to the points in the surface and ω_G to the center of the tetrahedron.

Remarks

1. As for triangles, the above formula, defined for the unit tetrahedron, remains valid for any tetrahedron if one multiplies the weights by six times the volume of the tetrahedron.
2. The \tilde{P}_2 element has 23 degrees of freedom instead of 10 for P_2 in 3D whereas it has 7 degrees of freedom instead of 6 in 2D. This remark shows that such a method is much more expensive in 3D than in 2D. It is not certain that the cost of the method constructed on the basis of the \tilde{P}_2 element is not equivalent to that using P_2 without mass-lumping and a good solver for the inversion of the mass matrix. The same remark holds for higher-order elements.
3. A \tilde{P}_3 tetrahedral element can be found in [23].

12.4.4 Non-Conforming Triangular Elements

A final direction of investigation is the use of non-conforming elements with mass-lumping to solve the wave equation. This point of view could perhaps lead to “lighter” triangular and especially tetrahedral elements. Here, the solution is not sought in $H^1(\Omega)$ but in $L^2(\Omega)$. The space of approximation can be written as

$$V_h^k = \{u_h \in L^2(\Omega) \mid u_h \text{ satisfies } \mathcal{C} \text{ and } \forall K \in \mathcal{T}, u_{h|_K} \in \check{P}_k\}, \quad (12.99)$$

where \check{P}_k is a polynomial space and

$$\mathcal{C} : \left\{ \begin{array}{l} \forall K_p \in \mathcal{T}, \forall K_q \in \mathcal{T} \text{ such that } K_p \cap K_q = a, \\ u_{h|_{K_p}} = u_{h|_{K_q}} \text{ on the interpolation points which belong to } a. \end{array} \right.$$

We search for a solution in $L^2(\Omega)$, which is, following Strang and Fix [106], a “variational crime”, since the space of approximation is not a subspace of the space to which the solution belongs. So, the consistency of such an approximation must be ensured by an additional condition. Here, the problem is that the values of the solution coincide only on a finite number of points on an edge of a triangle. The additional condition is given by the “patch test” which requires a good approximation of the integrals of the variational formulation along the edges [57, 72, 97, 106]. In the case of approximated integrals, the patch test says that we must have a quadrature formula on the

edges accurate enough to provide an exact value of the integral. Actually, in our case, the Gauss quadrature formulas are the most appropriate. For P_3 -like approximations, we obtain two elements:

- A P_3 element whose degrees of freedom are the three Gauss points on the edges and the center of the triangle (Fig. 12.33). For this element, the space of interpolation is exactly P_3 and the weights of the corresponding quadrature formula are:

$$\omega_G = \frac{9}{40}, \quad \omega_M = \frac{1}{20}, \quad \omega_a = \frac{1}{48}, \quad (12.100)$$

where ω_G corresponds to the center of the triangle, ω_M to the midpoint of the edges and ω_a to the other two Gauss points of the edges. Fortunately, all these weights are positive.

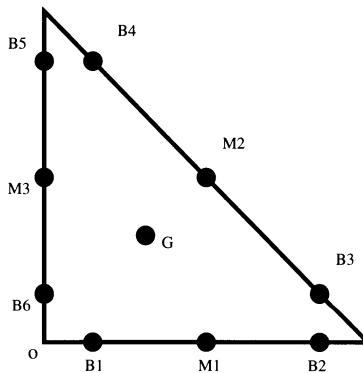


Fig. 12.33. The non-conforming P_3 triangular element

- A \tilde{P}_3 element whose degrees of freedom are the three Gauss points on the edges and three interior points C_j such that, if G is the center of the triangle and S_j one of its vertices ($j = 1..3$), we have

$$C_i = bS_i + (1 - b)G,$$

$$\text{where } b = -\frac{1}{180} \left(\frac{b_1^2 + 1390 - 40b_1 + i\sqrt{3}b_1^2 - 1390i\sqrt{3}}{b_1} \right) \simeq 0.385\,958,$$

$$\text{with } b_1 = (-29125 + 135i\sqrt{10815})^{\frac{1}{3}}.$$

(Fig. 12.34). For this element, the space of interpolation is the same as that of the conforming \tilde{P}_3 element and the weights of the corresponding quadrature formula are:

$$\begin{aligned}\omega_C &\simeq 0.112\,259\,227\,182\,220\,5, \\ \omega_M &\simeq 0.029\,202\,408\,032\,367\,31, \\ \omega_a &\simeq 0.012\,602\,515\,726\,039\,41,\end{aligned}\quad (12.101)$$

where ω_M and ω_a are defined as previously and ω_C corresponds to the points C_j . Here also, all these weights are positive.

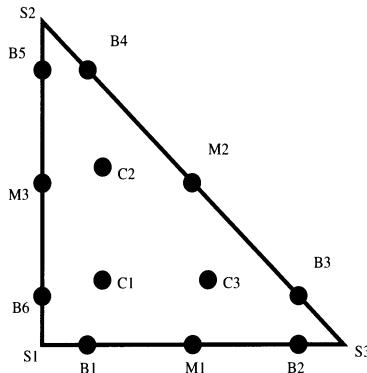


Fig. 12.34. The non-conforming \tilde{P}_3 triangular element

Although both elements have a dispersion relation in $O(h^4)$, the coefficient before the term in h^4 is much larger for P_3 than for \tilde{P}_3 and, therefore, the \tilde{P}_3 element provides a much more accurate approximation. On the other hand, both elements have a stability condition around 0.06 which is about four times more restrictive than that of the corresponding conforming element. All these remarks make this kind of approximation not very useful. Moreover, its extension to tetrahedra is not at all obvious.

So, the non-conforming approach was not as fruitful as we hoped but, as we shall see in the next chapter, it will be much more efficient for approximations of $H(\mathbf{curl}, \Omega)$ and $H(\text{div}, \Omega)$ for which we have no patch test since the natural approximations of these functional spaces is constructed on the basis of non-conforming elements.

12.5 A Numerical Illustration

As an illustration of the ability of the mass-lumped triangular elements to handle complex geometries, we have chosen to treat a problem of wave propagation in an exterior domain, which appears to be the complement of a “cone-sphere”-like obstacle. This is illustrated in Fig. 12.35, where we also present the computational mesh, which is, in fact, non-regular only in the

neighborhood of the obstacle. More precisely, we are interested in the diffraction of an incident wave generated by a pulse defined as in (12.73). This incident wave is emitted by a point source located at point (7.5,12). We present, in Fig. 12.35, a snapshot of the total field. We see, in particular, the diffraction phenomenon due to the vertex of the cone. We used for this computation a \tilde{P}_3 approximation.

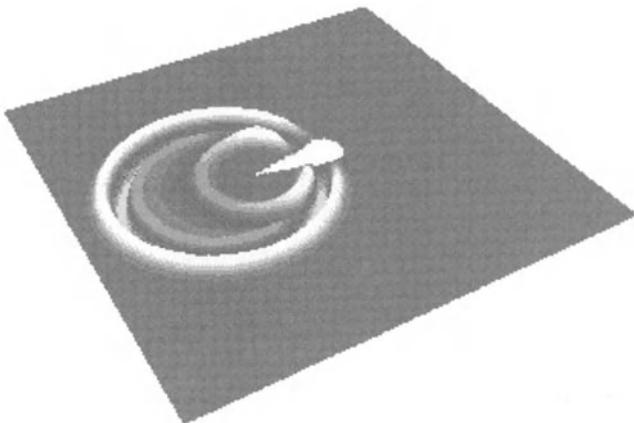
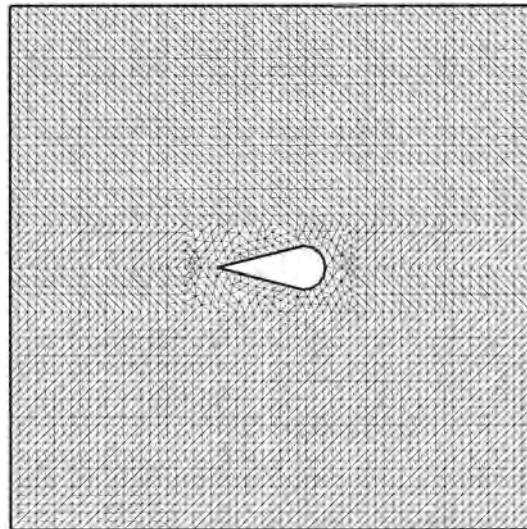


Fig. 12.35. The mesh around a cone-sphere obstacle (*above*) and the solution (*below*)

Find \mathbf{E} and \mathbf{B} such that $\mathbf{E}(., t) \in H_0(\mathbf{curl}, \Omega)$ and $\mathbf{B}(., t) \in V$ and

$$\frac{d}{dt} (\varepsilon \mathbf{E}, \boldsymbol{\varphi}) - (\mu^{-1} \mathbf{B}, \nabla \times \boldsymbol{\varphi}) = -(\mathbf{J}, \boldsymbol{\varphi}), \quad \forall \boldsymbol{\varphi} \in H_0(\mathbf{curl}, \Omega), \quad (13.2a)$$

$$\frac{d}{dt} (\mathbf{B}, \boldsymbol{\psi}) + (\nabla \times \mathbf{E}, \boldsymbol{\psi}) = 0, \quad \forall \boldsymbol{\psi} \in V, \quad (13.2b)$$

where

$$(\mathbf{u}, \mathbf{v}) = \int_{\Omega} \mathbf{u} \cdot \mathbf{v} \, dx$$

and V is a functional space.

As we saw in the second chapter, the natural definition of V would be $V = [L^2(\Omega)]^3$. However, in the following, we shall set $V = H(\text{div}, \Omega)$. This choice is legal since $\mathbf{B} \in H(\text{div}, \Omega)$. This feature can be derived from the fact that $\nabla \times \mathbf{E} \in H(\text{div}, \Omega)$ (since $\nabla \cdot (\nabla \times \mathbf{E}) = 0$).

13.1.2 A First Family of Hexahedral Edge Elements

Now, we define an orthogonal mesh \mathcal{Q} of Ω composed of rectangle parallelepipeds such that

$$\Omega = \bigcup_{j=1}^{N_e} K_j. \quad (13.3)$$

We are going to construct two spaces \mathbf{U}_{h0}^r and \mathbf{V}_h^r of finite elements such that $\mathbf{U}_{h0}^r \subset H_0(\mathbf{curl}, \Omega)$ and $\mathbf{V}_h^r \subset H(\text{div}, \Omega)$.

We first define the following polynomial space:

$$Q_{r_1, r_2, r_3} = \left\{ p(\mathbf{x}) = \sum_{\ell=0}^{r_1} \sum_{m=0}^{r_2} \sum_{n=0}^{r_3} a_{\ell, m, n} x_1^\ell x_2^m x_3^n, \quad a_{\ell, m, n} \in \mathbb{R} \right\}. \quad (13.4)$$

On the basis of this space, we have

$$\begin{aligned} \mathbf{U}_{h0}^r &= \{ \mathbf{v}_h \in H_0(\mathbf{curl}, \Omega) \text{ such that } \forall K_j \in \mathcal{Q}, \\ &\quad \mathbf{v}_h|_{K_j} \in Q_{r-1, r, r} \times Q_{r, r-1, r} \times Q_{r, r, r-1} \}. \end{aligned} \quad (13.5)$$

$$\begin{aligned} \mathbf{V}_h^r &= \{ \mathbf{v}_h \in H(\text{div}, \Omega) \text{ such that } \forall K_j \in \mathcal{Q}, \\ &\quad \mathbf{v}_h|_{K_j} \in Q_{r, r-1, r-1} \times Q_{r-1, r, r-1} \times Q_{r-1, r-1, r} \}. \end{aligned} \quad (13.6)$$

One can easily check that $\dim(Q_{r-1, r, r} \times Q_{r, r-1, r} \times Q_{r, r, r-1}) = 3r(r+1)^2$ and $\dim(Q_{r, r-1, r-1} \times Q_{r-1, r, r-1} \times Q_{r-1, r-1, r}) = 3r^2(r+1)$.

Of course, the fact that $\mathbf{v}_h \in H_0(\mathbf{curl}, \Omega)$ implies that the tangential component of \mathbf{v}_h is continuous and the fact that $\mathbf{v}_h \in H(\text{div}, \Omega)$ implies that the normal component of \mathbf{v}_h is continuous.

So, the finite element approximation of (13.2a) and (13.2b) can be written as:

Find \mathbf{E}_h and \mathbf{B}_h such that $\mathbf{E}_h(., t) \in \mathbf{U}_{h0}^r$ and $\mathbf{B}_h(., t) \in \mathbf{V}_h^r$ and

$$\frac{d}{dt} (\varepsilon \mathbf{E}_h, \varphi_h) - (\mu^{-1} \mathbf{B}_h, \nabla \times \varphi_h) = -(\mathbf{J}, \varphi_h), \quad \forall \varphi_h \in \mathbf{U}_{h0}^r, \quad (13.7a)$$

$$\frac{d}{dt} (\mathbf{B}_h, \psi_h) + (\nabla \times \mathbf{E}_h, \psi_h) = 0, \quad \forall \psi_h \in \mathbf{V}_h^r. \quad (13.7b)$$

Basically, the degrees of freedom for \mathbf{U}_{h0}^r are the circulations along the edges and the moments of different orders [92] and those of \mathbf{V}_h^r are the fluxes and the moments [99]. However, it will not be convenient to construct appropriate quadrature rules in order to mass-lump the mass matrix by using such degrees of freedom. It is easy to show that they can equivalently be defined by the values of the tangential components for \mathbf{U}_{h0}^r and of the normal components for \mathbf{V}_h^r . If these degrees of freedom are defined on regularly spaced points of an element, one can easily check that we obtain non-diagonal mass matrices for (13.7a) and (13.7b). In order to obtain mass-lumping, we must modify the locations of these degrees of freedom, i.e. define new interpolation points.

Let $0 = \xi_1^L < \xi_2^L < \dots < \xi_{r+1}^L = h$ and $\{\omega_\ell^L\}_{\ell=1}^{r+1}$ be the points and the weights of a Gauss-Lobatto quadrature rule² on the interval $[a, b]$. In the same way, we define the points $\{\xi_\ell^G\}_{\ell=1}^r$ and the weights $\{\omega_\ell^G\}_{\ell=1}^r$ of a Gauss quadrature formula [45] of the same order on the same interval. For these points, we define the Lagrange interpolation polynomials $\{l_\ell(x)\}_{\ell=1}^{r+1}$ and $\{g_\ell(x)\}_{\ell=1}^r$ such that $l_\ell(\xi_m^L) = \delta_{\ell m}$ and $g_\ell(\xi_m^G) = \delta_{\ell m}$. Of course, the Gauss points are all interior points whereas the Gauss-Lobatto points contain the ends of the interval. By using these functions, we can construct the 3D functions:

$$\varpi_{\ell,m,n}^{\alpha\beta\gamma}(x_1, x_2, x_3) = \alpha_\ell(x_1)\beta_m(x_2)\gamma_n(x_3), \quad (13.8)$$

where α , β and γ represent one of the two letters l and g . For instance, $\varpi_{\ell,m,n}^{glg}(x_1, x_2, x_3) = g_\ell(x_1)l_m(x_2)g_n(x_3)$.

From (13.8), we can deduce the basis functions corresponding to an element $K_j = [a_{1,j}, b_{1,j}] \times [a_{2,j}, b_{2,j}] \times [a_{3,j}, b_{3,j}]$. For \mathbf{U}_{h0}^r we have:

² In fact, $\xi_\ell^L = (b-a)\hat{\xi}_\ell^L + a$ and $\omega_\ell^L = (b-a)\hat{\omega}_\ell^L$ where $\hat{\xi}_\ell$ and $\hat{\omega}_\ell^L$ are the Gauss-Lobatto points and weights defined on $[0, 1]$.

$$\varphi_{\ell,m,n}^x = \left(\varpi_{\ell,m,n}^{gll}, 0, 0 \right), \ell = 1..r, m = 1..r+1, n = 1..r+1, \quad (13.9a)$$

$$\varphi_{\ell,m,n}^y = \left(0, \varpi_{\ell,m,n}^{lgl}, 0 \right), \ell = 1..r+1, m = 1..r, n = 1..r+1, \quad (13.9b)$$

$$\varphi_{\ell,m,n}^z = \left(0, 0, \varpi_{\ell,m,n}^{llg} \right), \ell = 1..r+1, m = 1..r+1, n = 1..r \quad (13.9c)$$

and, for \mathbf{V}_h^r ,

$$\psi_{\ell,m,n}^x = \left(\varpi_{\ell,m,n}^{lgg}, 0, 0 \right), \ell = 1..r+1, m = 1..r, n = 1..r, \quad (13.10a)$$

$$\psi_{\ell,m,n}^y = \left(0, \varpi_{\ell,m,n}^{glg}, 0 \right), \ell = 1..r, m = 1..r+1, n = 1..r, \quad (13.10b)$$

$$\psi_{\ell,m,n}^z = \left(0, 0, \varpi_{\ell,m,n}^{ggg} \right), \ell = 1..r, m = 1..r, n = 1..r+1. \quad (13.10c)$$

On a rectangular parallelepiped $[a_1, b_1] \times [a_2, b_2] \times [a_3, b_3]$, \mathbf{E}_h and \mathbf{B}_h can then be written as

$$\begin{aligned} \mathbf{E}_h &= \sum_{\ell=1}^r \sum_{m=1}^{r+1} \sum_{n=1}^{r+1} E_{\ell mn}^{(1)} \varphi_{\ell,m,n}^x + \sum_{\ell=1}^{r+1} \sum_{m=1}^r \sum_{n=1}^{r+1} E_{\ell mn}^{(2)} \varphi_{\ell,m,n}^y \\ &\quad + \sum_{\ell=1}^{r+1} \sum_{m=1}^{r+1} \sum_{n=1}^r E_{\ell mn}^{(3)} \varphi_{\ell,m,n}^z, \end{aligned} \quad (13.11)$$

$$\begin{aligned} \mathbf{B}_h &= \sum_{\ell=1}^{r+1} \sum_{m=1}^r \sum_{n=1}^r B_{\ell mn}^{(1)} \psi_{\ell,m,n}^x + \sum_{\ell=1}^r \sum_{m=1}^{r+1} \sum_{n=1}^r B_{\ell mn}^{(2)} \psi_{\ell,m,n}^y \\ &\quad + \sum_{\ell=1}^r \sum_{m=1}^r \sum_{n=1}^{r+1} B_{\ell mn}^{(3)} \psi_{\ell,m,n}^z, \end{aligned} \quad (13.12)$$

where $E_{\ell mn}^{(1)}$, $E_{\ell mn}^{(2)}$, $E_{\ell mn}^{(3)}$ and $B_{\ell mn}^{(1)}$, $B_{\ell mn}^{(2)}$, $B_{\ell mn}^{(3)}$ are the degrees of freedom for \mathbf{E}_h and \mathbf{B}_h .

An example of the locations of degrees of freedom $E_{\ell mn}^{(1)}$ and $B_{\ell mn}^{(1)}$ when $r = 3$ is given in Fig. 13.1.

By using the Gauss-Lobatto and Gauss rules, we construct, for two vector valued functions $\mathbf{u} = (u_1, u_2, u_3)$ and $\mathbf{v} = (v_1, v_2, v_3)$ on Ω , the following approximate scalar products on K_j :

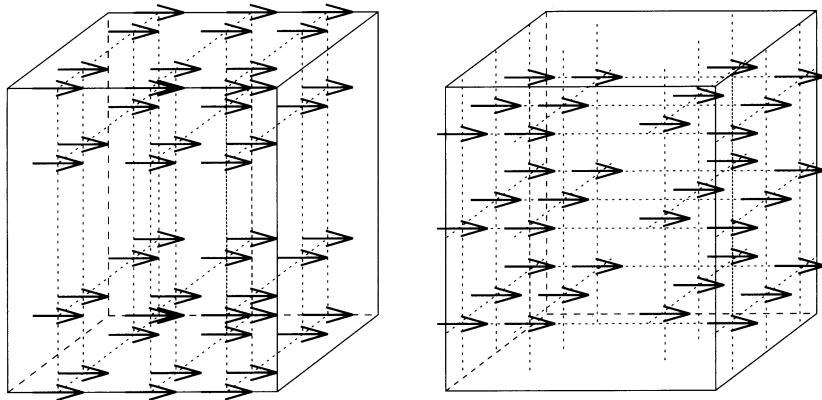


Fig. 13.1. The degrees of freedom for $E_{\ell m n}^{(1)}$ (left) and $B_{\ell m n}^{(1)}$ (right) when $r = 3$

$$(\mathbf{u}, \mathbf{v})_{h,j} =$$

$$\begin{aligned} & \sum_{\ell=1}^r \sum_{m=1}^{r+1} \sum_{n=1}^{r+1} \omega_\ell^G \omega_m^L \omega_n^L u_{1|K_j}(\xi_\ell^G, \xi_m^L, \xi_n^L) v_{1|K_j}(\xi_\ell^G, \xi_m^L, \xi_n^L) \\ & + \sum_{\ell=1}^{r+1} \sum_{m=1}^r \sum_{n=1}^{r+1} \omega_\ell^L \omega_m^G \omega_n^L u_{2|K_j}(\xi_\ell^L, \xi_m^G, \xi_n^L) v_{2|K_j}(\xi_\ell^L, \xi_m^G, \xi_n^L) \\ & + \sum_{\ell=1}^{r+1} \sum_{m=1}^{r+1} \sum_{n=1}^r \omega_\ell^L \omega_m^L \omega_n^G u_{3|K_j}(\xi_\ell^L, \xi_m^L, \xi_n^G) v_{3|K_j}(\xi_\ell^L, \xi_m^L, \xi_n^G), \end{aligned} \quad (13.13)$$

$$((\mathbf{u}, \mathbf{v}))_{h,j} =$$

$$\begin{aligned} & \sum_{\ell=1}^{r+1} \sum_{m=1}^r \sum_{n=1}^r \omega_\ell^L \omega_m^G \omega_n^G u_{1|K_j}(\xi_\ell^L, \xi_m^G, \xi_n^G) v_{1|K_j}(\xi_\ell^L, \xi_m^G, \xi_n^G) \\ & + \sum_{\ell=1}^r \sum_{m=1}^{r+1} \sum_{n=1}^r \omega_\ell^G \omega_m^L \omega_n^G u_{2|K_j}(\xi_\ell^G, \xi_m^L, \xi_n^G) v_{2|K_j}(\xi_\ell^G, \xi_m^L, \xi_n^G) \\ & + \sum_{\ell=1}^r \sum_{m=1}^r \sum_{n=1}^{r+1} \omega_\ell^G \omega_m^G \omega_n^L u_{3|K_j}(\xi_\ell^G, \xi_m^G, \xi_n^L) v_{3|K_j}(\xi_\ell^G, \xi_m^G, \xi_n^L). \end{aligned} \quad (13.14)$$

From the above scalar products, we derive the following scalar products on Ω :

$$(\mathbf{u}, \mathbf{v})_h = \sum_{j=1}^{N_e} (\mathbf{u}, \mathbf{v})_{h,j}, \quad (13.15)$$

$$((\mathbf{u}, \mathbf{v}))_h = \sum_{j=1}^{N_e} ((\mathbf{u}, \mathbf{v}))_{h,j}. \quad (13.16)$$

By using (13.15) and (13.16), we can write the discrete form of (13.7a) and (13.7b):

Find \mathbf{E}_h and \mathbf{B}_h such that $\mathbf{E}_h(., t) \in \mathbf{U}_{h0}^r$ and $\mathbf{B}_h(., t) \in \mathbf{V}_h^r$ and

$$\frac{d}{dt} (\varepsilon \mathbf{E}_h, \varphi_h)_h - ((\mu^{-1} \mathbf{B}_h, \nabla \times \varphi_h))_h = -(\mathbf{J}, \varphi_h)_h, \quad \forall \varphi_h \in \mathbf{U}_{h0}^r, \quad (13.17a)$$

$$\frac{d}{dt} ((\mathbf{B}_h, \psi_h))_h + ((\nabla \times \mathbf{E}_h, \psi_h))_h = 0, \quad \forall \psi_h \in \mathbf{V}_h^r. \quad (13.17b)$$

One can easily check that the basis functions defined in (13.9a)–(13.9c) form an orthogonal system for the scalar product $(,)_h$ and, in the same way, the basis functions defined in (13.10a)–(13.10c) form an orthogonal system for the scalar product $((,))_h$. So, by using the decompositions of \mathbf{E}_h and \mathbf{B}_h given in (13.11) and (13.12), one can see that the mass matrices obtained in (13.17a) and (13.17b) are both diagonal.

In particular, for $r = 1$, one obtains the following system on a regular mesh of space-step h :

$$\begin{aligned} & \varepsilon_{p+\frac{1}{2}, q, r} \frac{E_{1p+\frac{1}{2}, q, r}^{n+1} - E_{1p+\frac{1}{2}, q, r}^n}{\Delta t} \\ & + \frac{\mu_{p+\frac{1}{2}, q, r+\frac{1}{2}}^{-1} B_{2p+\frac{1}{2}, q, r+\frac{1}{2}}^{n+\frac{1}{2}} - \mu_{p+\frac{1}{2}, q, r-\frac{1}{2}}^{-1} B_{2p+\frac{1}{2}, q, r-\frac{1}{2}}^{n+\frac{1}{2}}}{h} \\ & - \frac{\mu_{p+\frac{1}{2}, q+\frac{1}{2}, r}^{-1} B_{3p+\frac{1}{2}, q+\frac{1}{2}, r}^{n+\frac{1}{2}} - \mu_{p+\frac{1}{2}, q-\frac{1}{2}, r}^{-1} B_{3p+\frac{1}{2}, q-\frac{1}{2}, r}^{n+\frac{1}{2}}}{h} = -J_{1p+\frac{1}{2}, q, r}^{n+\frac{1}{2}}, \end{aligned} \quad (13.18a)$$

$$\begin{aligned} & \varepsilon_{p, q+\frac{1}{2}, r} \frac{E_{2p, q+\frac{1}{2}, r}^{n+1} - E_{2p, q+\frac{1}{2}, r}^n}{\Delta t} \\ & + \frac{\mu_{p+\frac{1}{2}, q+\frac{1}{2}, r}^{-1} B_{3p+\frac{1}{2}, q+\frac{1}{2}, r}^{n+\frac{1}{2}} - \mu_{p-\frac{1}{2}, q+\frac{1}{2}, r}^{-1} B_{3p-\frac{1}{2}, q+\frac{1}{2}, r}^{n+\frac{1}{2}}}{h} \\ & - \frac{\mu_{p, q+\frac{1}{2}, r+\frac{1}{2}}^{-1} B_{1p, q+\frac{1}{2}, r+\frac{1}{2}}^{n+\frac{1}{2}} - \mu_{p, q+\frac{1}{2}, r-\frac{1}{2}}^{-1} B_{1p, q+\frac{1}{2}, r-\frac{1}{2}}^{n+\frac{1}{2}}}{h} = -J_{2p, q+\frac{1}{2}, r}^{n+\frac{1}{2}}, \end{aligned} \quad (13.18b)$$

$$\begin{aligned} & \varepsilon_{p,q,r+\frac{1}{2}} \frac{E_{3p,q,r+\frac{1}{2}}^{n+1} - E_{3p,q,r+\frac{1}{2}}^n}{\Delta t} \\ & + \frac{\mu_{p,q+\frac{1}{2},r+\frac{1}{2}}^{-1} B_{1p,q+\frac{1}{2},r+\frac{1}{2}}^{n+\frac{1}{2}} - \mu_{p,q-\frac{1}{2},r+\frac{1}{2}}^{-1} B_{1p,q-\frac{1}{2},r+\frac{1}{2}}^{n+\frac{1}{2}}}{h} \\ & - \frac{\mu_{p+\frac{1}{2},q,r+\frac{1}{2}}^{-1} B_{2p+\frac{1}{2},q,r+\frac{1}{2}}^{n+\frac{1}{2}} - \mu_{p-\frac{1}{2},q,r-\frac{1}{2}}^{-1} B_{2p-\frac{1}{2},q,r-\frac{1}{2}}^{n+\frac{1}{2}}}{h} = -J_{3p,q,r+\frac{1}{2}}^{n+\frac{1}{2}}, \end{aligned} \quad (13.18c)$$

$$\begin{aligned} & \frac{B_{1p,q+\frac{1}{2},r+\frac{1}{2}}^{n+\frac{1}{2}} - B_{1p,q+\frac{1}{2},r+\frac{1}{2}}^{n-\frac{1}{2}}}{\Delta t} - \frac{E_{2p,q+\frac{1}{2},r+1}^n - E_{2p,q+\frac{1}{2},r}^n}{h} \\ & + \frac{E_{3p,q+1,r+\frac{1}{2}}^n - E_{3p,q,r+\frac{1}{2}}^n}{h} = 0, \end{aligned} \quad (13.18d)$$

$$\begin{aligned} & \frac{B_{2p+\frac{1}{2},q,r+\frac{1}{2}}^{n+\frac{1}{2}} - B_{2p+\frac{1}{2},q,r+\frac{1}{2}}^{n-\frac{1}{2}}}{\Delta t} - \frac{E_{3p+1,q,r+\frac{1}{2}}^n - E_{3p,q,r+\frac{1}{2}}^n}{h} \\ & + \frac{E_{1p+\frac{1}{2},q,r+1}^n - E_{1p+\frac{1}{2},q,r}^n}{h} = 0, \end{aligned} \quad (13.18e)$$

$$\begin{aligned} & \frac{B_{3p+\frac{1}{2},q+\frac{1}{2},r}^{n+\frac{1}{2}} - B_{3p+\frac{1}{2},q+\frac{1}{2},r}^{n-\frac{1}{2}}}{\Delta t} - \frac{E_{1p+\frac{1}{2},q+1,r}^n - E_{1p+\frac{1}{2},q,r}^n}{h} \\ & + \frac{E_{2p+1,q+\frac{1}{2},r}^n - E_{2p,q+\frac{1}{2},r}^n}{h} = 0, \end{aligned} \quad (13.18f)$$

which provides the Yee scheme given in (4.80a)–(4.80f) by setting $\mathbf{B} = \mu \mathbf{H}$ when μ is a constant (Fig. 13.2).

A natural idea would be to use the same change of variables when μ depends on \mathbf{x} . Although natural, this idea is not legitimate since it would imply that \mathbf{H} belongs to $H(\text{div}, \Omega)$, which is not true. Beyond the mathematical issues of this statement, this change of variables would imply in practice that the normal component of \mathbf{H} is continuous, which is true as long as μ is itself continuous but is false for discontinuities of μ . Enforcing the normal component to be continuous in this case would provide an error of reflection-transmission in $O(h)$ instead of $O(h^2)$ ³.

³ For higher-order approximations, the loss of accuracy is even more important since we obtain $O(h)$ instead of $O(h^{2r})$.

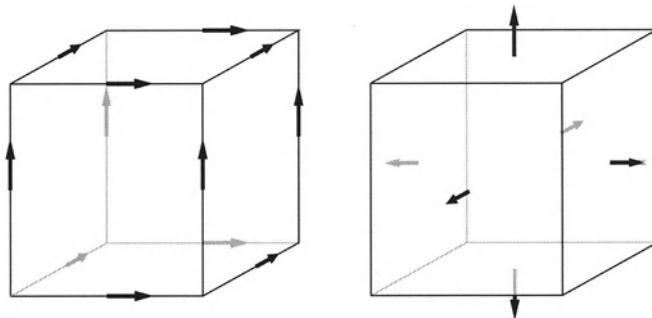


Fig. 13.2. The degrees of freedom for \mathbf{U}_{h0}^1 (left) and \mathbf{V}_h^1 (right)

So, the proper way to derive \mathbf{H} from \mathbf{B} is to “project” \mathbf{B} in $H(\mathbf{curl}, \Omega)$ by solving the following equation in \mathbf{H} :

$$(\mathbf{H}_h, \varphi_h)_h = (\mu^{-1} \mathbf{B}_h, \varphi_h)_h, \quad \forall \varphi_h \in \mathbf{U}_{h0}^r. \quad (13.19)$$

This equation provides a mean value of \mathbf{H} computed on the basis of the eight values of \mathbf{B} around it in the same direction. For instance, we have:

$$\begin{aligned} H_{1p+\frac{1}{2},q,r} = & \frac{1}{8} (\mu_{p,q+\frac{1}{2},r+\frac{1}{2}}^{-1} B_{1p,q+\frac{1}{2},r+\frac{1}{2}} + \mu_{p,q-\frac{1}{2},r+\frac{1}{2}}^{-1} B_{1p,q-\frac{1}{2},r+\frac{1}{2}} \\ & + \mu_{p,q-\frac{1}{2},r-\frac{1}{2}}^{-1} B_{1p,q-\frac{1}{2},r-\frac{1}{2}} + \mu_{p,q+\frac{1}{2},r-\frac{1}{2}}^{-1} B_{1p,q+\frac{1}{2},r-\frac{1}{2}} \\ & + \mu_{p+1,q+\frac{1}{2},r+\frac{1}{2}}^{-1} B_{1p+1,q+\frac{1}{2},r+\frac{1}{2}} + \mu_{p+1,q-\frac{1}{2},r+\frac{1}{2}}^{-1} B_{1p+1,q-\frac{1}{2},r+\frac{1}{2}} \\ & + \mu_{p+1,q-\frac{1}{2},r-\frac{1}{2}}^{-1} B_{1p+1,q-\frac{1}{2},r-\frac{1}{2}} + \mu_{p+1,q+\frac{1}{2},r-\frac{1}{2}}^{-1} B_{1p+1,q+\frac{1}{2},r-\frac{1}{2}}). \end{aligned} \quad (13.20)$$

Of course, the above formula is rather complicated and it would be time-consuming to derive \mathbf{H} from \mathbf{B} by using it at each degree of freedom of \mathbf{H} . Therefore, one should use $\mathbf{B} = \mu \mathbf{H}$ when μ is continuous and (13.20) holds at the discontinuities.

So, we have obtained a variational formulation of the Yee scheme for $r = 1$ and, for $r > 1$, we have variational extensions of this scheme.

Remarks

1. By searching for \mathbf{B} in $[L^2(\Omega)]^3$, one could derive \mathbf{H} from \mathbf{B} by setting $\mathbf{B} = \mu \mathbf{H}$, but, of course, we would not obtain the Yee scheme for $r = 1$. Moreover, the number of variables in \mathbf{B} would be much larger.

2. For $r = 1$, solving

$$(\varepsilon \mathbf{H}_h, \boldsymbol{\varphi}_h)_h = ((\mu^{-1} \mathbf{B}_h, \boldsymbol{\varphi}_h))_h, \quad \forall \boldsymbol{\varphi}_h \in \mathbf{U}_{h0}^r \quad (13.21)$$

would provide the same result as (13.19).

- 3. The functions l_j are actually the 1D basis functions defined in (11.11).
- 4. Mass-lumping is obtained here by using interpolation points which coincide with the quadrature points but also the fact that the supports of the degrees of freedom are orthogonal for the scalar product of \mathbb{R}^3 . This last feature will avoid the use of non-orthogonal meshes, as we shall see in the next section.

13.1.3 The 2D Case

As we saw it in Sect. 1.2.2, we have two versions of the Maxwell equations in 2D. By using the fields \mathbf{E} and \mathbf{B} and assuming that we are in an isotropic medium, we obtain:

- Transverse-magnetic (TM) equations:

$$\varepsilon(\mathbf{x}) \frac{\partial \mathbf{E}}{\partial t}(\mathbf{x}, t) - \mathbf{curl}(\mu^{-1}(\mathbf{x}) \mathbf{B}(\mathbf{x}, t)) = -\mathbf{J}(\mathbf{x}, t), \quad (13.22a)$$

$$\frac{\partial \mathbf{B}}{\partial t}(\mathbf{x}, t) + \mathbf{curl} \mathbf{E}(\mathbf{x}, t) = 0. \quad (13.22b)$$

- Transverse-electric (TE) equations:

$$\varepsilon(\mathbf{x}) \frac{\partial \mathbf{E}}{\partial t}(\mathbf{x}, t) - \mathbf{curl}(\mu^{-1}(\mathbf{x}) \mathbf{B}(\mathbf{x}, t)) = -\mathbf{J}(\mathbf{x}, t), \quad (13.23a)$$

$$\frac{\partial \mathbf{B}}{\partial t}(\mathbf{x}, t) + \mathbf{curl} \mathbf{E}(\mathbf{x}, t) = 0. \quad (13.23b)$$

In the TM case, $B = B_3$ and is independent of x_3 . So, we have no condition of regularity on the derivatives of B and, therefore, $B \in L^2(\Omega)$. On the other hand, \mathbf{E} is in $H_0(\mathbf{curl}, \Omega)$. If \mathcal{Q} is a mesh of Ω composed of rectangles⁴, the approximated solutions \mathbf{E}_h and B_h of (13.22a) and (13.22b) are searched in the following spaces:

$$\begin{aligned} \mathbf{U}_{h0}^r &= \{\mathbf{v}_h \in H_0(\mathbf{curl}, \Omega) \text{ such that } \forall K_j \in \mathcal{Q}, \\ &\quad \mathbf{v}_h|_{K_j} \in Q_{r-1,r} \times Q_{r,r-1}\}, \end{aligned} \quad (13.24)$$

$$V_h^r = \{v_h \in L^2(\Omega) \text{ such that } \forall K_j \in \mathcal{Q}, v_h|_{K_j} \in Q_{r-1}\}, \quad (13.25)$$

⁴ Ω is supposed to be a polygonal subdomain of \mathbb{R}^2 whose boundary is composed of segments parallel to the axes.

where Q_{r-1} is defined as in (12.3). Of course, the functions of \mathbf{U}_{h0}^r have continuous tangential components.

Now, if we denote, as in 3D,

$$\varpi_{\ell,m}^{\alpha\beta}(x_1, x_2) = \alpha_\ell(x_1)\beta_m(x_2), \quad (13.26)$$

where α and β represent one of the two letters l and g , the basis functions corresponding to \mathbf{U}_{h0}^r are

$$\boldsymbol{\varphi}_{\ell,m}^x = \left(\varpi_{\ell,m}^{gl}, 0 \right), \ell = 1..r, m = 1..r+1, \quad (13.27a)$$

$$\boldsymbol{\varphi}_{\ell,m}^y = \left(0, \varpi_{\ell,m}^{lg} \right), \ell = 1..r+1, m = 1..r, \quad (13.27b)$$

and those corresponding to V_h^r are

$$\varpi_{\ell,m}^{gg}, \ell = 1..r, m = 1..r. \quad (13.28)$$

In Fig. 13.3, we give the locations of the degrees of freedom for \mathbf{U}_{h0}^r and V_h^r when $r = 1$ and $r = 2$.

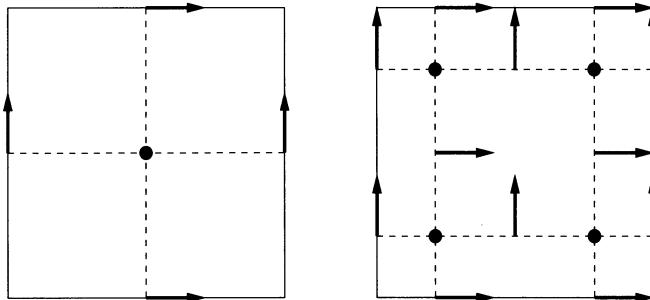


Fig. 13.3. The degrees of freedom for \mathbf{U}_{h0}^r (arrows) and V_h^r (black disks) when $r = 1$ (left) and $r = 2$ (right). The dashed lines indicate the Gauss points

In the TE case, E is such that $\mathbf{curl}E = (\partial E / \partial y, -\partial E / \partial x)^T \in [L^2(\Omega)]^2$, which implies that, actually, $E \in H_0^1(\Omega)$. On the other hand, $\mathbf{B} = (B_1, B_2)^T$ is such that $\partial B_1 / \partial x + \partial B_2 / \partial y \in L^2(\Omega)$ and, therefore, $\mathbf{B} \in H(\text{div}, \Omega)$. In this case, the spaces of approximation are

$$U_{h0}^r = \{v_h \in C^0(\Omega) \text{ such that } \forall K_j \in \mathcal{Q}, v_h|_{K_j} \in Q_r \text{ and } v_h|_{\partial\Omega} = 0\}, \quad (13.29)$$

$$\begin{aligned} \mathbf{V}_h^r &= \{\mathbf{v}_h \in H_0(\text{div}, \Omega) \text{ such that } \forall K_j \in \mathcal{Q}, \\ &\quad \mathbf{v}_h|_{K_j} \in Q_{r,r-1} \times Q_{r-1,r}\}. \end{aligned} \quad (13.30)$$

The basis functions corresponding to U_{h0}^r and \mathbf{V}_h^r are, respectively,

$$\varpi_{\ell,m}^l, \ell = 1..r+1, m = 1..r+1, \quad (13.31)$$

$$\boldsymbol{\varphi}_{\ell,m}^x = \left(\varpi_{\ell,m}^{lg}, 0 \right), \ell = 1..r+1, m = 1..r, \quad (13.32a)$$

$$\boldsymbol{\varphi}_{\ell,m}^y = \left(0, \varpi_{\ell,m}^{gl} \right), \ell = 1..r, m = 1..r+1. \quad (13.32b)$$

In Fig. 13.4, we give the locations of the degrees of freedom for U_{h0}^r and \mathbf{V}_h^r when $r = 1$ and $r = 2$.

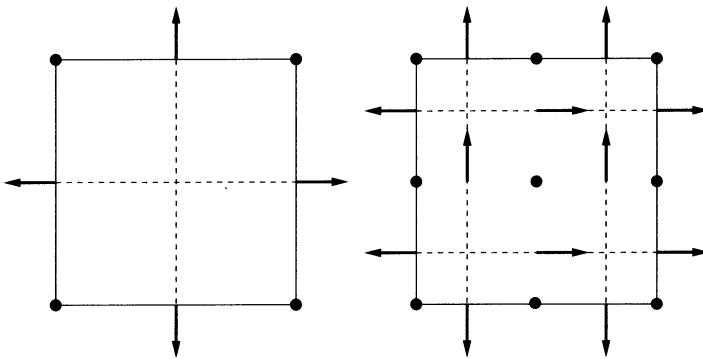


Fig. 13.4. The degrees of freedom for U_{h0}^r (black disks) and \mathbf{V}_h^r (arrows) when $r = 1$ (left) and $r = 2$ (right). The dashed lines indicate the Gauss points

Of course, in both cases, one obtains the variational version of the 2D Yee scheme. In the TM case, H is in U_{h0}^1 and should be computed as the mean value of the four values of $\mu_{-1}B$ around it. However, since $B \in L^2(\Omega)$ and has therefore no continuity properties, one can compute H at the same point as B . In the TE case, $\mathbf{H} \in \mathbf{U}_{h0}^r$ and must be computed as the mean value of the four normal components of $\mu_{-1}\mathbf{B}$ in the same direction.

A plane wave analysis shows that the stability and accuracy of the method in 2D and 3D are the same as those of the spectral elements of the same order applied to the wave equation [38].

13.2 Efficient Edge Elements for the Maxwell Equations

13.2.1 Extension to Anisotropic Media and Complex Geometries

Let us suppose that we want to solve (13.1a)–(13.1d) when ε is a (non-diagonal) symmetric definite positive matrix denoted $\underline{\varepsilon}$, on an orthogonal mesh. The mass matrix in \mathbf{E} , involved in (13.17a) leads to elementary numerical integrals of the basis functions defined in (13.9a)–(13.9c) of the form

$$I_M = (\underline{\varepsilon} \varphi_h, \varphi'_h)_{h,j}, \quad (13.33)$$

where φ_h and φ'_h are two basis functions on the orthogonal hexahedron K_j .

Unfortunately, the vector $\underline{\varepsilon} \varphi_h$ is here no longer orthogonal to φ'_h for the scalar product of \mathbb{R}^3 and, for this reason, I_M can not be equal to zero even when $\varphi_h \neq \varphi'_h$. This implies that, in such a case, the mass matrix may be non-diagonal.

As an illustration of this phenomenon, we set, for $r = 1$, $K_j = [0, 1]^3$ and

$$\underline{\varepsilon} = \begin{pmatrix} a & d & e \\ d & b & f \\ e & f & c \end{pmatrix}, \quad \varphi_h = \begin{pmatrix} yz \\ 0 \\ 0 \end{pmatrix}, \quad \varphi'_h = \begin{pmatrix} 0 \\ xz \\ 0 \end{pmatrix},$$

where $d \neq 0$.

For these values, we obtain $\underline{\varepsilon} \cdot \varphi_h, \varphi'_h = dxyz^2$ and

$$I_M = \frac{1}{4} \left[0 + 0 + 0 + \frac{d}{2} \right] = \frac{d}{8} \neq 0.$$

On the other hand, let us suppose that ε is a scalar function and the hexahedra K_j of \mathcal{Q} are such that $\mathbf{F}_j(\widehat{K}) = K_j$, where $\widehat{K} = [0, 1]^3$ and \mathbf{F}_j is the mapping defined in (12.38). One can show [49, 93] (cf. Appendix) that any function \mathbf{v}_h of \mathbf{U}_{h0}^r is such that

$$DF_j^* \mathbf{v}_h|_{K_j} \circ \mathbf{F}_j \in Q_{r-1,r,r} \times Q_{r,r-1,r} \times Q_{r,r,r-1}, \quad (13.34)$$

where DF_j^* is the adjoint matrix of the Jacobian matrix defined in (12.41).

Equation (13.34) implies that, for any basis function φ_h of \mathbf{U}_{h0}^r , there exists an interpolation function $\widehat{\varphi}_{\ell,m,n}^s$ ($s = x, y$ or z) on \widehat{K} , defined as in (13.9a)–(13.9c), such that

$$\varphi_h|_{K_j} \circ \mathbf{F}_j = DF_j^{*-1} \widehat{\varphi}_{\ell,m,n}^s. \quad (13.35)$$

The transform which associates $\widehat{\varphi}_{\ell,m,n}^s$ to $\varphi_h|_{K_j}$ is called the $H(\text{curl})$ -conforming transform because it ensures φ_h to be in $H(\text{curl}, \Omega)$, i.e. to have its tangential component continuous across two elements.

An integral of two basis functions multiplied by ε on a hexahedron K_j , which is a part of a mass matrix coefficient can be rewritten as

$$\begin{aligned}
\int_{K_j} \varepsilon \varphi_h \cdot \varphi'_h d\mathbf{x} &= \int_{\hat{K}} |J_j| \varepsilon \varphi_h|_{K_j} \circ \mathbf{F}_j \cdot \varphi'_h|_{K_j} \circ \mathbf{F}_j d\hat{\mathbf{x}} \\
&= \int_{\hat{K}} |J_j| \varepsilon DF_j^{*-1} \hat{\varphi}_{\ell,m,n}^s \cdot DF_j^{*-1} \hat{\varphi}_{\ell',m',n'}^{s'} d\hat{\mathbf{x}} \quad (13.36) \\
&= \int_{\hat{K}} |J_j| DF_j^{-1} \varepsilon DF_j^{*-1} \hat{\varphi}_{\ell,m,n}^s \cdot \hat{\varphi}_{\ell',m',n'}^{s'} d\hat{\mathbf{x}}
\end{aligned}$$

where J_j is the Jacobian defined in (12.21).

So, even in an isotropic medium, the mass matrix can be non-diagonal for a non-orthogonal mesh⁵, as we saw in the above example.

The mass matrix in \mathbf{B} , involved in (13.17b) suffers from the same problem since one can show [99, 114] that the *H(div)-conforming transform* is defined as

$$\psi_h|_{K_j} \circ \mathbf{F}_j = \frac{1}{|J_j|} DF_j \hat{\psi}_{\ell,m,n}^s \quad (13.37)$$

where J_j is the Jacobian and $\hat{\psi}_{\ell,m,n}^s$ is defined as in (13.10a)–(13.10c), and, therefore, we have, in the same way,

$$\int_{K_j} \mu \psi_h \cdot \psi'_h d\mathbf{x} = \int_{\hat{K}} \frac{1}{|J_j|} DF_j^* \mu DF_j \hat{\psi}_{\ell,m,n}^s \cdot \hat{\psi}_{\ell',m',n'}^{s'} d\hat{\mathbf{x}}. \quad (13.38)$$

In conclusion, this kind of approximation cannot be extended to non-orthogonal meshes or anisotropic media which both introduce a matrix in the mass integrals. This impossibility arises from the fact that mass-lumping is based, for this approximation, on the use of an adequate quadrature formula but also on the geometric orthogonality of the supports of the degrees of freedom which implies the orthogonality of the basis functions. This orthogonality is obviously destroyed when one multiplies a basis function by a non-diagonal matrix. So, in order to mass-lump the Maxwell equations in these cases, we must use another kind of approximation.

13.2.2 New Spaces of Approximation

So, we want to solve the Maxwell equations defined in (13.1a)–(13.1d), where ε and μ are replaced by (non-diagonal) symmetric definite positive matrices denoted $\underline{\varepsilon}$ and $\underline{\mu}$, on a domain Ω such that

⁵ We obtain a non-diagonal matrix $DF_j^{-1} \varepsilon DF_j^{*-1}$ in (13.36) even when the hexahedron is a non-rectangular parallelepiped, since DF_j^{-1} is a non-diagonal constant matrix in this case.

$$\Omega = \bigcup_{j=1}^{N_e} K_j, \quad K_j = \mathbf{F}_j(\widehat{K}). \quad (13.39)$$

Its variational formulation is the same as that given in (13.2a) and (13.2b) but its approximation is:

Find \mathbf{E}_h and \mathbf{B}_h such that $\mathbf{E}_h(., t) \in \widetilde{\mathbf{U}}_{h0}^r$ and $\mathbf{B}_h(., t) \in \widetilde{\mathbf{V}}_h^r$ and

$$\frac{d}{dt} (\underline{\varepsilon} \mathbf{E}_h, \boldsymbol{\varphi}_h) - (\mu^{-1} \mathbf{B}_h, \nabla \times \boldsymbol{\varphi}_h) = -(\mathbf{J}, \boldsymbol{\varphi}_h), \quad \forall \boldsymbol{\varphi}_h \in \widetilde{\mathbf{U}}_{h0}^r, \quad (13.40a)$$

$$\frac{d}{dt} (\mathbf{B}_h, \boldsymbol{\psi}_h) + (\nabla \times \mathbf{E}_h, \boldsymbol{\psi}_h) = 0, \quad \forall \boldsymbol{\psi}_h \in \widetilde{\mathbf{V}}_h^r, \quad (13.40b)$$

where

$$\begin{aligned} \widetilde{\mathbf{U}}_{h0}^r &= \{ \mathbf{v}_h \in H_0(\mathbf{curl}, \Omega) \\ &\text{such that } \forall K_j \in \mathcal{Q}, DF_j^* \mathbf{v}_h|_{K_j} \circ \mathbf{F}_j \in [Q_r]^3 \}, \end{aligned} \quad (13.41)$$

$$\begin{aligned} \widetilde{\mathbf{V}}_h^r &= \{ \mathbf{v}_h \in H(\text{div}, \Omega) \\ &\text{such that } \forall K_j \in \mathcal{Q}, |J_j| DF_j^{-1} \mathbf{v}_h|_{K_j} \circ \mathbf{F}_j \in [Q_r]^3 \} \end{aligned} \quad (13.42)$$

are spaces of approximation derived from [93].

13.2.3 Basis Functions and Degrees of Freedom

In order to describe the degrees of freedom for these spaces of approximation, we first define the basis functions on the unit cube $[0, 1]^3$.

As in (12.4), we define the Cartesian cube

$$\Xi_3 = \left\{ \hat{\xi}_{\ell,m,n} = (\hat{\xi}_\ell, \hat{\xi}_m, \hat{\xi}_n), \quad \ell = 1..r+1, m = 1..r+1, n = 1..r+1 \right\} \quad (13.43)$$

of the 1D set of Gauss-Lobatto quadrature points of order r .

At each point $\hat{\xi}_{\ell,m,n}$ of Ξ_3 , we define three vector-valued functions $\hat{\mathbf{v}} \in [Q_r]^3$ colinear to the canonical basis vectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$, at this point multiplied by $\epsilon = \pm 1$. The value of ϵ is chosen so that the basis functions are always towards the interior (or equivalently the exterior) of the reference element. The $3(r+1)^3$ basis functions are defined as follows:

$$\hat{\varphi}_{\ell,m,n}^x = (\epsilon_{\ell,m,n}^x \varpi_{\ell,m,n}^{lll}, 0, 0), \quad (13.44a)$$

$$\hat{\varphi}_{\ell,m,n}^y = (0, \epsilon_{\ell,m,n}^y \varpi_{\ell,m,n}^{lll}, 0), \quad (13.44b)$$

$$\hat{\varphi}_{\ell,m,n}^z = (0, 0, \epsilon_{\ell,m,n}^z \varpi_{\ell,m,n}^{lll}), \quad (13.44c)$$

where $\varpi_{\ell,m,n}^{lll}$ is defined as in (13.8) on the unit cube and $\epsilon_{\ell,m,n}^s = \pm 1$, $s = x, y$ or z .

Of course, we have

$$\begin{aligned}\hat{\varphi}_{\ell,m,n}^x(\hat{\xi}_{\ell',m',n'}) &= \left(\epsilon_{\ell,m,n}^x \delta_{\ell\ell'} \delta_{mm'} \delta_{nn'}, 0, 0 \right), \\ \hat{\varphi}_{\ell,m,n}^y(\hat{\xi}_{\ell',m',n'}) &= \left(0, \epsilon_{\ell,m,n}^y \delta_{\ell\ell'} \delta_{mm'} \delta_{nn'}, 0 \right), \\ \hat{\varphi}_{\ell,m,n}^z(\hat{\xi}_{\ell',m',n'}) &= \left(0, 0, \epsilon_{\ell,m,n}^z \delta_{\ell\ell'} \delta_{mm'} \delta_{nn'} \right).\end{aligned}$$

The restriction of the basis functions of $\tilde{\mathbf{U}}_{h0}^r$ to K_j is such that

$$\phi_{j,\ell,m,n}^s = DF_j^{*-1} \hat{\varphi}_{\ell,m,n}^s \circ F_j^{-1}, \quad s = x, y, \text{ or } z. \quad (13.45)$$

In the same way, we have for $\tilde{\mathbf{V}}_h^r$:

$$\psi_{j,\ell,m,n}^s = \frac{1}{|J_j|} DF_j \hat{\varphi}_{\ell,m,n}^s \circ F_j^{-1}, \quad s = x, y, \text{ or } z. \quad (13.46)$$

Although both are derived from the same basis functions on the unit cube, the basis functions of $\tilde{\mathbf{U}}_{h0}^r$ have their tangential components continuous but not their normal ones, whereas the basis functions of $\tilde{\mathbf{V}}_h^r$ have their normal components continuous but not their tangential ones. In particular, this means that the basis functions defined by (13.45) and (13.46) are only globally piecewise continuous.

Let us now define the degrees of freedom. At each point of K_j derived from a point of Ξ_3 , we have three degrees of freedom which are the components of the function on a local basis of \mathbb{R}^3 . This local basis is defined in different ways for $\tilde{\mathbf{U}}_{h0}^r$ and for $\tilde{\mathbf{V}}_h^r$. It contains:

- at a vertex, three vectors tangential to the three edges containing the vertex for $\tilde{\mathbf{U}}_{h0}^r$ and three vectors normal to the three faces containing this vertex for $\tilde{\mathbf{V}}_h^r$;
- at a point on an edge, one vector tangential to the edge and two vectors orthogonal to the first one and tangential to the two faces containing the edge for $\tilde{\mathbf{U}}_{h0}^r$ and two vectors normal to the faces containing the edge and a vector normal to these vectors for $\tilde{\mathbf{V}}_h^r$;
- at a point on a face, two vectors tangential to the face (which we can choose to be orthogonal to each other) and one vector orthogonal to these vectors for both spaces;
- at an interior point, three vectors parallel to the axes for both spaces.

On the other hand, for a function of $\tilde{\mathbf{U}}_{h0}^r$, the tangential components at a point are degrees of freedom defined continuously over all the mesh; the other components on the local basis are discontinuous, whereas for a function of $\tilde{\mathbf{V}}_h^r$, the normal components at a point are the continuous degrees of freedom.

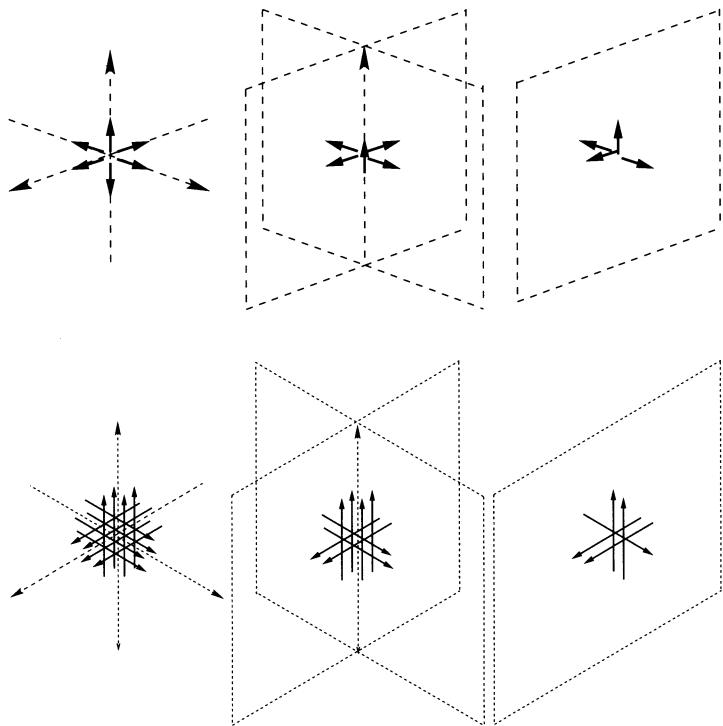


Fig. 13.5. The degrees of freedom for $\tilde{\mathbf{U}}_{h0}^r$ (above) and $\tilde{\mathbf{V}}_h^r$ (below) around a vertex (left), a point on an edge (center) and a point on a face (right) when the mesh is orthogonal

As a practical consequence of all these properties, we obtain the following number of degrees of freedom for each kind of point:

- For $\tilde{\mathbf{U}}_{h0}^r$, we have:
 - around a vertex, one degree of freedom per edge containing the vertex;
 - around a point on an edge, one degree of freedom on the edge and one degree of freedom per face containing the edge;
 - around a point on a face, two degrees of freedom on the face and one degree of freedom per element containing the face;
 - around an interior point, three degrees of freedom.

- For $\tilde{\mathbf{V}}_h^r$, we have:
 - around a vertex, one degree of freedom per face containing the vertex;
 - around a point on an edge, two degrees of freedom per face containing the edge;
 - around a point on a face, one degree of freedom on the face and two degrees of freedom per element containing the face;
 - around an interior point, three degrees of freedom.

In particular, for a structured mesh (i.e. on a mesh which can be derived from an orthogonal mesh by a conform mapping), we obtain (Fig. 13.5)⁶:

- six degrees of freedom for $\tilde{\mathbf{U}}_{h0}^r$ and twelve degrees of freedom for $\tilde{\mathbf{V}}_h^r$ around a vertex;
- five degrees of freedom for $\tilde{\mathbf{U}}_{h0}^r$ and eight degrees of freedom for $\tilde{\mathbf{V}}_h^r$ around a point on an edge;
- four degrees of freedom for $\tilde{\mathbf{U}}_{h0}^r$ and five degrees of freedom for $\tilde{\mathbf{V}}_h^r$ around a point on a face;
- three degrees of freedom for both spaces around an interior point.

Remarks

1. The $H(\text{curl})$ -conforming transform maps a function orthogonal to $\partial\hat{K}$ into a function orthogonal to ∂K_j . The $H(\text{div})$ -conforming transform has the same property for tangential functions. On these properties is based the choice of the non-continuous degrees of freedom.
2. The definitions of the basis functions given in (13.44a)–(13.44c) ensure the continuity of the normal components of the $H(\text{div})$ -conforming transformed functions and that of the tangential components of the $H(\text{curl})$ -conforming transformed functions and, on the other hand, they keep the invariance by symmetry or rotation of the basis functions on the unit element. However, the invariance by axial symmetry does not hold for tangential basis functions located at the center of a face or at the midpoint of an edge. This phenomenon induces some problems of compatibility between the degrees of freedom of two elements in some cases for $H(\text{curl})$ approximations. These problems necessitate the definition of a pointer which specifies the signs of the degrees of freedom in each element, which then implies extra storage and a more complicated way of programming. To avoid this trouble, it is better to use odd orders of approximation (for which the Gauss-Lobatto points do not contain midpoints) for $H(\text{curl})$ elements.

13.2.4 The Mass Integral

Now, let φ_h be a basis function of $\tilde{\mathbf{U}}_{h0}^r$ whose support is \mathcal{S} . We can write, for the mass integral of (13.40a):

$$\int_{\Omega} \underline{\underline{\varepsilon}} \mathbf{E}_h \cdot \varphi_h \, dx = \int_{\mathcal{S}} \underline{\underline{\varepsilon}} \mathbf{E}_h \cdot \varphi_h \, dx = \sum_{K_j \in \mathcal{S}} \int_{K_j} \underline{\underline{\varepsilon}} \mathbf{E}_h \cdot \varphi_h \, dx. \quad (13.47)$$

⁶ For non-structured meshes, one can obtain fewer or more degrees of freedom for a vertex and for a point on an edge.

Let us define the function ζ such that, for any point $F_j(\hat{\xi}_{\ell,m,n})$ of K_j , $\zeta(j, \ell, m, n)$ represents the number of this point in the mesh \mathcal{Q} . By using this function, we obtain, on each hexahedron K_j :

$$\begin{aligned} I_j &= \int_{K_j} \underline{\underline{\varepsilon}} \cdot \mathbf{E}_h \cdot \boldsymbol{\varphi}_h \, d\mathbf{x} = \\ &\sum_{s \in \{x, y, z\}} \sum_{\ell=1}^{r+1} \sum_{m=1}^{r+1} \sum_{n=1}^{r+1} E_{\zeta(j, \ell, m, n)}^s \times \\ &\int_{\widehat{K}} |J_j| \underline{\underline{\varepsilon}} \circ \mathbf{F}_j \boldsymbol{\phi}_{j, \ell, m, n}^s \circ \mathbf{F}_j \cdot \boldsymbol{\phi}_{j, \ell_j, m_j, n_j}^{s_j} \circ \mathbf{F}_j \, d\widehat{\mathbf{x}}, \end{aligned} \quad (13.48)$$

where $\boldsymbol{\phi}_{j, \ell_j, m_j, n_j}^{s_j} = \boldsymbol{\varphi}_h|_{K_j}$ and $E_{\zeta(j, \ell, m, n)}^s$ is a degree of freedom of \mathbf{E}_h .

By taking into account (13.45), we obtain:

$$\begin{aligned} &\int_{\widehat{K}} |J_j| \underline{\underline{\varepsilon}} \circ \mathbf{F}_j \boldsymbol{\phi}_{j, \ell, m, n}^s \circ \mathbf{F}_j \cdot \boldsymbol{\phi}_{j, \ell_j, m_j, n_j}^{s_j} \circ \mathbf{F}_j \, d\widehat{\mathbf{x}} = \\ &\int_{\widehat{K}} |J_j| \underline{\underline{\varepsilon}} \circ \mathbf{F}_j D\mathbf{F}_j^{*-1} \hat{\boldsymbol{\varphi}}_{\ell, m, n}^s \cdot D\mathbf{F}_j^{*-1} \hat{\boldsymbol{\varphi}}_{\ell_j, m_j, n_j}^{s_j} \, d\widehat{\mathbf{x}} = \\ &\int_{\widehat{K}} |J_j| D\mathbf{F}_j^{-1} \underline{\underline{\varepsilon}} \circ \mathbf{F}_j D\mathbf{F}_j^{*-1} \hat{\boldsymbol{\varphi}}_{\ell, m, n}^s \cdot \hat{\boldsymbol{\varphi}}_{\ell_j, m_j, n_j}^{s_j} \, d\widehat{\mathbf{x}} \simeq \\ &\sum_{\ell'=1}^{r+1} \sum_{m'=1}^{r+1} \sum_{n'=1}^{r+1} [|J_j| D\mathbf{F}_j^{-1} \underline{\underline{\varepsilon}} \circ \mathbf{F}_j D\mathbf{F}_j^{*-1} \hat{\boldsymbol{\varphi}}_{\ell, m, n}^s \cdot \hat{\boldsymbol{\varphi}}_{\ell_j, m_j, n_j}^{s_j}] (\hat{\xi}_{\ell', m', n'}). \end{aligned} \quad (13.49)$$

Now, if we set $M_j = |J_j| D\mathbf{F}_j^{-1} \underline{\underline{\varepsilon}} \circ \mathbf{F}_j D\mathbf{F}_j^{*-1}$, by definition of the basis functions of \widehat{K} , we have

$$\begin{aligned} &[M_j \hat{\boldsymbol{\varphi}}_{\ell, m, n}^s \cdot \hat{\boldsymbol{\varphi}}_{\ell_j, m_j, n_j}^{s_j}] (\hat{\xi}_{\ell', m', n'}) = \\ &M_{\mathcal{I}(s) \mathcal{I}(s_j)}^{(j)} \delta_{\ell \ell'} \delta_{mm'} \delta_{nn'} \delta_{\ell_j \ell'} \delta_{m_j m'} \delta_{n_j n'}, \end{aligned} \quad (13.50)$$

where $M_{pq}^{(j)}$ is the current term of M_j and $\mathcal{I}(x) = 1$, $\mathcal{I}(y) = 2$, $\mathcal{I}(z) = 3$.

Obviously, (13.50) is not equal to zero if and only if $\ell = \ell_j = \ell'$, $m = m_j = m'$ and $n = n_j = n'$. In particular, when $\hat{\boldsymbol{\varphi}}_{\ell, m, n}^s$ and $\hat{\boldsymbol{\varphi}}_{\ell_j, m_j, n_j}^{s_j}$ do not correspond to degrees of freedom located at the same point (i.e. $\ell = \ell_j$, $m = m_j$ and $n = n_j$), (13.50) is equal to zero for any value of $\hat{\xi}_{\ell', m', n'}$. In other words, the interactions of two basis functions corresponding to two

different points of the mesh are always equal to zero.

When $\ell = \ell_j = \ell'$, $m = m_j = m'$ and $n = n_j = n'$, we have

$$[M_j \hat{\varphi}_{\ell, m, n}^s \cdot \hat{\varphi}_{\ell_j, m_j, n_j}^{s_j}] (\hat{\xi}_{\ell', m', n'}) = M_{\mathcal{I}(s) \mathcal{I}(s_j)}^{(j)} \quad (13.51)$$

and, therefore,

$$\begin{aligned} I_j &= M_{1\mathcal{I}(s_j)}^{(j)} E_{\zeta(j, \ell_j, m_j, n_j)}^x + M_{2\mathcal{I}(s_j)}^{(j)} E_{\zeta(j, \ell_j, m_j, n_j)}^y \\ &\quad + M_{3\mathcal{I}(s_j)}^{(j)} E_{\zeta(j, \ell_j, m_j, n_j)}^z. \end{aligned} \quad (13.52)$$

Since the interactions of two basis functions corresponding to two different points of the mesh are always equal to zero, the mass matrix is *block-diagonal*. (13.47) and (13.52) show that each block corresponds to a point of the mesh and its terms are the interactions of the basis functions around this point. So, the size of a block is equal to the square of the number of degrees of freedom for $\tilde{\mathbf{U}}_{h0}^r$ located at the corresponding point of the mesh (Fig. 13.5).

A similar computation shows that the mass matrix derived from the mass integral of (13.40b) is also block-diagonal and the size of its blocks is equal to the square of the number of degrees of freedom for $\tilde{\mathbf{V}}_h^r$ around a point of the mesh (Fig. 13.5).

So, we replaced a global mass matrix by a block-diagonal one. From the computational point of view, the gain is valuable since the inverse of each block can be stored and the inversion of the mass matrix at each time-step boils to the product by a sequence of little block matrices. Moreover, since the mass matrix is symmetric, only a triangular part of the inverse of each block must be stored. However, one must notice that the size of the blocks is larger for $\tilde{\mathbf{V}}_h^r$ than for $\tilde{\mathbf{U}}_{h0}^r$. In particular, around a vertex, one obtains, on a structured mesh, a 12×12 block for $\tilde{\mathbf{V}}_h^r$ versus 6×6 for $\tilde{\mathbf{U}}_{h0}^r$.

13.2.5 The Stiffness Integral

The computation of the stiffness integrals is based on the following property:

If one denotes

$$\hat{\nabla} = \left(\frac{\partial}{\partial \hat{x}_1}, \frac{\partial}{\partial \hat{x}_2}, \frac{\partial}{\partial \hat{x}_3} \right)^T,$$

we have

$$\mathbf{v} \circ \mathbf{F}_j = D\mathbf{F}_j^{*-1} \hat{\mathbf{v}} \Rightarrow (\nabla \times \mathbf{v}) \circ \mathbf{F}_j = \frac{1}{J_i} \left(D\mathbf{F}_j \hat{\nabla} \times \hat{\mathbf{v}} \right). \quad (13.53)$$

In other words, the curl of a function behaves as a function of $H(\text{div}, \Omega)$ ⁷.

Now, let us consider the stiffness integral of (13.40a). As for the mass integral, we have, for any basis function φ_h of $\tilde{\mathbf{U}}_{h0}^r$:

$$\begin{aligned} \int_{\Omega} \underline{\underline{\mu}}^{-1} \mathbf{B}_h \cdot \nabla \times \varphi_h \, d\mathbf{x} &= \int_{\mathcal{S}} \underline{\underline{\mu}}^{-1} \mathbf{B}_h \cdot \nabla \times \varphi_h \, d\mathbf{x} \\ &= \sum_{K_j \in \mathcal{S}} \int_{K_j} \underline{\underline{\mu}}^{-1} \mathbf{B}_h \cdot \nabla \times \varphi_h \, d\mathbf{x} \end{aligned} \quad (13.54)$$

and

$$\begin{aligned} I'_j &= \int_{K_j} \underline{\underline{\mu}}^{-1} \mathbf{B}_h \cdot \nabla \times \varphi_h \, d\mathbf{x} = \\ &\sum_{s \in \{x, y, z\}} \sum_{\ell=1}^{r+1} \sum_{m=1}^{r+1} \sum_{n=1}^{r+1} B_{\zeta(j, \ell, m, n)}^s \times \end{aligned} \quad (13.55)$$

$$\int_{\hat{K}} |J_j| \underline{\underline{\mu}}^{-1} \circ \mathbf{F}_j \psi_{j, \ell, m, n}^s \circ \mathbf{F}_j \cdot (\nabla \times \phi_{j, \ell_j, m_j, n_j}^{s_j}) \circ F_j \, d\hat{\mathbf{x}},$$

where $\psi_{j, \ell, m, n}^s$ and $\phi_{j, \ell_j, m_j, n_j}^{s_j}$ are defined as in (13.45) and (13.46).

So, by using (13.45) and (13.53), we obtain:

$$\begin{aligned} \int_{\hat{K}} |J_j| \underline{\underline{\mu}}^{-1} \circ \mathbf{F}_j \psi_{j, \ell, m, n}^s \circ \mathbf{F}_j \cdot (\nabla \times \phi_{j, \ell_j, m_j, n_j}^{s_j}) \circ F_j \, d\hat{\mathbf{x}} &= \\ \int_{\hat{K}} \frac{1}{|J_j|} DF_j^* \underline{\underline{\mu}}^{-1} \circ \mathbf{F}_j DF_j \hat{\varphi}_{\ell, m, n}^s \cdot \hat{\nabla} \times \hat{\varphi}_{\ell_j, m_j, n_j}^{s_j} \, d\hat{\mathbf{x}}. \end{aligned} \quad (13.56)$$

Equation (13.56) shows that the stiffness matrix depends on the geometry and the physics of the problem and, therefore, one must store it for each degree of freedom. On the other hand, the presence of $DF_j^* \underline{\underline{\mu}}^{-1} \circ \mathbf{F}_j DF_j$ in front of $\hat{\varphi}_{\ell, m, n}^s$ implies that, in general, none of the interactions between the basis functions is equal to zero. All these remarks show that, finally, we have a huge storage for the stiffness matrix.

The question is: would it be possible to avoid it?

13.2.6 An Efficient Alternative

In order to avoid the storage of the stiffness matrix, one must first reformulate the Maxwell equations by using the fact that $\mathbf{B} = \underline{\underline{\mu}} \mathbf{H}$. By applying this

⁷ This is, by the way, an additional reason for taking \mathbf{B} in $H(\text{div}, \Omega)$.

change of variables, we obtain:

$$\underline{\underline{\varepsilon}}(\mathbf{x}) \frac{\partial \mathbf{E}}{\partial t}(\mathbf{x}, t) - \nabla \times \mathbf{H}(\mathbf{x}, t) = -\mathbf{J}(\mathbf{x}, t) \quad \text{in } \Omega, \quad (13.57a)$$

$$\underline{\underline{\mu}}(\mathbf{x}) \frac{\partial \mathbf{H}}{\partial t}(\mathbf{x}, t) + \nabla \times \mathbf{E}(\mathbf{x}, t) = 0 \quad \text{in } \Omega. \quad (13.57b)$$

The second step is to define the following variational formulation of (13.57a) and (13.57b):

Find \mathbf{E} and \mathbf{H} such that $\mathbf{E}(., t) \in H_0(\mathbf{curl}, \Omega)$ and $\mathbf{H}(., t) \in [L^2(\Omega)]^3$ and

$$\frac{d}{dt} (\underline{\underline{\varepsilon}} \mathbf{E}, \boldsymbol{\varphi}) - (\mathbf{H}, \nabla \times \boldsymbol{\varphi}) = -(\mathbf{J}, \boldsymbol{\varphi}), \quad \forall \boldsymbol{\varphi} \in H_0(\mathbf{curl}, \Omega), \quad (13.58a)$$

$$\frac{d}{dt} (\underline{\underline{\mu}} \mathbf{H}, \boldsymbol{\psi}) + (\nabla \times \mathbf{E}, \boldsymbol{\psi}) = 0, \quad \forall \boldsymbol{\psi} \in [L^2(\Omega)]^3. \quad (13.58b)$$

The third step is to define the following space of approximation:

$$\widetilde{\mathbf{W}}_h^r = \{\mathbf{v}_h \in [L^2(\Omega)]^3 \text{ such that } \forall K_j \in \mathcal{Q}, DF_j^* \mathbf{v}_h|_{K_j} \circ \mathbf{F}_j \in [Q_r]^3\}, \quad (13.59)$$

which is, of course, a subspace of $[L^2(\Omega)]^3$, i.e. whose functions are discontinuous at the boundaries of the hexahedra, in which we use locally the $H(\mathbf{curl})$ -conforming transform.

By using this space, we can write the following approximated formulation of (13.58a) and (13.58b):

Find \mathbf{E}_h and \mathbf{H}_h such that $\mathbf{E}_h(., t) \in \widetilde{\mathbf{U}}_{h0}^r$ and $\mathbf{H}_h(., t) \in \widetilde{\mathbf{W}}_h^r$ and

$$\frac{d}{dt} (\underline{\underline{\varepsilon}} \mathbf{E}_h, \boldsymbol{\varphi}_h) - (\mathbf{H}_h, \nabla \times \boldsymbol{\varphi}_h) = -(\mathbf{J}, \boldsymbol{\varphi}_h), \quad \forall \boldsymbol{\varphi} \in \widetilde{\mathbf{U}}_{h0}^r, \quad (13.60a)$$

$$\frac{d}{dt} (\underline{\underline{\mu}} \mathbf{H}_h, \boldsymbol{\psi}_h) + (\nabla \times \mathbf{E}_h, \boldsymbol{\psi}_h) = 0, \quad \forall \boldsymbol{\psi} \in \widetilde{\mathbf{W}}_h^r. \quad (13.60b)$$

The advantage of the somehow strange definition of $\widetilde{\mathbf{W}}_h^r$ lies in the following:

If we consider the elementary stiffness integral given in (13.56), in which now $\boldsymbol{\psi}_{j,\ell,m,n}^s \in \widetilde{\mathbf{W}}_h^r$, $\boldsymbol{\phi}_{j,\ell_j,m_j,n_j}^{s_j} \in \widetilde{\mathbf{U}}_{h0}^r$ and we suppressed $\underline{\underline{\mu}}^{-1} \circ \mathbf{F}_j$, by using (13.53) and the definition of $\widetilde{\mathbf{W}}_h^r$, we obtain:

$$\int_{K_j} \boldsymbol{\psi}_{j,\ell,m,n}^s \cdot \nabla \times \boldsymbol{\phi}_{j,\ell_j,m_j,n_j}^{s_j} d\mathbf{x} =$$

$$\int_{\widehat{K}} |J_j| \boldsymbol{\psi}_{j,\ell,m,n}^s \circ \mathbf{F}_j \cdot (\nabla \times \boldsymbol{\phi}_{j,\ell_j,m_j,n_j}^{s_j}) \circ \mathbf{F}_j d\hat{\mathbf{x}} =$$

$$\int_{\widehat{K}} |J_j| DF_j^{*-1} \hat{\boldsymbol{\varphi}}_{\ell,m,n}^s \cdot \frac{1}{J_j} DF_j (\widehat{\nabla} \times \hat{\boldsymbol{\varphi}}_{\ell_j,m_j,n_j}^{s_j}) d\hat{\mathbf{x}} =$$

$$\int_{\hat{K}} \frac{|J_j|}{J_j} DF_j^* DF_j^{*-1} \hat{\varphi}_{\ell,m,n}^s \cdot \hat{\nabla} \times \hat{\varphi}_{\ell_j,m_j,n_j}^{s_j} d\hat{x} = \\ \text{sgn}(J_j) \int_{\hat{K}} \hat{\varphi}_{\ell,m,n}^s \cdot \hat{\nabla} \times \hat{\varphi}_{\ell_j,m_j,n_j}^{s_j} d\hat{x}. \quad (13.61)$$

In other words, by knowing the stiffness integrals on \hat{K} only and the sign $\text{sgn}(J_j)$ of J_j on each element, we know the stiffness matrix on the whole mesh \mathcal{Q} . This means that we replace the huge storage of the stiffness matrix by the storage of one value of $\text{sgn}(J_j)$ per element and the (negligible) storage of the local interactions of the basis functions on \hat{K} ! The whole information about physics and geometry is now contained in the block-diagonal mass matrices.

Besides this dramatical gain of storage, this formulation induces also a significant gain of time of computation. One can show that a Lagrange basis function of $[Q_r]^3$ has at most $2r+1$ non-null interactions only on \hat{K} whereas the restriction of a basis function of $\tilde{\mathbf{V}}_h^r$ to K_j had $3(r+1)^3$ non-null interactions on K_j . So, for the $3(r+1)^3$ basis functions of one element, we have $6(r+1)^4$ non-null interactions for the stiffness matrix derived from (13.61) instead of $9(r+1)^6$ for that derived from (13.54), i.e. a ratio at least equal to $3/2(r+1)^2$. For instance, for $r=3$, we obtain a ratio of 24 and, for $r=5$, a ratio of 54, which implies a significant gain of computation time!

The fact that $\tilde{\mathbf{W}}_h^r$ is a space of discontinuous functions whereas the functions of $\tilde{\mathbf{V}}_h^r$ had continuous normal components does not lead to a significant number of additional variables. Actually, one can show that the number of degrees of freedom of the two spaces tends to be equivalent when r increases. For instance, when $r=3$, we obtain, for a cube containing N^3 elements, $192N^3$ degrees of freedom for $\tilde{\mathbf{W}}_h^r$ instead of $168N^3$ degrees of freedom for $\tilde{\mathbf{V}}_h^r$. On the other hand, the elementary block matrices of the mass matrix are all 3×3 for $\tilde{\mathbf{W}}_h^r$, whereas these matrices can be 12×12 and even more for $\tilde{\mathbf{V}}_h^r$.

A concluding remark:

It would be natural to replace $[L^2(\Omega)]^3$ by $H(\mathbf{curl}, \Omega)$ in the definition of $\tilde{\mathbf{W}}_h^r$. This definition would lead to relation (13.61) and the number of degrees of freedom involved for $H(\mathbf{curl}, \Omega)$ is much smaller than that provided by $[L^2(\Omega)]^3$ (for instance, on the N^3 cube, we would have $96N^3$ degrees of freedom instead of $192N^3$). However, such a definition leads to dispersive approximations whereas searching for the solution in $[L^2(\Omega)]^3$ leads to a numerical dispersion equivalent to that of (13.40a) and (13.40b). Actually, one can show that these approximations have the same numerical dispersion as that of the first family of edge elements and, therefore, the same as that of the

spectral element method applied to the wave equation [38]. This dispersion seems to be optimal.

13.2.7 The 2D Case

The 2D case is, of course, based on the set Ξ_2 defined in (12.4). For the TM equations, which are now those given in (1.16a) and (1.16b), the spaces of approximation are

$$\begin{aligned} \tilde{\mathbf{U}}_{h0}^r &= \{\mathbf{v}_h \in H_0(\text{curl}, \Omega) \\ &\quad \text{such that } \forall K_j \in \mathcal{Q}, DF_j^* \mathbf{v}_h|_{K_j} \circ \mathbf{F}_j \in [Q_r]^2\}, \end{aligned} \quad (13.62)$$

$$\tilde{W}_h^r = \{v_h \in L^2(\Omega) \text{ such that } \forall K_j \in \mathcal{Q}, v_h|_{K_j} \circ F_j \in Q_r\} \quad (13.63)$$

and, for the TE equations, given in (1.19a) and (1.19b), we have

$$\tilde{U}_{h0}^r = \{v_h \in H_0^1(\Omega) \text{ such that } \forall K_j \in \mathcal{Q}, v_h|_{K_j} \circ F_j \in Q_r\}, \quad (13.64)$$

$$\tilde{\mathbf{W}}_h^r = \{\mathbf{v}_h \in [L^2(\Omega)]^2 \text{ such that } \forall K_j \in \mathcal{Q}, DF_j^* \mathbf{v}_h|_{K_j} \circ \mathbf{F}_j \in [Q_r]^2\}. \quad (13.65)$$

A study of the numerical dispersion on a periodical non-regular mesh can be found in [39].

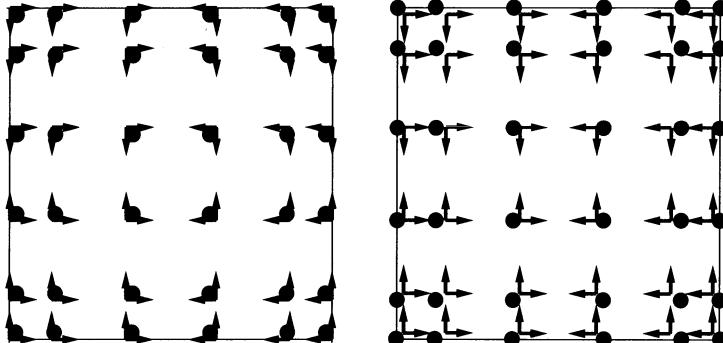


Fig. 13.6. The degrees of freedom for the TM equations (left) and the TE equations (right) in 2D when $r = 5$. The *points* indicate the degrees of freedom for the scalar unknown whereas the *arrows* represent the degrees of freedom for the vector valued unknown

From the stability point of view, both edge element methods have the same stability conditions as the quadrilateral and hexahedral spectral elements defined in the previous chapter.

13.2.8 A 2D Numerical Experiment

In order to illustrate the possibilities of the method, we give, in this section, a numerical experiment concerning propagation of the magnetic field in a 2D heterogeneous domain with anisotropy (for a TM model).

In the domain $\Omega =]0, 30[^2$, we define a subdomain

$$\Omega' = \{(x_1, x_2) \in \mathbb{R}^2 \text{ such that } (x_1 - 15)^2 + (x_2 - 15)^2 - 25 < 0\}.$$

In Ω , we solve the 2D Maxwell equations given in (1.16b) with the following values of the parameters:

$$\begin{aligned} \mu &= 1, \\ \varepsilon &= \begin{cases} I_2 & \text{if } (x_1, x_2) \in \Omega', \\ & = A \text{ otherwise,} \end{cases} \end{aligned}$$

where

$$A = \begin{pmatrix} 1 & 1/4 \\ 1/4 & 2 \end{pmatrix}, \quad I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The source, located at the center of the domain, is the pulse defined in (12.73) and is parallel to the first component of \mathbf{E} .

Ω is meshed by a mesh \mathcal{M} (8384 quadrilaterals, 201 920 degrees of freedom in \mathbf{E} and 134 144 degrees of freedom in H), as shown in Fig. 13.7 and we use Q_3 elements. The approximation in time is made by a leapfrog scheme with a time-step such that $c_M \Delta t / h \leq \alpha_M / 2$, where c_M is the maximum velocity in the domain and α_M is the stability condition ($\simeq 0.164$). The CPU time for this computation is about 36 s⁸.

In Fig. 13.8, we give the snapshots for $t = 2$ (when the wave is at the boundary of Ω') $t = 3$, $t = 4$ and $t = 5$. The effect of the anisotropy appears clearly.

13.3 Triangular and Tetrahedral Edge Elements

A natural extension to triangles and tetrahedra of the approximation defined on the basis of quadrilaterals and hexahedra in the previous section, would be to use the scalar mass-lumped finite elements constructed for the wave equation in Sect. 12.4 and to introduce d degrees of freedom at each point

⁸ On a DEC AlphaStation 500, 1 processor 21164 (500 MHz), 256 MB, 4.3 GB.

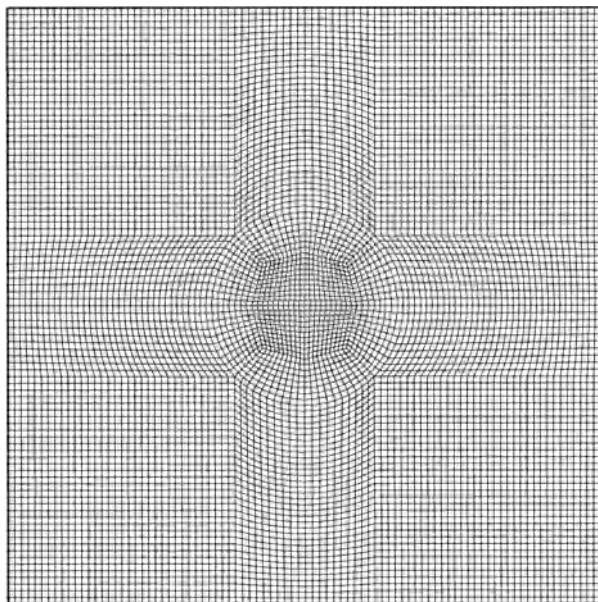


Fig. 13.7. The mesh \mathcal{M}

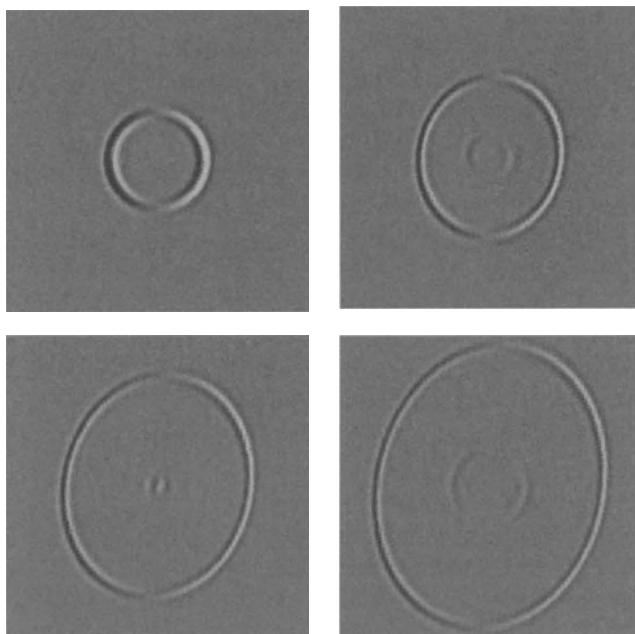


Fig. 13.8. Snapshots of the magnetic field from $t = 5$ (above left) to $t = 12.5$ (below right). One can notice the effect of anisotropy and the reflection in the interior disk

We also introduce the set of functions

$$\mathcal{R}_1 = \{\mathbf{u} = \mathbf{a} + b(x_2, -x_1)^T, \mathbf{a} \in \mathbb{R}^2, b \in \mathbb{R}\} \quad (13.67)$$

and the three basis functions

$$\mathbf{w}_1 = \nabla \lambda_1 \lambda_2 \lambda_3, \quad (13.68a)$$

$$\mathbf{w}_2 = \lambda_1 \nabla \lambda_2 \lambda_3, \quad (13.68b)$$

$$\mathbf{w}_3 = \lambda_1 \lambda_2 \nabla \lambda_3. \quad (13.68c)$$

So, we can define

$$\tilde{\mathcal{R}}_1 = \mathcal{R}_1 \oplus \{\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3\}, \quad (13.69)$$

for which the set of degrees of freedom is unisolvant.

Of course, we have here two kinds of basis functions:

- 3 basis functions corresponding to the tangential components:

$$\mathbf{w}_j^\tau = \frac{4}{\|\nabla \lambda_j\|} \mathbf{w}_j, \quad j = 1..3, \quad (13.70)$$

- 3 basis functions corresponding to the normal components (which are actually the basis functions of \mathcal{R}_1):

$$\mathbf{w}_j^\nu = \mathbf{u}_j + \sum_{\ell=1}^3 \alpha_{j\ell} \mathbf{w}_\ell^\tau, \quad (13.71)$$

where $\mathbf{u}_j = \nabla \lambda_m \lambda_n - \nabla \lambda_n \lambda_m$, $(m, n) \in \{1, 2, 3\}^2$, $m \neq j$, $n \neq j$, $\alpha_{j\ell} = -(\mathbf{u}_j \cdot \boldsymbol{\nu})(M_\ell)$. $\boldsymbol{\nu}$ is the unit outward normal to ∂T and M_ℓ is the midpoint of the edge opposite to the vertex S_ℓ .

The appropriate quadrature formula is, of course, given by

$$\int_T f(\mathbf{x}) d\mathbf{x} \simeq \frac{\text{mes}(T)}{3} \sum_{\ell=1}^3 f(M_\ell), \quad (13.72)$$

where $\text{mes}(T)$ is the measure of the triangle.

The Second-Order Element. In order to define the second-order element, we shall first introduce some additional notations. For each edge of T , we denote by M_{pq} the point of the edge $S_p S_q$ such that

$$S_p M_{pq} = \alpha S_p S_q, \quad (13.73)$$

where α is a given real constant. At each of the points M_{pq} , we define two degrees of freedom which are the values of the normal and tangential components of a function and at G , the two components of the functions in the two directions of space are the degrees of freedom (Fig. 13.10).

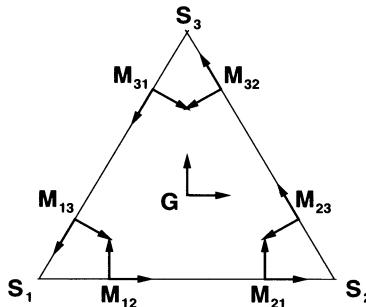


Fig. 13.10. The second-order triangular edge element with mass-lumping

Let \mathcal{V}_2 be the space of polynomials generated by the vectors $(x_2^2, -x_1x_2)^T$, $(-x_1x_2, x_1^2)^T$,

$$\mathcal{R}_2 = (P_1(T))^2 \oplus \mathcal{V}_2, \quad (13.74)$$

and

$$\mathbf{w}_{mn} = \nabla \lambda_\ell \lambda_m \lambda_n \phi_{nm}, \quad (13.75a)$$

$$\mathbf{w}_{nm} = \nabla \lambda_\ell \lambda_m \lambda_n \phi_{mn}, \quad (13.75b)$$

where $(\ell, m, n) \in \{1, 2, 3\}^3$ and are all different, and ϕ_{nm} is the polynomial of P_1 which vanishes at G and M_{nm} .

With the above notations, we can define the space of polynomials $\tilde{\mathcal{R}}_2$ for which the set of degrees of freedom is unisolvant:

$$\tilde{\mathcal{R}}_2 = \mathcal{R}_2 \oplus \{\mathbf{w}_{12}, \mathbf{w}_{21}, \mathbf{w}_{13}, \mathbf{w}_{31}, \mathbf{w}_{23}, \mathbf{w}_{32}\}. \quad (13.76)$$

The corresponding quadrature rule to obtain mass-lumping is given by

$$\int_T f(\mathbf{x}) d\mathbf{x} \simeq \text{mes}(T) \left[\frac{9}{40} f(G) + \frac{11}{240} \sum_{(p,q)} f(M_{pq}) \right], \quad (13.77)$$

where $(p, q) \in \{1, 2, 3\}^2$ and $p \neq q$.

The stability conditions for a regular mesh in a homogeneous isotropic medium such that $c = \sqrt{\varepsilon\mu}$ are, for a leapfrog scheme on a regular mesh composed of rectangle triangles [50]:

- $c\Delta t/h \leq 0.2654$ for the first-order element (versus 0.7071 for quadrilateral elements),
- $c\Delta t/h \leq 0.1167$ for the second-order element (versus 0.2886 for quadrilateral elements).

Remark

A third-order element was studied in [50] but was less efficient than the second-order one because of its large number of degrees of freedom. For this reason, we do not describe it here.

13.3.2 Tetrahedral Edge Elements

For tetrahedral elements, the problem is slightly more complicated since one must obtain the continuity of the tangential component over all the faces. A sufficient condition for ensuring this property is that the space of approximation on a face of the tetrahedron is a subspace of the 2D $H(\text{curl}, \Omega)$ continuous space. As for the 2D case, we shall solve the second-order formulation given in (1.9) or (1.10).

The First-Order Element. Let T now be a tetrahedron whose vertices are (S_1, S_2, S_3, S_4) and $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ the corresponding barycentric coordinates. An edge is defined by its ends and each face is affected of the index of its opposite vertex. So, the face Φ_j is such that $S_j \notin \Phi_j$. At each midpoint M_{pq} of an edge $S_p S_q$, we define the three unit vectors $(e_{pq,1}, e_{pq,2}, e_{pq,3})$ such that $e_{pq,1}$ is colinear to $S_p S_q$, and $e_{pq,2}, e_{pq,3}$ are the two vector orthogonal to $e_{pq,1}$ and parallel to the two faces containing $S_p S_q$. The degrees of freedom are the three components of a function at the point M_{pq} in this local basis, for each edge of the tetrahedron T (Fig. 13.11). So, we have 18 degrees of freedom.

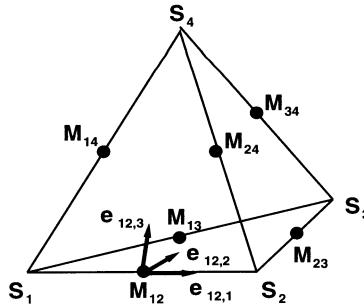


Fig. 13.11. The locations of the degrees of freedom and the three degrees of freedom at the point M_{12} for $\tilde{\mathcal{S}}_1$

Now, let us set $I = \{(\{\ell, m\}, n)\}$ such that $(\ell, m) \in \{1, 2, 3, 4\}^2, \ell \neq m, n \neq \ell, n \neq m\}$. We can define the following spaces of polynomial vector valued functions:

$$\mathcal{S}_1 = \{u(x) = \mathbf{a} \times x + \mathbf{b}, (\mathbf{a}, \mathbf{b}) \in \mathbb{R}^3 \times \mathbb{R}^3\}, \quad (13.78)$$

$$\tilde{\mathcal{S}}_1 = \mathcal{S}_1 \oplus \{\mathbf{w}_{\ell m}^n\}_{(\{\ell, m\}, n) \in I}, \quad (13.79)$$

where

$$\mathbf{w}_{\ell m}^n = \lambda_\ell \lambda_m \nabla \lambda_n. \quad (13.80)$$

Let $\boldsymbol{\nu}_j$ be the unit outward normal to a face $\Phi_j = S_\ell S_m S_n$. If we set

$$\nabla_j \mathbf{w} = \boldsymbol{\nu}_j \times (\nabla \mathbf{w} \times \boldsymbol{\nu}_j), \quad (13.81)$$

which is actually the gradient of the restriction of \mathbf{w} to Φ_j , we have

$$\boldsymbol{\nu}_j \times (\mathbf{w}_{\ell m}^n \times \boldsymbol{\nu}_j)|_{\Phi_j} = \mathbf{t}_{\ell m}^n = \lambda_\ell \lambda_m \nabla_j \lambda_j. \quad (13.82)$$

Let

$$\mathcal{R}_1(\Phi_j) = \{\boldsymbol{\nu}_j \times (\nabla \mathbf{u} \times \boldsymbol{\nu}_j), \mathbf{u} \in \mathcal{S}_1\} \quad (13.83)$$

be the space of the tangential traces of the functions of \mathcal{S}_1 and

$$\tilde{\mathcal{R}}_1(\Phi_j) = \mathcal{R}_1(\Phi_j) \oplus \{\mathbf{t}_{\ell m}^n, \mathbf{t}_{n\ell}^m, \mathbf{t}_{mn}^\ell\}. \quad (13.84)$$

Equation (13.82) implies that $\tilde{\mathcal{R}}_1(\Phi_j)$ is the space of the tangential traces of the functions of $\tilde{\mathcal{S}}_1$. One can easily see that $\tilde{\mathcal{R}}_1(\Phi_j)$ is isomorphic to the space \mathcal{R}_1 defined in (13.69), which shows that the restriction to a face of the space of approximation is isomorphic to a subspace of the 2D $H(\text{curl}, \Omega)$ space.

Of course, the corresponding quadrature formula is

$$\int_T f(\mathbf{x}) \, d\mathbf{x} \simeq \frac{\text{mes}(T)}{6} \sum_{(\ell,m) \in \mathcal{I}} f(M_{\ell m}), \quad (13.85)$$

where $\mathcal{I} = \{(p, q) \in \{1, 2, 3, 4\}^2 \text{ such that } p < q\}$.

The Second-Order Element. Paradoxically, the second-order element is easier to define than the first-order one. Its degrees of freedom are the three components of a function in the basis composed of the three unit vectors at each vertex and, on the other hand, its three components on an orthonormal basis at the center G_j of a face Φ_j , composed of the unit outward normal at this point and two arbitrary orthogonal unit vectors parallel to the face (Fig. 13.12). The space of polynomials for which this set of degrees of freedom is unisolvant can be constructed as follows:

Let \mathcal{V}_2 be the eight-dimensional space generated by the vectors

$$\begin{pmatrix} x_2^2 \\ -x_1 x_2 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ -x_2 x_3 \\ x_2^2 \end{pmatrix}, \begin{pmatrix} -x_1 x_2 \\ x_1^2 \\ 0 \end{pmatrix}, \begin{pmatrix} -x_2 x_3 \\ 0 \\ x_1^2 \end{pmatrix},$$

$$\begin{pmatrix} x_3^2 \\ 0 \\ -x_1x_3 \end{pmatrix}, \begin{pmatrix} 0 \\ x_3^2 \\ -x_2x_3 \end{pmatrix}, \begin{pmatrix} x_2x_3 \\ -x_1x_3 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ x_1x_3 \\ -x_1x_2 \end{pmatrix}$$

and

$$\mathcal{S}_2 = [P_3(T)]^3 \oplus \mathcal{V}_2. \quad (13.86)$$

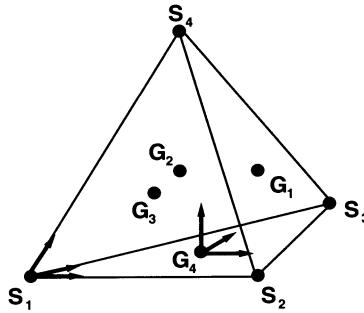


Fig. 13.12. The locations of the degrees of freedom and the three degrees of freedom at the points S_1 and G_4 for $\tilde{\mathcal{S}}_2$

The polynomial space $\tilde{\mathcal{S}}_2$ of dimension 24, for which the set of degrees of freedom is unisolvant, is

$$\tilde{\mathcal{S}}_2 = \mathcal{S}_2 \oplus \{\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \mathbf{w}_4\}, \quad (13.87)$$

where

$$\mathbf{w}_p = \lambda_\ell \lambda_m \lambda_n \nabla \lambda_p$$

and $\{\ell, m, n, p\}$ is a circular permutation of $\{1, 2, 3, 4\}$.

The corresponding quadrature formula is

$$\int_T f(\mathbf{x}) d\mathbf{x} \simeq \text{mes}(T) \left[\frac{3}{80} \sum_{p=1}^4 f(G_p) + \frac{1}{720} \sum_{q=1}^4 f(S_q) \right]. \quad (13.88)$$

One could a priori expect to obtain two points per edge with three degrees of freedom at each point. The quadrature formula was actually sought for this arrangement of points but the unique computed result was that the two points on the edges were moved to its ends. This element (also called a *vertex element*) is a generalization of the second kind of Nédélec's edge elements [93] also given by Mur [91]. A similar element can be constructed in 2D.

The stability conditions for a regular mesh in a homogeneous isotropic medium such that $c = \sqrt{\varepsilon\mu}$ are, for a leapfrog scheme on a regular mesh composed of cubic cells divided into six tetrahedra (which provides 43 classes of degrees of freedom for $\tilde{\mathcal{S}}_1$ and 62 for $\tilde{\mathcal{S}}_2$) [50]:

- $c\Delta t/h \leq 0.18$ for the first-order element (versus 0.5773 for quadrilateral elements),
- $c\Delta t/h \leq 0.21$ for the second-order element (versus 0.2356 for quadrilateral elements).

One can notice that the second-order element has a surprisingly large stability condition.

The numerical dispersion of all these elements was studied in a semi-numerical way in 2D and 3D [50]. The first-order elements have a dispersion error $O(h^2)$ and the second-order one is $O(h^4)$.

13.3.3 Spaces of Approximation

Construction Based on Elements of Any Shape. In order to construct the spaces of approximation of the solution in a domain $\Omega \subset \mathbb{R}^d$ ($d = 2, 3$) meshed by a mesh \mathcal{T} , we shall denote $\tilde{\mathcal{R}}_r^T$ and $\tilde{\mathcal{S}}_r^T$, $r = 1, 2$, the above polynomial spaces corresponding to a triangle or a tetrahedron T . With these notations, one can define the following spaces of approximation

– in 2D:

$$\mathbf{X}_{h0}^r = \left\{ \mathbf{v}_h \in H_0(\mathbf{curl}, \Omega) \text{ such that } \forall T_j \in \mathcal{T}, \mathbf{v}_h|_{T_j} \in \tilde{\mathcal{R}}_r^T \right\}, \quad (13.89)$$

– in 3D:

$$\mathbf{X}_{h0}^r = \left\{ \mathbf{v}_h \in H_0(\mathbf{curl}, \Omega) \text{ such that } \forall T_j \in \mathcal{T}, \mathbf{v}_h|_{T_j} \in \tilde{\mathcal{S}}_r^T \right\}. \quad (13.90)$$

On the basis of these spaces, one can write the following approximated problem in 2D:

Find \mathbf{E}_h such that $\mathbf{E}_h(., t) \in \mathbf{X}_{h0}^r$ and

$$\frac{d^2}{dt^2} \int_{\Omega} \underline{\varepsilon} \mathbf{E}_h \cdot \boldsymbol{\varphi}_h \, dx - \int_{\Omega} \frac{1}{\mu} \mathbf{curl} \mathbf{E}_h \mathbf{curl} \boldsymbol{\varphi}_h \, dx = - \int_{\Omega} \mathbf{j} \cdot \boldsymbol{\varphi}_h \, dx, \quad (13.91)$$

$$\forall \boldsymbol{\varphi}_h \in \mathbf{X}_{h0}^r$$

and, in 3D:

Find \mathbf{E}_h such that $\mathbf{E}_h(., t) \in \mathbf{X}_{h0}^r$ and

$$\begin{aligned} \frac{d^2}{dt^2} \int_{\Omega} \underline{\varepsilon} \mathbf{E}_h \cdot \boldsymbol{\varphi}_h \, d\mathbf{x} - \int_{\Omega} \underline{\mu}^{-1} \nabla \times \mathbf{E}_h \cdot \nabla \times \boldsymbol{\varphi}_h \, d\mathbf{x} = \\ - \int_{\Omega} \mathbf{j} \cdot \boldsymbol{\varphi}_h \, d\mathbf{x}, \quad \forall \boldsymbol{\varphi}_h \in \mathbf{X}_{h0}^r. \end{aligned} \quad (13.92)$$

Construction Based on the Unit Element. From the computational point of view (and also in order to obtain error estimates), it is useful to derive the approximation from the unit element \widehat{T} whose vertices are $(0, 0)$, $(1, 0)$ and $(0, 1)$ in 2D and $(0, 0, 0)$, $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$ in 3D. As for quadrilaterals and hexahedra, this construction is based on the $H(\text{curl})$ -conforming transform and the spaces of approximation are

– In 2D:

$$\begin{aligned} \tilde{\mathbf{X}}_{h0}^r &= \{ \mathbf{v}_h \in H_0(\text{curl}, \Omega) \\ &\quad \text{such that } \forall T_j \in \mathcal{T}, DF_j^* \mathbf{v}_h|_{T_j} \circ \mathbf{F}_j \in \tilde{\mathcal{R}}_r^{\widehat{T}} \}. \end{aligned} \quad (13.93)$$

– In 3D:

$$\begin{aligned} \tilde{\mathbf{X}}_{h0}^r &= \{ \mathbf{v}_h \in H_0(\text{curl}, \Omega) \\ &\quad \text{such that } \forall T_j \in \mathcal{T}, DF_j^* \mathbf{v}_h|_{T_j} \circ \mathbf{F}_j \in \tilde{\mathcal{S}}_r^{\widehat{T}} \}. \end{aligned} \quad (13.94)$$

For the first definition, the basis functions are parallel or orthogonal to the edges or to the faces and those parallel to the edges or the faces are continuous. As for quadrilateral and hexahedral elements, in the second definition these basis functions can have any direction and only their tangential components are continuous.

Remarks

1. Of course, one can easily construct similar $H(\text{div})$ elements.
2. Other points of view for solving the Maxwell equations were developed in [8, 26, 63, 104]. The first one was motivated by the study of the behavior of a flow of particles in an electromagnetic field. In this case, the use of edge elements raises some difficulties which justify the use of continuous finite elements. Of course, additional conditions ensure the well-posedness of the problem. Anyhow, this kind of approach was the source of interesting problems [7]. The second point of view is derived from methods initially applied to fluid mechanics but leads to uncentered approximations which, of course, introduce numerical dissipation. The third approach is based on diagonalization of the linear forms derived from the mass integral but leads to conditionally stable approximations.

13.4 A New Formulation of Spectral Elements

13.4.1 A New Approximation of the Wave Equation

The dramatical gain of storage and CPU time induced by the new mixed formulation described in Sect. 13.2 led us to construct a similar approach for the wave equation in the following way:

Let us consider the formulation of the wave equation (on a domain Ω with Dirichlet boundary conditions) as a first-order system:

$$\eta(\mathbf{x}) \frac{\partial u}{\partial t}(\mathbf{x}, t) = \nabla \cdot \mathbf{v}(\mathbf{x}, t) + F(\mathbf{x}, t), \quad (13.95a)$$

$$\frac{\partial \mathbf{v}}{\partial t}(\mathbf{x}, t) = \gamma(\mathbf{x}) \nabla u(\mathbf{x}, t), \quad (13.95b)$$

whose variational formulation can be written as:

Find u and \mathbf{v} such that $u(., t) \in H_0^1(\Omega)$ and $\mathbf{v}(., t) \in [L^2(\Omega)]^d$ ($d = 2, 3$) and

$$\frac{d}{dt} \int_{\Omega} \eta u \varphi \, d\mathbf{x} = - \int_{\Omega} \mathbf{v} \cdot \nabla \varphi \, d\mathbf{x} + \int_{\Omega} F \varphi \, d\mathbf{x} \quad \forall \varphi \in H_0^1(\Omega), \quad (13.96a)$$

$$\frac{d}{dt} \int_{\Omega} \gamma^{-1} \mathbf{v} \cdot \psi \, d\mathbf{x} = \int_{\Omega} \nabla u \cdot \psi \, d\mathbf{x} \quad \forall \psi \in [L^2(\Omega)]^d. \quad (13.96b)$$

In this formulation, the divergence of \mathbf{v} is involved and so the natural way to obtain an approximation on a mesh \mathcal{Q} , composed of quadrilaterals in 2D and of hexahedra in 3D, similar to that of the Maxwell equations is given by the following approximate formulation:

Find u_h and \mathbf{v}_h such that $u_h(., t) \in U_h^r(\Omega)$ and $\mathbf{v}_h(., t) \in \tilde{\mathbf{Y}}_{h0}^r$ and

$$\frac{d}{dt} \int_{\Omega} \eta u_h \varphi_h \, d\mathbf{x} = - \int_{\Omega} \mathbf{v}_h \cdot \nabla \varphi_h \, d\mathbf{x} + \int_{\Omega} F \varphi_h \, d\mathbf{x} \quad \forall \varphi_h \in U_h^r(\Omega), \quad (13.97a)$$

$$\frac{d}{dt} \int_{\Omega} \gamma^{-1} \mathbf{v}_h \cdot \psi_h \, d\mathbf{x} = \int_{\Omega} \nabla u_h \cdot \psi_h \, d\mathbf{x} \quad \forall \psi_h \in \tilde{\mathbf{Y}}_{h0}^r, \quad (13.97b)$$

where $U_h^r(\Omega)$ is defined as in (12.16) and

$$\begin{aligned} \tilde{\mathbf{Y}}_h^r &= \{\mathbf{v}_h \in [L^2(\Omega)]^2 \\ &\text{such that } \forall K_j \in \mathcal{Q}, |J_j| D F_j^{-1} \mathbf{v}_{h|_{K_j}} \circ \mathbf{F}_j \in [Q_r]^d\}. \end{aligned} \quad (13.98)$$

For any function $\varphi_h \in U_h^r(\Omega)$ such that $\varphi_{h|_{K_j}} \circ \mathbf{F}_j = \hat{\varphi}_h$ and any $\psi_h \in \tilde{\mathbf{Y}}_{h0}^r$ such that $\psi_{h|_{K_j}} \circ \mathbf{F}_j = 1/|J_j| D F_j \hat{\psi}_h$, we have, in the same way as in (13.61):

$$\begin{aligned}
\int_{K_j} \boldsymbol{\psi}_h \cdot \nabla \varphi_h \, dx &= \int_{\hat{K}} |J_j| \boldsymbol{\psi}_h \circ \mathbf{F}_j \cdot \nabla \varphi_h \circ \mathbf{F}_j \, d\hat{x} = \\
\int_{\hat{K}} \frac{|J_j|}{|J_j|} DF_j \hat{\boldsymbol{\psi}}_h \cdot DF_j^{*-1} \hat{\nabla} \hat{\varphi}_h \, d\hat{x} &= \\
\int_{\hat{K}} DF_j^{-1} DF_j \hat{\boldsymbol{\psi}}_h \cdot \hat{\nabla} \hat{\varphi}_h \, d\hat{x} &= \int_{\hat{K}} \hat{\boldsymbol{\psi}}_h \cdot \hat{\nabla} \hat{\varphi}_h \, d\hat{x}.
\end{aligned} \tag{13.99}$$

By taking, as previously, the basis functions on \hat{K} at the Gauss-Lobatto points for $U_h^r(\Omega)$ and $\tilde{\mathbf{Y}}_{h0}^r$, we obtain the following discrete system:

$$D_h \frac{d\mathbf{U}}{dt} = -R_h \mathbf{V} + \mathbf{F}_h, \tag{13.100a}$$

$$B_h \frac{d\mathbf{V}}{dt} = R_h^* \mathbf{U}, \tag{13.100b}$$

where \mathbf{U} is the vector of the components of u_h on the basis of $U_h^r(\Omega)$ and \mathbf{V} , the vector of the components of v_h on the basis of $\tilde{\mathbf{Y}}_{h0}^r$ and R_h^* the transposed matrix of R_h .

In this system, D_h is a diagonal matrix and B_h a $d \times d$ block-diagonal matrix which contain the whole geometry and physics, whereas, thanks to (13.99), R_h is a stiffness matrix computed only on \hat{K} .

The gain of storage for this approximation versus the spectral element approximation formulated as in the previous chapter is represented in Fig. 13.13.

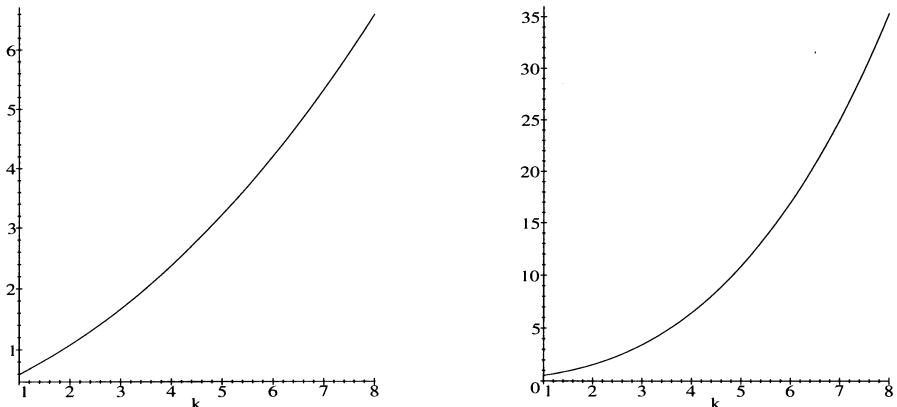


Fig. 13.13. The ratios between the storages required by classical spectral FEM and mixed FEM in 2D (left) and 3D (right). The abscissae represent the order of the polynomial approximation. One can notice that the new method is more expensive for the first-order approximation and that the gain increases with the order and the dimension

In these curves, we take into account both matrices and variables. Actually, the storage of the whole stiffness matrix for spectral elements is in $O(r^4)$ in 2D and in $O(r^6)$ in 3D whereas the new approach provides a storage in $O(r^2)$ in 2D and in $O(r^3)$ in 3D.

On the other hand, the matrix R_h is very sparse and this leads to a substantial gain of CPU time. This gain was evaluated in terms of multiplications and additions and the comparison between the algorithm provided by the spectral approximation and the new formulation is represented in Fig. 13.14. Following this evaluation, the algorithms are in $O(r^4)$ in 2D and in $O(r^6)$ in 3D for spectral elements and the new approach provides a cost in $O(r^3)$ in 2D and in $O(r^4)$ in 3D. Actually, the gain of time is even better. For instance, we obtained, in 2D, a ratio of 1.1 for Q_3 and 1.9 for Q_5 which is about 25% better than expected. This additional gain is due to the computation of array addresses, since the arrays are larger for classical elements.

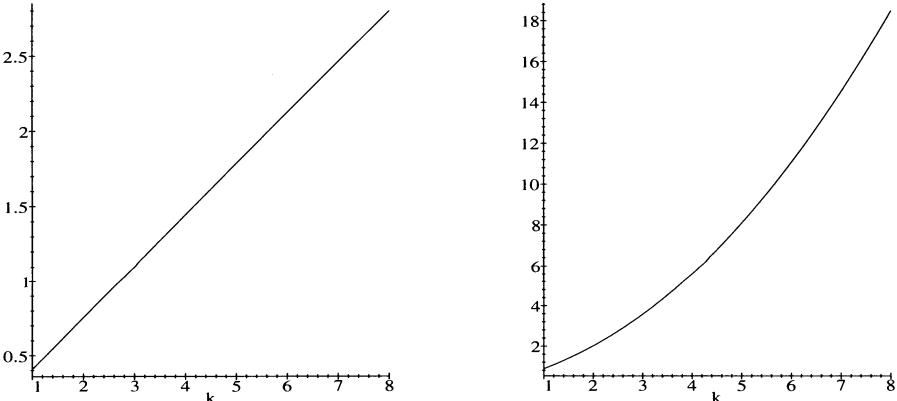


Fig. 13.14. The ratios between the computations required by classical spectral FEM and mixed FEM in 2D (left) and 3D (right). The abscissae represent the order of the polynomial approximation. The new method seems to be more expensive for the first-order approximation in 3D and for the first and second-order approximations in 2D

All the above features of the method show its undisputed superiority on spectral elements. The only question that one can ask is: Is it as accurate as spectral elements? This question will be answered in the following section.

13.4.2 A Theorem of Equivalence

Theorem 5. *If B_h , R_h and R_h^* are the matrices defined in (13.100a) and (13.100b) and K_h is the stiffness matrix of the spectral element method defined by (12.34), we have the factorization:*

$$K_h = R_h B_h^{-1} R_h^* \quad (13.101)$$

which ensures that

$$\mathbf{U} = \tilde{\mathbf{U}}, \quad (13.102)$$

where \mathbf{U} is the solution of (13.100a) and (13.100b) and $\tilde{\mathbf{U}}$ the solution given by the spectral element method [31].

Proof. Let us first give some definitions

- $\tau = \{K_j\}_{j=1,\dots,N_e}$, $K_j \subset \Omega$
- $\forall j = 1, \dots, N_d$, where N_d is the number of degrees of freedom in $\Omega - \partial\Omega$, we set

$$\begin{aligned} S_j &= \{i \text{ such that } K_i \subset \text{Supp}(\varphi_j)\} \\ &= \{ \text{index of all the elements in the support of } \varphi_j \} \end{aligned} \quad (13.103)$$

and $\forall i = 1, \dots, N_e$, the functions θ_i and ζ_i such that :

$$\forall j = 1, \dots, N_d, \theta_i(j) = p \text{ and } \zeta_i(p) = j \text{ with } \mathbf{F}_i(\hat{\xi}_p) = x_j. \quad (13.104)$$

- $\tilde{\mathbf{Y}}_h^r$ is defined in a more general framework which will provide a better understanding of the theorem:

$$\tilde{\mathbf{Y}}_h^r = \left\{ \mathbf{w} \in (L^2(\Omega))^d \text{ such that } (P_i^{-1}\mathbf{w})_{|_{K_i}} \circ \mathbf{F}_i \in \left[Q_r(\hat{K})\right]^d \right\}, \quad (13.105)$$

where, $\forall i$, P_i is an invertible $d \times d$ -matrix.

- Let $(\varphi_\ell)_{\ell=1,\dots,N_d}$ be the basis functions of $U_h^r(\Omega)$, $(\hat{\varphi}_j)_{j=1,\dots,\nu}$ the basis functions of $Q_r(\hat{K})$ and $(\hat{\psi}_{j,l})_{j=1,\dots,\nu; l=1,\dots,N}$ those of $\left[Q_r(\hat{K})\right]^d$.

From these functions, we define φ_j^i and $\psi_{j,l}^i$ such that $\varphi_j^i = \varphi_{\zeta_i(j)}|_{K_i}$, $\varphi_j^i \circ \mathbf{F}_i = \hat{\varphi}_j$ and $\psi_{j,l}^i \circ \mathbf{F}_i = P_i \hat{\psi}_{j,l}$ with:

$$K_i = \text{Supp}(\varphi_j^i) = \text{Supp}(\psi_{j,l}^i).$$

- With this notation, we obtain:

$$u_h = \sum_{i=1}^{N_d} u_i \varphi_i = \sum_{i=1}^{N_d} \sum_{j=1}^{\nu} u_{\zeta_i(j)} \varphi_j^i, \quad (13.106)$$

$$\mathbf{v}_h = \sum_{i=1}^{N_d} \sum_{j=1}^{\nu} \sum_{l=1}^d v_{j,l}^i \psi_{j,l}^i, \quad (13.107)$$

where ν is the number of points of interpolation in one element.

Remarks

1. This notation takes into account the fact that \mathbf{v}_h is discontinuous and that u_h vanishes on the boundary of Ω .
 2. In order to simplify the proof, we shall set $\hat{\psi}_{p,l}$ such that:
- $\forall \hat{\xi}_q, q = 1, \dots, \nu, \hat{\psi}_{j,l}(\hat{\xi}_q) = \delta_{jq} \mathbf{e}_l$ where $(\mathbf{e}_1, \dots, \mathbf{e}_d)$ is the canonical basis of \mathbb{R}^d and δ_{jq} is the Kronecker symbol.

– Decomposition of $\int_{\Omega} \mathbf{v}_h \cdot \psi_{p,m}^i \, d\mathbf{x}$:

$$\int_{\Omega} \mathbf{v}_h \cdot \psi_{p,m}^i \, d\mathbf{x} = \sum_{\ell=1}^{N_e} \int_{K_{\ell}} \mathbf{v}_h \cdot \psi_{p,m}^i \, d\mathbf{x}. \quad (13.108)$$

By using (13.107), we obtain:

$$\int_{\Omega} \mathbf{v}_h \cdot \psi_{p,m}^i \, d\mathbf{x} = \int_{K_i} \sum_{j=1}^{\nu} \sum_{l=1}^d v_{j,l}^i \psi_{j,l}^i \cdot \psi_{p,m}^i \, d\mathbf{x}. \quad (13.109)$$

Let us set: $\mathbf{x} = \mathbf{F}_i(\hat{\mathbf{x}})$.

$$\int_{\Omega} \mathbf{v}_h \cdot \psi_{p,m}^i \, d\mathbf{x} = \sum_{j=1}^{\nu} \sum_{l=1}^d v_{j,l}^i \int_{\hat{K}} |J_i| P_i \hat{\psi}_{j,l} \cdot P_i \hat{\psi}_{p,m} \, d\hat{\mathbf{x}}. \quad (13.110)$$

After transposition, we obtain:

$$\int_{\Omega} \mathbf{v}_h \cdot \psi_{p,m}^i \, d\mathbf{x} = \sum_{j=1}^{\nu} \sum_{l=1}^d v_{j,l}^i \int_{\hat{K}} |J_i| P_i^* P_i \hat{\psi}_{j,l} \cdot \hat{\psi}_{p,m} \, d\hat{\mathbf{x}}. \quad (13.111)$$

The use of the Gauss-Lobatto quadrature rule combined with the properties of orthogonality of the functions $\hat{\psi}_{p,m}$ leads to:

$$\begin{aligned} \int_{\Omega} \mathbf{v}_h \cdot \psi_{p,m}^i \, d\mathbf{x} &\simeq \sum_{j=1}^{\nu} \sum_{l=1}^d v_{j,l}^i \sum_{q=1}^{\nu} \omega_q |J_i(\hat{\xi}_q)| P_i^*(\hat{\xi}_q) \times \\ &\quad P_i(\hat{\xi}_q) \hat{\psi}_{j,l}(\hat{\xi}_q) \cdot \hat{\psi}_{p,m}(\hat{\xi}_q) \end{aligned}$$

which can be rewritten as

$$\begin{aligned} \int_{\Omega} \mathbf{v}_h \cdot \psi_{p,m}^i \, d\mathbf{x} &\simeq \sum_{j=1}^{\nu} \sum_{l=1}^d v_{j,l}^i \sum_{q=1}^{\nu} \omega_q |J_i(\hat{\xi}_q)| P_i^*(\hat{\xi}_q) \times \\ &\quad P_i(\hat{\xi}_q) \delta_{j,q} \mathbf{e}_l \cdot \delta_{p,q} \mathbf{e}_m \\ &\simeq \omega_p \sum_{l=1}^d v_{p,l}^i |J_i(\hat{\xi}_p)| P_i^*(\hat{\xi}_p) P_i(\hat{\xi}_p) \mathbf{e}_l \cdot \mathbf{e}_m. \end{aligned}$$

Remark

The above formula shows that the matrix B_h is $d \times d$ -block diagonal.

- Decomposition of $\int_{\Omega} \nabla u_h \cdot \psi_{p,m}^i \, dx$:

We use the same method as for $\int_{\Omega} v_h \cdot \psi_{p,m}^i \, dx$. This method provides:

$$\int_{\Omega} \nabla u_h \cdot \psi_{p,m}^i \, dx =$$

$$\sum_{j=1}^{\nu} u_{\zeta_i(j)} \int_{\hat{K}} |J_i| P_i^* DF_i^{*-1}(\hat{\xi}_j) \hat{\nabla} \hat{\varphi}_j \cdot \hat{\psi}_{p,m} \, d\hat{\xi} \simeq$$

$$\sum_{j=1}^{\nu} u_{\zeta_i(j)} \sum_{q=1}^{\nu} \omega_q |J_i(\hat{\xi}_q)| P_i^*(\hat{\xi}_q) DF_i^{*-1}(\hat{\xi}_q) \hat{\nabla} \hat{\varphi}_j(\hat{\xi}_q) \cdot \hat{\psi}_{p,m}(\hat{\xi}_q) \simeq$$

$$\sum_{j=1}^{\nu} u_{\zeta_i(j)} \omega_p |J_i(\hat{\xi}_p)| P_i^*(\hat{\xi}_p) DF_i^{*-1}(\hat{\xi}_p) \hat{\nabla} \hat{\varphi}_j(\hat{\xi}_p) \cdot e_m.$$

Finally, we obtain the following relation:

$$\begin{aligned} \sum_{l=1}^d v_{p,l}^i |J_i(\hat{\xi}_p)| P_i^*(\hat{\xi}_p) P_i(\hat{\xi}_p) e_l \cdot e_m \\ = \end{aligned} \tag{13.112}$$

$$\sum_{j=1}^{\nu} u_{\zeta_i(j)} |J_i(\hat{\xi}_p)| P_i^*(\hat{\xi}_p) DF_i^{*-1}(\hat{\xi}_p) \hat{\nabla} \hat{\varphi}_j(\hat{\xi}_p) \cdot e_m.$$

- Decomposition of $\int_{\Omega} v_h \cdot \nabla \varphi_n \, dx$:

We use the relation: $\varphi_n = \sum_{i \in S_n} \varphi_n|_{K_i} = \sum_{i \in S_n} \varphi_{\theta_i(n)}^i$.

The same method leads to:

$$\begin{aligned} \int_{\Omega} v_h \cdot \nabla \varphi_n \, dx &= \sum_{i \in S_n} \sum_{p=1}^{\nu} \sum_{l=1}^d v_{p,l}^i \int_{K_i} \psi_{p,l}^i \cdot \nabla \varphi_{\theta_i(n)}^i \, dx \\ &= \sum_{i \in S_n} \sum_{p=1}^{\nu} \sum_{l=1}^d v_{p,l}^i \int_{\hat{K}} |J_i| DF_i^{-1} P_i \hat{\psi}_{p,l} \cdot \hat{\nabla} \hat{\varphi}_{\theta_i(n)} \, d\hat{x} \\ &\simeq \sum_{i \in S_n} \sum_{p=1}^{\nu} \sum_{l=1}^d v_{p,l}^i \omega_p |J_i(\hat{\xi}_p)| DF_i^{-1}(\hat{\xi}_p) \times \\ &\quad P_i(\hat{\xi}_p) e_l \cdot \hat{\nabla} \hat{\varphi}_{\theta_i(n)}(\hat{\xi}_p). \end{aligned}$$

Now, we introduce $P_i^*(\hat{\xi}_p)$ $P_i(\hat{\xi}_p)$ in order to use (13.112). We can write:

$$\int_{\Omega} \mathbf{v}_h \cdot \nabla \varphi_n \, d\mathbf{x} \simeq \sum_{i \in S_n} \sum_{p=1}^{\nu} \omega_p DF_i^{-1}(\hat{\xi}_p) P_i^{*-1}(\hat{\xi}_p) \times \\ \sum_{l=1}^d v_{p,l}^i |J_i(\hat{\xi}_p)| P_i^*(\hat{\xi}_p) P_i(\hat{\xi}_p) e_l \cdot \hat{\nabla} \hat{\varphi}_{\theta_i(n)}(\hat{\xi}_p).$$

Let us set $\hat{\nabla} \hat{\varphi}_{\theta_i(n)}(\hat{\xi}_p) = \sum_{m=1}^d \frac{\partial \hat{\varphi}_{\theta_i(n)}}{\partial \hat{x}_m}(\hat{\xi}_p) e_m$. We obtain:

$$\int_{\Omega} \mathbf{v}_h \cdot \nabla \varphi_n \, d\mathbf{x} \simeq \sum_{i \in S_n} \sum_{p=1}^{\nu} \omega_p DF_i^{-1}(\hat{\xi}_p) P_i^{*-1}(\hat{\xi}_p) \sum_{m=1}^d \frac{\partial \hat{\varphi}_{\theta_i(n)}}{\partial \hat{x}_m}(\hat{\xi}_p) \times \\ \sum_{l=1}^d v_{p,l}^i |J_i(\hat{\xi}_p)| P_i^*(\hat{\xi}_p) P_i(\hat{\xi}_p) e_l \cdot e_m.$$

Now, let us set $A_i^p = |J_i(\hat{\xi}_p)| P_i^*(\hat{\xi}_p) DF_i^{*-1}(\hat{\xi}_p)$. By using (13.112), we obtain:

$$\int_{\Omega} \mathbf{v}_h \cdot \nabla \varphi_n \, d\mathbf{x} \simeq \sum_{i \in S_n} \sum_{p=1}^{\nu} \omega_p DF_i^{-1}(\hat{\xi}_p) P_i^{*-1}(\hat{\xi}_p) \times \\ \sum_{m=1}^d \frac{\partial \hat{\varphi}_{\theta_i(n)}}{\partial \hat{x}_m}(\hat{\xi}_p) \sum_{j=1}^{\nu} u_{\zeta_i(j)} A_i^p \hat{\nabla} \hat{\varphi}_j(\hat{\xi}_p) \cdot e_m.$$

By switching the sums in p and j and after simplication, we have:

$$\int_{\Omega} \mathbf{v}_h \cdot \nabla \varphi_n \, d\mathbf{x} \simeq \sum_{i \in S_n} \sum_{j=1}^{\nu} u_{\zeta_i(j)} \sum_{p=1}^{\nu} \omega_p |J_i(\hat{\xi}_p)| DF_i^{-1}(\hat{\xi}_p) \times \\ DF_i^{*-1}(\hat{\xi}_p) \hat{\nabla} \hat{\varphi}_{\theta_i(n)}(\hat{\xi}_p) \cdot \hat{\nabla} \hat{\varphi}_j(\hat{\xi}_p),$$

which is the approximation by the Gauss-Lobatto quadrature rule of

$$\int_{\hat{K}} |J_i| DF_i^{-1} DF_i^{*-1} \hat{\nabla} \hat{\varphi}_{\theta_i(n)} \cdot \hat{\nabla} \hat{\varphi}_j \, d\hat{\mathbf{x}}. \quad (13.113)$$

So, if we denote $\int_Q f(\mathbf{x}) \, d\mathbf{x}$ the approximation by a Gauss-Lobatto quadrature rule of $\int_Q f(\mathbf{x}) \, d\mathbf{x}$, we obtain:

$$\begin{aligned}
\int_{\Omega} \mathbf{v}_h \cdot \nabla \varphi_n \, d\mathbf{x} &= \sum_{i \in S_n} \sum_{j=1}^{\nu} u_{\zeta_i(j)} \int_{\hat{K}} |J_i| DF_i^{-1} DF_i^{*-1} \hat{\nabla} \hat{\varphi}_j \cdot \hat{\nabla} \hat{\varphi}_{\theta_i(n)} \, d\hat{\mathbf{x}} \\
&= \sum_{i \in S_n} \sum_{j=1}^{\nu} u_{\zeta_i(j)} \int_{\hat{K}} |J_i| DF_i^{*-1} \hat{\nabla} \hat{\varphi}_j \cdot DF_i^{*-1} \hat{\nabla} \hat{\varphi}_{\theta_i(n)} \, d\hat{\mathbf{x}} \\
&= \sum_{i \in S_n} \sum_{j=1}^{\nu} u_{\zeta_i(j)} \int_{K_i} \nabla \varphi_j^i \cdot \nabla \varphi_{\theta_i(n)}^i \, d\mathbf{x},
\end{aligned}$$

which finally provides:

$$\begin{aligned}
\int_{\Omega} \mathbf{v}_h \cdot \nabla \varphi_n \, d\mathbf{x} &= \sum_{i \in S_n} \int_{K_i} \nabla u_h|_{K_i} \cdot \nabla \varphi_n^i|_{K_i} \, d\mathbf{x} \\
&= \int_{\Omega} \nabla u_h \cdot \nabla \varphi_n \, d\mathbf{x}.
\end{aligned} \tag{13.114}$$

So, by taking for φ_n the elements of \mathcal{B}_U , we obtain (13.101). \diamond

Remarks:

1. For reasons of simplicity, this proof was performed when $\eta = \gamma = 1$ but it remains the same when these parameters depend on \mathbf{x} .
2. The definition (13.105) of $\tilde{\mathbf{Y}}_h^r$ shows that this theorem holds for any approximation of $[L^2(\Omega)]^d$.
3. Actually, this proof holds for any quadrature rule, as soon as the quadrature points coincide with the degrees of freedom.
4. The factorization given in (13.101) can be interpreted as a discrete expression of:

$$\begin{cases} \eta \frac{\partial^2 u}{\partial t^2} = \nabla \cdot \mathbf{v} + f & \text{in } \Omega \times [0, T], \\ \gamma^{-1} \mathbf{v} = \nabla u & \text{in } \Omega \times [0, T]. \end{cases} \tag{13.115}$$

Actually, this formulation can also be used for implementing the method but requires a little more storage. The approximation of this formulation provides an algorithm close to that given in [84].

13.4.3 Extension to the Elastics System

The general formulation of the elastics system given in Sect. 1.3.1 is a second-order system which must be transformed into a first-order one to use a mixed formulation. Several first-order systems can be derived from (1.24a) and (1.24b) but, in our case, only one provides the appropriate formulation

[33, 55]. We now describe its construction in 2D in order to simplify the description.

The tensor c given in Sect. 1.3.1 can be expressed in the following form:

$$\check{C} = \begin{pmatrix} c_{1,1,1,1} & c_{1,1,1,2} & c_{1,1,2,1} & c_{1,1,2,2} \\ c_{1,2,1,1} & c_{1,2,1,2} & c_{1,2,2,1} & c_{1,2,2,2} \\ c_{2,1,1,1} & c_{2,1,1,2} & c_{2,1,2,1} & c_{2,1,2,2} \\ c_{2,2,1,1} & c_{2,2,1,2} & c_{2,2,2,1} & c_{2,2,2,2} \end{pmatrix}. \quad (13.116)$$

Now,

- $c_{i,j,k,l} = c_{l,i,j}$ implies that \check{C} is symmetric.
- $c_{i,j,k,l} = c_{j,i,k,l}$ implies that the second and third rows (and also columns) of \check{C} are the same.

Let m be a function from $\{1, 2\}^2$ to $\{1, 2, 3\}$ such that $m(1, 1) = 1$, $m(2, 2) = 2$, $m(1, 2) = m(2, 1) = 3$. We set

$$\hat{c}_{m(i,j)m(k,\ell)} = c_{ijk\ell}, \quad (13.117)$$

so that we can define the matrix

$$\hat{C} = \begin{pmatrix} \hat{c}_{1,1} & \hat{c}_{1,2} & \hat{c}_{1,3} \\ \hat{c}_{1,2} & \hat{c}_{2,2} & \hat{c}_{2,3} \\ \hat{c}_{3,1} & \hat{c}_{3,2} & \hat{c}_{3,3} \end{pmatrix}. \quad (13.118)$$

By using Hooke's law (Sect. 1.3.1), the stress tensor $\underline{\tau} = (\tau_{1,1}, \tau_{1,2}, \tau_{2,1}, \tau_{2,2})^T$ is related to the displacement vector $\mathbf{v} = (v_1, v_2)^T$ in the following way:

$$\tau_{1,1} = \hat{c}_{1,1} \frac{\partial v_1}{\partial x} + \hat{c}_{1,2} \frac{\partial v_2}{\partial y} + \hat{c}_{1,3} \left(\frac{\partial v_1}{\partial y} + \frac{\partial v_2}{\partial x} \right), \quad (13.119a)$$

$$\tau_{1,2} = \hat{c}_{1,3} \frac{\partial v_1}{\partial x} + \hat{c}_{2,3} \frac{\partial v_2}{\partial y} + \hat{c}_{3,3} \left(\frac{\partial v_1}{\partial y} + \frac{\partial v_2}{\partial x} \right), \quad (13.119b)$$

$$\tau_{2,1} = \hat{c}_{1,3} \frac{\partial v_1}{\partial x} + \hat{c}_{2,3} \frac{\partial v_2}{\partial y} + \hat{c}_{3,3} \left(\frac{\partial v_1}{\partial y} + \frac{\partial v_2}{\partial x} \right), \quad (13.119c)$$

$$\tau_{2,2} = \hat{c}_{1,2} \frac{\partial v_1}{\partial x} + \hat{c}_{2,2} \frac{\partial v_2}{\partial y} + \hat{c}_{2,3} \left(\frac{\partial v_1}{\partial y} + \frac{\partial v_2}{\partial x} \right). \quad (13.119d)$$

Now, by using the fact that

$$\nabla \cdot \left[\begin{pmatrix} a & b \\ c & d \end{pmatrix} \nabla u \right] = \frac{\partial}{\partial x} \left[a \frac{\partial u}{\partial x} + b \frac{\partial u}{\partial y} \right] + \frac{\partial}{\partial y} \left[c \frac{\partial u}{\partial x} + d \frac{\partial u}{\partial y} \right], \quad (13.120)$$

we obtain:

$$\mathbf{div}_{\underline{\underline{\tau}}} = \begin{pmatrix} \nabla \cdot [A_1 \nabla v_1] + \nabla \cdot [A_2 \nabla v_2] \\ \nabla \cdot [A_3 \nabla v_1] + \nabla \cdot [A_4 \nabla v_2] \end{pmatrix}, \quad (13.121)$$

where

$$A_1 = \begin{pmatrix} \hat{c}_{1,1} & \hat{c}_{1,3} \\ \hat{c}_{1,3} & \hat{c}_{3,3} \end{pmatrix}, \quad A_2 = \begin{pmatrix} \hat{c}_{1,3} & \hat{c}_{1,2} \\ \hat{c}_{3,3} & \hat{c}_{2,3} \end{pmatrix},$$

$$A_3 = \begin{pmatrix} \hat{c}_{1,3} & \hat{c}_{3,3} \\ \hat{c}_{1,2} & \hat{c}_{2,3} \end{pmatrix}, \quad A_4 = \begin{pmatrix} \hat{c}_{3,3} & \hat{c}_{2,3} \\ \hat{c}_{2,3} & \hat{c}_{2,2} \end{pmatrix}$$

and \mathbf{div} is defined as in Sect. 1.3.1.

One can notice that $A_1 = A_1^*$, $A_4 = A_4^*$, $A_2 = A_3^*$ and

$$\check{C} = \begin{pmatrix} A_1 & A_2 \\ A_3 & A_4 \end{pmatrix}.$$

So, by setting

$$\tau_{1,1} = \frac{\partial \sigma_{1,1}}{\partial t} + \frac{\partial \sigma_{2,1}}{\partial t}, \quad (13.122a)$$

$$\tau_{1,2} = \frac{\partial \sigma_{1,2}}{\partial t} + \frac{\partial \sigma_{2,2}}{\partial t}, \quad (13.122b)$$

$$\tau_{2,1} = \frac{\partial \sigma_{3,1}}{\partial t} + \frac{\partial \sigma_{4,1}}{\partial t}, \quad (13.122c)$$

$$\tau_{2,2} = \frac{\partial \sigma_{3,2}}{\partial t} + \frac{\partial \sigma_{4,2}}{\partial t}, \quad (13.122d)$$

and

$$\boldsymbol{\sigma}_j = \begin{pmatrix} \sigma_{j,1} \\ \sigma_{j,2} \end{pmatrix}, \quad j = 1..4, \quad (13.123)$$

the initial elastics system,

$$\rho \frac{\partial^2 \mathbf{v}}{\partial t^2} = \mathbf{div}_{\underline{\underline{\tau}}} + \mathbf{f}, \quad (13.124)$$

can be rewritten as

$$\rho \frac{\partial v_1}{\partial t} - \nabla \cdot \boldsymbol{\sigma}_1 - \nabla \cdot \boldsymbol{\sigma}_2 = f_1, \quad (13.125a)$$

$$\rho \frac{\partial v_2}{\partial t} - \nabla \cdot \boldsymbol{\sigma}_3 - \nabla \cdot \boldsymbol{\sigma}_4 = f_2, \quad (13.125b)$$

$$\frac{\partial \boldsymbol{\sigma}_1}{\partial t} = A_1 \nabla v_1, \quad (13.125c)$$

$$\frac{\partial \boldsymbol{\sigma}_2}{\partial t} = A_2 \nabla v_2, \quad (13.125d)$$

$$\frac{\partial \boldsymbol{\sigma}_3}{\partial t} = A_3 \nabla v_1, \quad (13.125e)$$

$$\frac{\partial \boldsymbol{\sigma}_4}{\partial t} = A_4 \nabla v_2. \quad (13.125f)$$

The edge element approximation defined for the wave equation can then be applied to the new formulation of the elastics system in the following way:

Find \mathbf{v}_h and $\boldsymbol{\sigma}_{jh}$, $j = 1..4$ such that $\mathbf{v}_h(., t) \in [U_h^r(\Omega)]^2$, $\boldsymbol{\sigma}_{jh}(., t) \in \tilde{\mathbf{Y}}_{h0}^r$ and

$$\begin{aligned} \frac{d}{dt} \int_{\Omega} \rho v_{1h} \varphi_h d\mathbf{x} &= - \int_{\Omega} \boldsymbol{\sigma}_{1h} \cdot \nabla \varphi_h d\mathbf{x} - \int_{\Omega} \boldsymbol{\sigma}_{2h} \cdot \nabla \varphi_h d\mathbf{x} \\ &\quad + \int_{\Omega} f_1 \varphi_h d\mathbf{x} \quad \forall \varphi_h \in U_h^r(\Omega), \end{aligned} \quad (13.126a)$$

$$\begin{aligned} \frac{d}{dt} \int_{\Omega} \rho v_{2h} \varphi_h d\mathbf{x} &= - \int_{\Omega} \boldsymbol{\sigma}_{3h} \cdot \nabla \varphi_h d\mathbf{x} - \int_{\Omega} \boldsymbol{\sigma}_{4h} \cdot \nabla \varphi_h d\mathbf{x} \\ &\quad + \int_{\Omega} f_2 \varphi_h d\mathbf{x} \quad \forall \varphi_h \in U_h^r(\Omega), \end{aligned} \quad (13.126b)$$

$$\frac{d}{dt} \int_{\Omega} A_1^{-1} \boldsymbol{\sigma}_{1h} \cdot \boldsymbol{\psi}_h d\mathbf{x} = \int_{\Omega} \nabla v_{1h} \cdot \boldsymbol{\psi}_h d\mathbf{x} \quad \forall \boldsymbol{\psi}_h \in \tilde{\mathbf{Y}}_{h0}^r, \quad (13.126c)$$

$$\frac{d}{dt} \int_{\Omega} A_2^{-1} \boldsymbol{\sigma}_{2h} \cdot \boldsymbol{\psi}_h d\mathbf{x} = \int_{\Omega} \nabla v_{2h} \cdot \boldsymbol{\psi}_h d\mathbf{x} \quad \forall \boldsymbol{\psi}_h \in \tilde{\mathbf{Y}}_{h0}^r, \quad (13.126d)$$

$$\frac{d}{dt} \int_{\Omega} A_3^{-1} \boldsymbol{\sigma}_{3h} \cdot \boldsymbol{\psi}_h d\mathbf{x} = \int_{\Omega} \nabla v_{1h} \cdot \boldsymbol{\psi}_h d\mathbf{x} \quad \forall \boldsymbol{\psi}_h \in \tilde{\mathbf{Y}}_{h0}^r, \quad (13.126e)$$

$$\frac{d}{dt} \int_{\Omega} A_4^{-1} \boldsymbol{\sigma}_{4h} \cdot \boldsymbol{\psi}_h d\mathbf{x} = \int_{\Omega} \nabla v_{2h} \cdot \boldsymbol{\psi}_h d\mathbf{x} \quad \forall \boldsymbol{\psi}_h \in \tilde{\mathbf{Y}}_{h0}^r. \quad (13.126f)$$

Remarks

1. The formulation (13.125a)–(13.125f) satisfies the energy identity and is therefore well-posed [55].
2. The free surface boundary condition can here be written as

$$(\boldsymbol{\sigma}_1 + \boldsymbol{\sigma}_2) \cdot \boldsymbol{\nu} = 0, \quad (13.127a)$$

$$(\boldsymbol{\sigma}_3 + \boldsymbol{\sigma}_4) \cdot \boldsymbol{\nu} = 0, \quad (13.127b)$$

where $\boldsymbol{\nu}$ is the unit outward normal to the boundary.

3. The 3D system can be constructed in the same way [55]. It involves nine elementary matrices A_ℓ . By using their symmetry properties, only six of them must be stored.
4. A priori, the above formulation seems to be legal only if the inverses of the matrices A_ℓ exist. In fact, the use of these inverses is an artifact introduced in order to obtain the variational formulation. From the algorithmic

point of view, these matrices are included in the blocks of the block mass matrices derived from (13.126c)–(13.126f). These blocks are actually proportional to the values of the matrix functions $DF_j^{-1}A_\ell^{-1}DF_j^{*-1}$ at the Gauss-Lobatto points. So, only their inverses $DF_j^*A_\ell DF_j$, which always exist, are stored and used.

5. A theorem of equivalence with the spectral element approximation can also be proved for the elastics system [55]. It seems that such a theorem can also be proved for the Maxwell equations. It would show the equivalence of the approximation defined in (13.60a) and (13.60b) with the approximation of the second-order formulation (curl-curl formulation) of the Maxwell equations obtained by using $\tilde{\mathbf{U}}_{h0}^r$.
6. The gain of storage and CPU over the spectral elements evaluated for the wave equation are roughly the same for the elastics system.
7. One could take, as in [17], $\mathbf{v}(.,t)$ in $[L^2(\Omega)]^2$ and $\boldsymbol{\sigma}_j(.,t)$ in $H(\text{div}, \Omega)$. Although more classical, this approach leads to an algorithm that is more difficult to implement and is not equivalent to the approximation by spectral elements. On the other hand, the choice of the quadrature points in [17] leads to very expensive algorithms.
8. All the approximations by finite elements can be used as variational finite difference schemes of regular grids. In this fall, second-order formulations in time of the equations provide more efficient methods. However, the use of first-order systems could be useful for modeling unbounded domains, as we shall see in the next chapter.

14. Modeling Unbounded Domains

14.1 History of the Problem

Although the physical world is finite, from a mathematical point of view, very large domains must often be modeled as infinite ones. This is particularly true for waves which can be propagated on distances which correspond to several hundreds or even several thousands of wavelengths. Generally, the subdomain of propagation which provides interesting results is much smaller than the whole domain. For instance, in a scattering experiment, the interesting part of the solution is often located around the obstacle which is a small part of the domain of propagation. On the other hand, the whole domain of propagation is generally too large to be taken into account by a numerical model on a computer.

All these considerations led researchers and engineers to look for ways to replace the whole domain of propagation by a more restricted one without destroying the solution. The first step in that direction was made by Engquist and Majda [53] who introduced the idea of *absorbing boundary conditions* (ABC) for the wave equation. The problem was to find boundary conditions for the restricted subdomain which would be able to absorb the waves when they reached them. The main difficulty of this approach came from the fact that the perfectly absorbing boundary condition or *transparent* condition is very easy to write in 1D but leads to a very complicated non-linear condition dependent on time and space in higher dimensions. So, in order to get reasonable conditions, one must approximate the transparent condition to a certain order. An additional difficulty comes from the fact that such an approximation based on Taylor expansions is unstable for order greater than 2 and so, for higher-order ABC, one must use Padé [12] approximations [53]. Since the approximation of the continuous transparent condition was understood, it became important to discretize them and to study their numerical behavior, i.e. their stability and accuracy. These problems and the extension of the ABC to other equations [18, 27, 90] were the source of numerous papers. One can see [59, 67] for instance for a survey of these papers.

Another direction of investigation was that of the *absorbing* or *damping layers*. This idea, which is more physical, consists in adding some layers around the domain in which one solves the wave equations with a damping term. This term should be able to reduce the amplitude of the wave during

its propagation in the layer so that one obtains almost no reflection in the effective domain. Various models of dampers were proposed (see for instance [21]) but all had the same and major drawback: At the entrance of the layer, the wave sees a contrast of impedances which generates non-negligible artificial reflected waves growing with the angle of incidence of the wave and which are very difficult to remove. For this reason, the ABC remained, for a long time, the main means of modeling unbounded domains. Then, in 1994, a French engineer, Jean-Pierre Bérenger, made the brilliant discovery of the *perfectly matched layer* (PML), i.e. a layer without reflection for any angle of propagation [19, 20]. The PML generated a true “gold rush” of both researcher and engineers. In fact, this miraculous damping layer raises many mathematical and numerical problems which were partially treated in several papers [1, 2, 40, 78, 85, 86, 123]. On the other hand, other studies were devoted to the extension of this technique, which was initially defined for the Maxwell system, to other models of wave equations [41, 55, 66, 112].

This chapter will be restricted to PML because, besides its modern character, this technique provides easy modeling of unbounded domains for all kinds of wave equations whereas the ABC do not work very well for elastics and, to a less extent, for the Maxwell equations. The main drawback of PML is that the question of their stability is still an open problem and for complex models (such as anisotropic media), some instabilities could appear [55]. Anyhow, this technique remains the most convenient way to model unbounded domains.

14.2 Perfectly Matched Layers

14.2.1 Presentation of the Method

Let us consider the first-order system formulation of the wave equation given in (13.95a) and (13.95b) in 2D with $\eta = \gamma = 1$ and $F = 0$:

$$\frac{\partial u}{\partial t} = \nabla \cdot \mathbf{v}, \quad (14.1a)$$

$$\frac{\partial \mathbf{v}}{\partial t} = \nabla u. \quad (14.1b)$$

The Bérenger’s formulation of this system, which is based on the decomposition of u into $u_x + u_y$, can be written as

$$\frac{\partial u_x}{\partial t} = \frac{\partial v_1}{\partial x_1}, \quad (14.2a)$$

$$\frac{\partial u_y}{\partial t} = \frac{\partial v_2}{\partial x_2}, \quad (14.2b)$$

$$\frac{\partial \mathbf{v}}{\partial t} = \nabla u. \quad (14.2c)$$

In a second, step, one adds a damping term multiplied by a continuous function $\zeta(x_1)$ such that $\zeta(x_1) = 0$ for $x_1 \leq 0$. We obtain:

$$\frac{\partial u_x}{\partial t} + \zeta u_x = \frac{\partial v_1}{\partial x_1}, \quad (14.3a)$$

$$\frac{\partial u_y}{\partial t} = \frac{\partial v_2}{\partial x_2}, \quad (14.3b)$$

$$\frac{\partial v_1}{\partial t} + \zeta v_1 = \frac{\partial u}{\partial x_1}, \quad (14.3c)$$

$$\frac{\partial v_2}{\partial t} = \frac{\partial u}{\partial x_2}. \quad (14.3d)$$

To understand these equations, we start by applying the Fourier transform in time to (14.3a) and (14.3d). We obtain the following equations:

$$(-i\omega + \zeta)\hat{u}_x = \frac{\partial \hat{v}_1}{\partial x_1}, \quad (14.4a)$$

$$-i\omega \hat{u}_y = \frac{\partial \hat{v}_2}{\partial x_2}, \quad (14.4b)$$

$$(-i\omega + \zeta)\hat{v}_1 = \frac{\partial \hat{u}}{\partial x_1}, \quad (14.4c)$$

$$-i\omega \hat{v}_2 = \frac{\partial \hat{u}}{\partial x_2}, \quad (14.4d)$$

which can be rewritten as

$$(i\omega)^2(\hat{u}_x + \hat{u}_y) + i\omega \frac{\partial \hat{v}_2}{\partial x_2} - \frac{(i\omega)^2}{-i\omega + \zeta} \frac{\partial \hat{v}_1}{\partial x_1} = 0 \quad (14.5)$$

or

$$\omega^2 \hat{u} + \frac{\partial^2 \hat{u}}{\partial x_2^2} + \frac{i\omega}{-i\omega + \zeta} \frac{\partial}{\partial x_1} \left(\frac{i\omega}{-i\omega + \zeta} \frac{\partial \hat{u}}{\partial x_1} \right) = 0. \quad (14.6)$$

By setting

$$\tilde{x}_1 = x_1 + \frac{i}{\omega} \int_0^{x_1} \zeta(s) ds, \quad (14.7)$$

(14.6) takes the form of the Helmholtz equation:

$$\omega^2 \hat{u} + \frac{\partial^2 \hat{u}}{\partial x_1^2} + \frac{\partial^2 \hat{u}}{\partial \tilde{x}_2^2} = 0. \quad (14.8)$$

By applying the Fourier transform in x_2 to (14.8), we obtain the following ODE:

$$\frac{\partial^2 \check{u}}{\partial \tilde{x}_1^2} + (\omega^2 - k_2^2) \check{u} = 0. \quad (14.9)$$

By setting $k_1 = \sqrt{\omega^2 - k_2^2}$, we obtain the equation

$$\frac{\partial^2 \check{u}}{\partial \tilde{x}_1^2} + k_1^2 \check{u} = 0, \quad (14.10)$$

whose solution can be written as

$$\check{u}(\tilde{x}_1) = A(k_2, \omega) e^{ik_1 \tilde{x}_1} + B(k_2, \omega) e^{-ik_1 \tilde{x}_1}. \quad (14.11)$$

After applying the inverse Fourier transform in x_2 , we obtain:

$$\begin{aligned} \hat{u}(\tilde{x}_1, x_2, \omega) &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} A(k_2, \omega) e^{ik_1 \tilde{x}_1} e^{ik_2 x_2} dk_2 \\ &+ \frac{1}{2\pi} \int_{-\infty}^{+\infty} B(k_2, \omega) e^{-ik_1 \tilde{x}_1} e^{ik_2 x_2} dk_2. \end{aligned} \quad (14.12)$$

Now, by replacing \tilde{x}_1 with its value given in (14.7), one can easily check that for both $k_2^2 < \omega^2$ or $k_2^2 > \omega^2$, the second integral of (14.12) provides an exponentially increasing wave which is not an acceptable solution of the problem. For this reason, we shall set $B = 0$ in the following.

From (14.11), we obtain:

$$A(k_2, \omega) = \check{u}(0) = \int_{-\infty}^{+\infty} \hat{u}(0, x_2, \omega) e^{-ik_2 x_2} dk_2, \quad (14.13)$$

so that, finally:

$$\hat{u}(\tilde{x}_1, x_2, \omega) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \check{u}(0) e^{ik_1 \tilde{x}_1} e^{ik_2 x_2} dk_2, \quad (14.14)$$

with $\check{u}(0)$ defined in (14.13).

Now, by using (14.7), we can rewrite (14.14) in the following form:

$$\hat{u}(\tilde{x}_1, x_2, \omega) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \check{u}(0) e^{i(k_1 x_1 + k_2 x_2)} e^{-k_1 \theta(x_1, \omega)} dk_2, \quad (14.15)$$

where $\theta(x_1, \omega) = \frac{1}{\omega} \int_0^{x_1} \zeta(s) ds$.

So, for $x_1 \leq 0$, the solution is exactly that of the wave equation and, for $x_1 > 0$, it is an evanescent wave exponentially decreasing and no reflection appears at the interface. This shows that the PML absorbs the incident wave without any reflection and that is true for any value of (k_1, k_2) , i.e. for any angle of incidence.

Remark

The change of variable (14.7) can be interpreted as the extension of the variable x_1 to a path of \mathbb{C} .

14.2.2 Construction of the PML in 2D

The previously defined change of variable is not simply a convenient mean of interpretation of Bérenger's PML, it is actually a powerful way to construct PMLs that are easy to implement. This approach was introduced by Chew and Weedon [22] (and can be found in another form in [123]) and we shall use its form given in [39].

The Wave Equation. Let us first introduce this approach on the simplest equation which is the wave equation [32]. We take the first-order system defined in (13.95a) and (13.95b) with $F = 0^1$ in $\Omega =]-\infty, 0[^2$:

$$\eta \frac{\partial u}{\partial t} = \nabla \cdot \mathbf{v}, \quad (14.16a)$$

$$\frac{\partial \mathbf{v}}{\partial t} = \gamma \nabla u. \quad (14.16b)$$

After applying the Fourier transform in time to (14.16a) and (14.16b), we obtain:

$$-\mathrm{i}\omega\eta\hat{u} - \frac{\partial\hat{v}_1}{\partial x_1} - \frac{\partial\hat{v}_2}{\partial x_2} = 0, \quad (14.17a)$$

$$-\mathrm{i}\omega\gamma^{-1}\hat{v}_1 - \frac{\partial\hat{u}}{\partial x_1} = 0, \quad (14.17b)$$

$$-\mathrm{i}\omega\gamma^{-1}\hat{v}_2 - \frac{\partial\hat{u}}{\partial x_2} = 0. \quad (14.17c)$$

¹ The fact of setting $F = 0$ is equivalent to the definition of an initial value of the unknowns.

Now, we introduce the change of variables

$$\tilde{x}_1 = \begin{cases} x_1 & \text{if } x_1 < 0, \\ x_1 + \frac{i}{\omega} \int_0^{x_1} \zeta_1(s) ds & \text{otherwise,} \end{cases} \quad (14.18a)$$

$$\tilde{x}_2 = \begin{cases} x_2 & \text{if } x_2 < 0, \\ x_2 + \frac{i}{\omega} \int_0^{x_2} \zeta_2(s) ds & \text{otherwise,} \end{cases} \quad (14.18b)$$

where, as in the previous section, $\zeta_1(x_1) = \zeta_2(x_2) = 0$ in Ω .

Then, we extend (14.17a)–(14.17c) to \mathbb{C} by using this change of variables. We obtain:

$$-i\omega\eta\hat{u} - \frac{\partial\hat{v}_1}{\partial\tilde{x}_1} - \frac{\partial\hat{v}_2}{\partial\tilde{x}_2} = 0, \quad (14.19a)$$

$$-i\omega\gamma^{-1}\hat{v}_1 - \frac{\partial\hat{u}}{\partial\tilde{x}_1} = 0, \quad (14.19b)$$

$$-i\omega\gamma^{-1}\hat{v}_2 - \frac{\partial\hat{u}}{\partial\tilde{x}_2} = 0. \quad (14.19c)$$

Now, we have:

$$d\hat{v}_1 = \frac{\partial\hat{v}_1}{\partial x_1} dx_1 + \frac{\partial\hat{v}_1}{\partial x_2} dx_2 = \frac{\partial\hat{v}_1}{\partial\tilde{x}_1} d\tilde{x}_1 + \frac{\partial\hat{v}_1}{\partial\tilde{x}_2} d\tilde{x}_2.$$

Since

$$d\tilde{x}_1 = \begin{cases} dx_1 & \text{if } x_1 < 0, \\ dx_1 + \frac{i\zeta_1}{\omega} dx_1 & \text{if } x_1 \geq 0 \end{cases}$$

and $\zeta_1(s) = 0$ for $s < 0$, we obtain the following relation:

$$\frac{\partial\hat{v}_1}{\partial x_1} = \left(1 + \frac{i\zeta_1}{\omega}\right) \frac{\partial\hat{v}_1}{\partial\tilde{x}_1}.$$

Therefore

$$\frac{\partial}{\partial\tilde{x}_1} = \frac{1}{1 + i\zeta_1/\omega} \frac{\partial}{\partial x_1}. \quad (14.20)$$

By the same process we obtain:

$$\frac{\partial}{\partial\tilde{x}_2} = \frac{1}{1 + i\zeta_2/\omega} \frac{\partial}{\partial x_2}. \quad (14.21)$$

So, (14.19a)–(14.19c) can be then rewritten as

$$-\mathrm{i}\omega\eta\hat{u} - \frac{1}{1 + \mathrm{i}\zeta_1/\omega} \frac{\partial\hat{v}_1}{\partial x_1} - \frac{1}{1 + \mathrm{i}\zeta_2/\omega} \frac{\partial\hat{v}_2}{\partial x_2} = 0, \quad (14.22a)$$

$$-\mathrm{i}\omega\gamma^{-1}\hat{v}_1 - \frac{1}{1 + \mathrm{i}\zeta_1/\omega} \frac{\partial\hat{u}}{\partial x_1} = 0, \quad (14.22b)$$

$$-\mathrm{i}\omega\gamma^{-1}\hat{v}_2 - \frac{1}{1 + \mathrm{i}\zeta_2/\omega} \frac{\partial\hat{u}}{\partial x_2} = 0, \quad (14.22c)$$

which can be set into the following form

$$-\left(1 + \frac{\mathrm{i}\zeta_1}{\omega}\right) \left(1 + \frac{\mathrm{i}\zeta_2}{\omega}\right) \mathrm{i}\omega\eta\hat{u} \quad (14.23a)$$

$$-\left(1 + \frac{\mathrm{i}\zeta_2}{\omega}\right) \frac{\partial\hat{v}_1}{\partial x_1} - \left(1 + \frac{\mathrm{i}\zeta_1}{\omega}\right) \frac{\partial\hat{v}_2}{\partial x_2} = 0,$$

$$-\left(1 + \frac{\mathrm{i}\zeta_1}{\omega}\right) \mathrm{i}\omega\gamma^{-1}\hat{v}_1 - \frac{\partial\hat{u}}{\partial x_1} = 0, \quad (14.23b)$$

$$-\left(1 + \frac{\mathrm{i}\zeta_2}{\omega}\right) \mathrm{i}\omega\gamma^{-1}\hat{v}_2 - \frac{\partial\hat{u}}{\partial x_2} = 0. \quad (14.23c)$$

By setting

$$\tilde{v}_1 = \left(1 + \frac{\mathrm{i}\zeta_2}{\omega}\right) \hat{v}_1, \quad (14.24a)$$

$$\tilde{v}_2 = \left(1 + \frac{\mathrm{i}\zeta_1}{\omega}\right) \hat{v}_2, \quad (14.24b)$$

we obtain:

$$-\left(1 + \frac{\mathrm{i}\zeta_1}{\omega}\right) \left(1 + \frac{\mathrm{i}\zeta_2}{\omega}\right) \mathrm{i}\omega\eta\hat{u} - \frac{\partial\tilde{v}_1}{\partial x_1} - \frac{\partial\tilde{v}_2}{\partial x_2} = 0, \quad (14.25a)$$

$$-\frac{1 + \mathrm{i}\zeta_1/\omega}{1 + \mathrm{i}\zeta_2/\omega} \mathrm{i}\omega\gamma^{-1}\tilde{v}_1 - \frac{\partial\hat{u}}{\partial x_1} = 0, \quad (14.25b)$$

$$-\frac{1 + \mathrm{i}\zeta_2/\omega}{1 + \mathrm{i}\zeta_1/\omega} \mathrm{i}\omega\gamma^{-1}\tilde{v}_2 - \frac{\partial\hat{u}}{\partial x_2} = 0. \quad (14.25c)$$

We can now define the following Bérenger-like variables:

$$\hat{v}_1^* = \frac{1 + i\zeta_1/\omega}{1 + i\zeta_2/\omega} \tilde{v}_1, \quad (14.26a)$$

$$\hat{v}_2^* = \frac{1 + i\zeta_2/\omega}{1 + i\zeta_1/\omega} \tilde{v}_2, \quad (14.26b)$$

$$\hat{u}^* = (1 + i\zeta_1/\omega)(1 + i\zeta_2/\omega)\hat{u}, \quad (14.26c)$$

which can be rewritten as

$$(-i\omega + \zeta_2)\hat{v}_1^* = (-i\omega + \zeta_1)\tilde{v}_1, \quad (14.27a)$$

$$(-i\omega + \zeta_1)\hat{v}_2^* = (-i\omega + \zeta_2)\tilde{v}_2, \quad (14.27b)$$

$$-\omega^2\hat{u}^* = (i\omega - \zeta_1)(i\omega - \zeta_2)\hat{u} = (-\omega^2 - i\omega(\zeta_1 + \zeta_2) + \zeta_1\zeta_2)\hat{u}. \quad (14.27c)$$

By combining (14.26a)–(14.26c) with (14.25a)–(14.25c) and (14.27a)–(14.27c), we obtain the following system:

$$-i\omega\eta\hat{u}^* - \frac{\partial\tilde{v}_1}{\partial x_1} - \frac{\partial\tilde{v}_2}{\partial x_2} = 0, \quad (14.28a)$$

$$-i\omega\gamma^{-1}\hat{v}_1^* - \frac{\partial\hat{u}}{\partial x_1} = 0, \quad (14.28b)$$

$$-i\omega\gamma^{-1}\hat{v}_2^* - \frac{\partial\hat{u}}{\partial x_2} = 0, \quad (14.28c)$$

$$(-i\omega + \zeta_2)\hat{v}_1^* = (-i\omega + \zeta_1)\tilde{v}_1, \quad (14.28d)$$

$$(-i\omega + \zeta_1)\hat{v}_2^* = (-i\omega + \zeta_2)\tilde{v}_2, \quad (14.28e)$$

$$-\omega^2\hat{u}^* = (-\omega^2 - i\omega(\zeta_1 + \zeta_2) + \zeta_1\zeta_2)\hat{u}. \quad (14.28f)$$

Now, by calling v_1 and v_2 the inverse Fourier transforms of \tilde{v}_1 and \tilde{v}_2 (this nomenclature is justified by the fact that \tilde{v}_1 and \tilde{v}_2 coincide in Ω with \hat{v}_1 and \hat{v}_2 as shown in (14.24a) and (14.24b)), we obtain the following system in time domain:

$$\eta \frac{\partial u^*}{\partial t} - \frac{\partial v_1}{\partial x_1} - \frac{\partial v_2}{\partial x_2} = 0, \quad (14.29a)$$

$$\gamma^{-1} \frac{\partial v_1^*}{\partial t} - \frac{\partial u}{\partial x_1} = 0, \quad (14.29b)$$

$$\gamma^{-1} \frac{\partial v_2^*}{\partial t} - \frac{\partial u}{\partial x_2} = 0, \quad (14.29c)$$

$$\frac{\partial v_1}{\partial t} + \zeta_1 v_1 = \frac{\partial v_1^*}{\partial t} + \zeta_2 v_1^*, \quad (14.29d)$$

$$\frac{\partial v_2}{\partial t} + \zeta_2 v_2 = \frac{\partial v_2^*}{\partial t} + \zeta_1 v_2^*, \quad (14.29e)$$

$$\frac{\partial^2 u}{\partial t^2} + (\zeta_1 + \zeta_2) \frac{\partial u}{\partial t} + \zeta_1\zeta_2 u = \frac{\partial^2 u^*}{\partial t^2}. \quad (14.29f)$$

So, the PML consists in solving the same system (14.29a)–(14.29c) as that defined in Ω plus three ODEs in u , v_1 and v_2 defined in (14.29d)–(14.29f).

The practical definition of ζ_j , $j = 1, 2$ can be of the form

$$\zeta_j = \begin{cases} 0 & \text{if } x \leq 0, \\ \zeta \left(\frac{x}{a}\right)^2 & \text{otherwise,} \end{cases} \quad (14.30)$$

with

$$\zeta = \frac{3c_0}{2a} \log(R),$$

where a is the thickness of the layer, $c_0 = \sqrt{\gamma/\eta}$ is the velocity in this layer and $R = 1000$.

Generally, a thickness between 1 or 2 wavelengths provides an excellent absorption, i.e. such that the reflected waves are of the order of the dispersion of the numerical method.

The Maxwell Equations. A similar process provides, for the TM Maxwell equations [39], the following PML:

$$\varepsilon \frac{\partial E_x^*}{\partial t} - \frac{\partial H}{\partial x_2} = 0, \quad (14.31a)$$

$$\varepsilon \frac{\partial E_y^*}{\partial t} + \frac{\partial H}{\partial x_1} = 0, \quad (14.31b)$$

$$\mu \frac{\partial H^*}{\partial t} + \frac{\partial E_y}{\partial x_1} - \frac{\partial E_x}{\partial x_2} = 0, \quad (14.31c)$$

$$\frac{\partial E_x}{\partial t} + \zeta_2 E_x = \frac{\partial E_x^*}{\partial t} + \zeta_1 E_x^*, \quad (14.31d)$$

$$\frac{\partial E_y}{\partial t} + \zeta_1 E_y = \frac{\partial E_y^*}{\partial t} + \zeta_2 E_y^*, \quad (14.31e)$$

$$\frac{\partial^2 H}{\partial t^2} + (\zeta_1 + \zeta_2) \frac{\partial H}{\partial t} + \zeta_1 \zeta_2 H = \frac{\partial^2 H^*}{\partial t^2}. \quad (14.31f)$$

Here, $c_0 = 1/\sqrt{\varepsilon\mu}$.

The Elastics System. By applying the Fourier transform in time and the extension of variables defined in (14.18a), (14.18b), (14.20) and (14.21) to the elastics system defined in (13.125a)–(13.125f) with $f_1 = f_2 = 0$, we obtain [55]:

$$\begin{aligned} -\rho i\omega \hat{v}_1 &- \frac{1}{1 + i\zeta_1/\omega} \left(\frac{\partial \hat{\sigma}_{1,1}}{\partial x_1} + \frac{\partial \hat{\sigma}_{2,1}}{\partial x_1} \right) \\ &- \frac{1}{1 + i\zeta_2/\omega} \left(\frac{\partial \hat{\sigma}_{1,2}}{\partial x_2} + \frac{\partial \hat{\sigma}_{2,2}}{\partial x_2} \right) = 0, \end{aligned} \quad (14.32a)$$

$$\begin{aligned} -\rho i\omega \hat{v}_2 &= -\frac{1}{1+i\zeta_1/\omega} \left(\frac{\partial \hat{\sigma}_{3,1}}{\partial x_1} + \frac{\partial \hat{\sigma}_{4,1}}{\partial x_1} \right) \\ &\quad - \frac{1}{1+i\zeta_2/\omega} \left(\frac{\partial \hat{\sigma}_{3,2}}{\partial x_2} + \frac{\partial \hat{\sigma}_{4,2}}{\partial x_2} \right) = 0, \end{aligned} \quad (14.32b)$$

$$-\mathrm{i}\omega \tilde{\boldsymbol{\sigma}}_1 = A_1 \left(\frac{1}{1+i\zeta_1/\omega} \frac{\partial \hat{v}_1}{\partial x_1}, \frac{1}{1+i\zeta_2/\omega} \frac{\partial \hat{v}_1}{\partial x_2} \right)^T, \quad (14.32c)$$

$$-\mathrm{i}\omega \tilde{\boldsymbol{\sigma}}_2 = A_2 \left(\frac{1}{1+i\zeta_1/\omega} \frac{\partial \hat{v}_2}{\partial x_1}, \frac{1}{1+i\zeta_2/\omega} \frac{\partial \hat{v}_2}{\partial x_2} \right)^T, \quad (14.32d)$$

$$-\mathrm{i}\omega \tilde{\boldsymbol{\sigma}}_3 = A_3 \left(\frac{1}{1+i\zeta_1/\omega} \frac{\partial \hat{v}_1}{\partial x_1}, \frac{1}{1+i\zeta_2/\omega} \frac{\partial \hat{v}_1}{\partial x_2} \right)^T, \quad (14.32e)$$

$$-\mathrm{i}\omega \tilde{\boldsymbol{\sigma}}_4 = A_4 \left(\frac{1}{1+i\zeta_1/\omega} \frac{\partial \hat{v}_2}{\partial x_1}, \frac{1}{1+i\zeta_2/\omega} \frac{\partial \hat{v}_2}{\partial x_2} \right)^T. \quad (14.32f)$$

Now, by setting

$$M_2 = \begin{pmatrix} 1 + \frac{i\zeta_2}{\omega} & 0 \\ 0 & 1 + \frac{i\zeta_1}{\omega} \end{pmatrix}, \quad M_1 = \begin{pmatrix} 1 + \frac{i\zeta_1}{\omega} & 0 \\ 0 & 1 + \frac{i\zeta_2}{\omega} \end{pmatrix} \quad (14.33)$$

and

$$\tilde{\boldsymbol{\sigma}}_j = M_2 \tilde{\boldsymbol{\sigma}}_j, \quad j = 1..4, \quad (14.34)$$

we derive from (14.32a)–(14.32f) the following form of the system:

$$-\rho i\omega \left(1 + \frac{i\zeta_1}{\omega} \right) \left(1 + \frac{i\zeta_1}{\omega} \right) \hat{v}_1 - \nabla \cdot \tilde{\boldsymbol{\sigma}}_1 - \nabla \cdot \tilde{\boldsymbol{\sigma}}_2 = 0, \quad (14.35a)$$

$$-\rho i\omega \left(1 + \frac{i\zeta_1}{\omega} \right) \left(1 + \frac{i\zeta_1}{\omega} \right) \hat{v}_2 - \nabla \cdot \tilde{\boldsymbol{\sigma}}_3 - \nabla \cdot \tilde{\boldsymbol{\sigma}}_4 = 0, \quad (14.35b)$$

$$-\mathrm{i}\omega \tilde{\boldsymbol{\sigma}}_1 = M_2 A_1 M_1^{-1} \nabla \hat{v}_1, \quad (14.35c)$$

$$-\mathrm{i}\omega \tilde{\boldsymbol{\sigma}}_2 = M_2 A_2 M_1^{-1} \nabla \hat{v}_2, \quad (14.35d)$$

$$-\mathrm{i}\omega \tilde{\boldsymbol{\sigma}}_3 = M_2 A_3 M_1^{-1} \nabla \hat{v}_1, \quad (14.35e)$$

$$-\mathrm{i}\omega \tilde{\boldsymbol{\sigma}}_4 = M_2 A_4 M_1^{-1} \nabla \hat{v}_2. \quad (14.35f)$$

By introducing the Bérenger-like variables

$$\hat{v}_1^* = \left(1 + \frac{i\zeta_1}{\omega} \right) \left(1 + \frac{i\zeta_1}{\omega} \right) \hat{v}_1 \quad (14.36a)$$

$$\hat{v}_2^* = \left(1 + \frac{i\zeta_1}{\omega} \right) \left(1 + \frac{i\zeta_1}{\omega} \right) \hat{v}_2 \quad (14.36b)$$

$$\tilde{\boldsymbol{\sigma}}_j^* = M_1 A_j^{-1} M_2^{-1} \tilde{\boldsymbol{\sigma}}_j, \quad j = 1..4 \quad (14.36c)$$

and applying the inverse Fourier transform in time to the modified system, we finally obtain:

$$\rho \frac{\partial v_1^*}{\partial t} - \nabla \cdot \boldsymbol{\sigma}_1 - \nabla \cdot \boldsymbol{\sigma}_2 = f_1, \quad (14.37a)$$

$$\rho \frac{\partial v_2^*}{\partial t} - \nabla \cdot \boldsymbol{\sigma}_3 - \nabla \cdot \boldsymbol{\sigma}_4 = f_2, \quad (14.37b)$$

$$\frac{\partial \boldsymbol{\sigma}_1^*}{\partial t} = A_1 \nabla v_1, \quad (14.37c)$$

$$\frac{\partial \boldsymbol{\sigma}_2^*}{\partial t} = A_2 \nabla v_2, \quad (14.37d)$$

$$\frac{\partial \boldsymbol{\sigma}_3^*}{\partial t} = A_3 \nabla v_1, \quad (14.37e)$$

$$\frac{\partial \boldsymbol{\sigma}_4^*}{\partial t} = A_4 \nabla v_2, \quad (14.37f)$$

$$\frac{\partial^2 v_1}{\partial t^2} + (\zeta_1 + \zeta_2) \frac{\partial v_1}{\partial t} + \zeta_1 \zeta_2 v_1 = \frac{\partial^2 v_1^*}{\partial t^2}, \quad (14.37g)$$

$$\frac{\partial^2 v_2}{\partial t^2} + (\zeta_1 + \zeta_2) \frac{\partial v_2}{\partial t} + \zeta_1 \zeta_2 v_2 = \frac{\partial^2 v_2^*}{\partial t^2}, \quad (14.37h)$$

$$\frac{\partial \boldsymbol{\sigma}_j}{\partial t} + N_1 \boldsymbol{\sigma}_j = A_j \left(\frac{\partial \boldsymbol{\sigma}_j^*}{\partial t} + N_2 \boldsymbol{\sigma}_j^* \right), \quad j = 1..4, \quad (14.37i)$$

where

$$N_1 = \begin{pmatrix} \zeta_1 & 0 \\ 0 & \zeta_2 \end{pmatrix}, \quad N_2 = \begin{pmatrix} \zeta_2 & 0 \\ 0 & \zeta_1 \end{pmatrix}. \quad (14.38)$$

14.2.3 The Three-Dimensional Case

We now give guidelines for the construction of PML for the wave equation in 3D. Their extension to systems can be constructed in a similar way.

Let the system (14.16a) and (14.16b) be set in the open set $\Omega =]-\infty, 0[^3$. After applying the Fourier transform in time to this system, we transform it, as in 2D, by using the change of variables defined in (14.18a) and (14.18b) to which we add

$$\tilde{x}_3 = \begin{cases} x_3 & \text{if } x_3 < 0, \\ x_3 + \frac{i}{\omega} \int_0^{x_3} \zeta_3(s) ds & \text{otherwise.} \end{cases} \quad (14.39)$$

Then, we obtain:

$$-i\omega\eta\hat{u} - \frac{1}{1+i\zeta_1/\omega} \frac{\partial \hat{v}_1}{\partial x_1} - \frac{1}{1+i\zeta_2/\omega} \frac{\partial \hat{v}_2}{\partial x_2} - \frac{1}{1+i\zeta_3/\omega} \frac{\partial \hat{v}_3}{\partial x_3} = 0, \quad (14.40a)$$

$$-i\omega\gamma^{-1}\hat{v}_1 - \frac{1}{1+i\zeta_1/\omega} \frac{\partial \hat{u}}{\partial x_1} = 0, \quad (14.40b)$$

$$-i\omega\gamma^{-1}\hat{v}_2 - \frac{1}{1+i\zeta_2/\omega} \frac{\partial \hat{u}}{\partial x_2} = 0, \quad (14.40c)$$

$$-\mathrm{i}\omega\gamma^{-1}\hat{v}_3 - \frac{1}{1+\mathrm{i}\zeta_3/\omega} \frac{\partial\hat{u}}{\partial x_3} = 0, \quad (14.40d)$$

which can be rewritten as

$$\begin{aligned} & -\left(1 + \frac{\mathrm{i}\zeta_1}{\omega}\right) \left(1 + \frac{\mathrm{i}\zeta_2}{\omega}\right) \left(1 + \frac{\mathrm{i}\zeta_3}{\omega}\right) \mathrm{i}\omega\eta\hat{u} \\ & - \left(1 + \frac{\mathrm{i}\zeta_2}{\omega}\right) \left(1 + \frac{\mathrm{i}\zeta_3}{\omega}\right) \frac{\partial\hat{v}_1}{\partial x_1} \end{aligned} \quad (14.41a)$$

$$\begin{aligned} & -\left(1 + \frac{\mathrm{i}\zeta_1}{\omega}\right) \left(1 + \frac{\mathrm{i}\zeta_3}{\omega}\right) \frac{\partial\hat{v}_2}{\partial x_2} \\ & - \left(1 + \frac{\mathrm{i}\zeta_1}{\omega}\right) \left(1 + \frac{\mathrm{i}\zeta_2}{\omega}\right) \frac{\partial\hat{v}_3}{\partial x_3} = 0, \end{aligned}$$

$$-\left(1 + \frac{\mathrm{i}\zeta_1}{\omega}\right) \mathrm{i}\omega\gamma^{-1}\hat{v}_1 - \frac{\partial\hat{u}}{\partial x_1} = 0, \quad (14.41b)$$

$$-\left(1 + \frac{\mathrm{i}\zeta_2}{\omega}\right) \mathrm{i}\omega\gamma^{-1}\hat{v}_2 - \frac{\partial\hat{u}}{\partial x_2} = 0. \quad (14.41c)$$

$$-\left(1 + \frac{\mathrm{i}\zeta_3}{\omega}\right) \mathrm{i}\omega\gamma^{-1}\hat{v}_3 - \frac{\partial\hat{u}}{\partial x_3} = 0. \quad (14.41d)$$

Now, by setting

$$\tilde{v}_1 = \left(1 + \frac{\mathrm{i}\zeta_2}{\omega}\right) \left(1 + \frac{\mathrm{i}\zeta_3}{\omega}\right) \hat{v}_1, \quad (14.42a)$$

$$\tilde{v}_2 = \left(1 + \frac{\mathrm{i}\zeta_1}{\omega}\right) \left(1 + \frac{\mathrm{i}\zeta_3}{\omega}\right) \hat{v}_2, \quad (14.42b)$$

$$\tilde{v}_3 = \left(1 + \frac{\mathrm{i}\zeta_1}{\omega}\right) \left(1 + \frac{\mathrm{i}\zeta_2}{\omega}\right) \hat{v}_3, \quad (14.42c)$$

we obtain:

$$\begin{aligned} & -\left(1 + \frac{\mathrm{i}\zeta_1}{\omega}\right) \left(1 + \frac{\mathrm{i}\zeta_2}{\omega}\right) \left(1 + \frac{\mathrm{i}\zeta_3}{\omega}\right) \mathrm{i}\omega\eta\hat{u} \\ & - \frac{\partial\tilde{v}_1}{\partial x_1} - \frac{\partial\tilde{v}_2}{\partial x_2} - \frac{\partial\tilde{v}_3}{\partial x_3} = 0, \end{aligned} \quad (14.43a)$$

$$-\frac{1 + \mathrm{i}\zeta_1/\omega}{(1 + \mathrm{i}\zeta_2/\omega)(1 + \mathrm{i}\zeta_3/\omega)} \mathrm{i}\omega\gamma^{-1}\tilde{v}_1 - \frac{\partial\hat{u}}{\partial x_1} = 0, \quad (14.43b)$$

$$-\frac{1 + \mathrm{i}\zeta_2/\omega}{(1 + \mathrm{i}\zeta_1/\omega)(1 + \mathrm{i}\zeta_3/\omega)} \mathrm{i}\omega\gamma^{-1}\tilde{v}_2 - \frac{\partial\hat{u}}{\partial x_2} = 0. \quad (14.43c)$$

$$-\frac{1 + \mathrm{i}\zeta_3/\omega}{(1 + \mathrm{i}\zeta_1/\omega)(1 + \mathrm{i}\zeta_2/\omega)} \mathrm{i}\omega\gamma^{-1}\tilde{v}_3 - \frac{\partial\hat{u}}{\partial x_3} = 0. \quad (14.43d)$$

We can now define the following Bérenger-like variables:

$$\hat{v}_1^* = \frac{1 + i\zeta_1/\omega}{(1 + i\zeta_2/\omega)(1 + i\zeta_3/\omega)} \tilde{v}_1, \quad (14.44a)$$

$$\hat{v}_2^* = \frac{1 + i\zeta_2/\omega}{(1 + i\zeta_1/\omega)(1 + i\zeta_3/\omega)} \tilde{v}_2, \quad (14.44b)$$

$$\hat{v}_3^* = \frac{1 + i\zeta_3/\omega}{(1 + i\zeta_1/\omega)(1 + i\zeta_2/\omega)} \tilde{v}_3, \quad (14.44c)$$

$$\hat{u}^* = (1 + i\zeta_1/\omega)(1 + i\zeta_2/\omega)(1 + i\zeta_3/\omega)\hat{u}, \quad (14.44d)$$

which can be rewritten as

$$(-i\omega + \zeta_2)(-i\omega + \zeta_3)\hat{v}_1^* = (-i\omega + \zeta_1)\tilde{v}_1, \quad (14.45a)$$

$$(-i\omega + \zeta_1)(-i\omega + \zeta_3)\hat{v}_2^* = (-i\omega + \zeta_2)\tilde{v}_2, \quad (14.45b)$$

$$(-i\omega + \zeta_1)(-i\omega + \zeta_2)\hat{v}_3^* = (-i\omega + \zeta_3)\tilde{v}_3, \quad (14.45c)$$

$$\begin{aligned} -i\omega^3\hat{u}^* &= (i\omega - \zeta_1)(i\omega - \zeta_2)(i\omega - \zeta_3)\hat{u} \\ &= (-i\omega^3 + \omega^2(\zeta_1 + \zeta_2 + \zeta_3) \\ &\quad + i\omega(\zeta_1\zeta_2 + \zeta_2\zeta_3 + \zeta_1\zeta_3) - \zeta_1\zeta_2\zeta_3)\hat{u}. \end{aligned} \quad (14.45d)$$

By combining (14.44a)–(14.44d) with (14.43a)–(14.43d) and (14.45a)–(14.45d) and by applying the inverse Fourier transform in time, we obtain the following system in time domain:

$$\eta \frac{\partial u^*}{\partial t} - \frac{\partial v_1}{\partial x_1} - \frac{\partial v_2}{\partial x_2} - \frac{\partial v_3}{\partial x_3} = 0, \quad (14.46a)$$

$$\gamma^{-1} \frac{\partial v_1^*}{\partial t} - \frac{\partial u}{\partial x_1} = 0, \quad (14.46b)$$

$$\gamma^{-1} \frac{\partial v_2^*}{\partial t} - \frac{\partial u}{\partial x_2} = 0, \quad (14.46c)$$

$$\gamma^{-1} \frac{\partial v_3^*}{\partial t} - \frac{\partial u}{\partial x_3} = 0, \quad (14.46d)$$

$$\frac{\partial v_1}{\partial t} + \zeta_1 v_1 = \frac{\partial^2 v_1^*}{\partial t^2} + (\zeta_2 + \zeta_3) \frac{\partial v_1^*}{\partial t} + \zeta_2 \zeta_3 v_1^*, \quad (14.46e)$$

$$\frac{\partial v_2}{\partial t} + \zeta_2 v_2 = \frac{\partial^2 v_2^*}{\partial t^2} + (\zeta_1 + \zeta_3) \frac{\partial v_2^*}{\partial t} + \zeta_1 \zeta_3 v_2^*, \quad (14.46f)$$

$$\frac{\partial v_3}{\partial t} + \zeta_3 v_3 = \frac{\partial^2 v_3^*}{\partial t^2} + (\zeta_1 + \zeta_2) \frac{\partial v_3^*}{\partial t} + \zeta_1 \zeta_2 v_3^*, \quad (14.46g)$$

$$\begin{aligned} \frac{\partial^3 u}{\partial t^3} - (\zeta_1 + \zeta_2 + \zeta_3) \frac{\partial^2 u}{\partial t^2} - (\zeta_1 \zeta_2 + \zeta_2 \zeta_3 + \zeta_1 \zeta_3) \frac{\partial u}{\partial t} \\ - \zeta_1 \zeta_2 \zeta_3 u = \frac{\partial^3 u^*}{\partial t^3}. \end{aligned} \quad (14.46h)$$

So, the ODE in v_1 , v_2 , v_3 remained of first-order with a second-order right-hand side whereas the ODE in u became of third-order instead of second-order

in 2D. However, when one of the ζ_j is equal to 0, one can obtain a second-order system. Actually, all the ζ_j are different from 0 only in the corners of a cubic domain. In the other cases, the equation is the same as in the 2D case.

Remark

The exterior boundary condition in the PML can be reflecting boundary conditions or even first-order ABC that are easy to implement. However, for the elastics system, the free surface boundary condition seems to lead to unstable approximations for long times of resolution. The condition $\mathbf{v} = 0$ seems to provide stable algorithms [55].

14.2.4 Finite Element Approximation

The system (14.29a)–(14.29f) can be rewritten as

$$\eta \frac{\partial u^*}{\partial t} = \nabla \cdot \mathbf{v}, \quad (14.47a)$$

$$\frac{\partial \mathbf{v}^*}{\partial t} = \gamma \nabla u, \quad (14.47b)$$

$$\frac{\partial \mathbf{v}}{\partial t} + N_1 \mathbf{v} = \frac{\partial \mathbf{v}^*}{\partial t} + N_2 \mathbf{v}^*, \quad (14.47c)$$

$$\frac{\partial^2 u}{\partial t^2} + (\zeta_1 + \zeta_2) \frac{\partial u}{\partial t} + \zeta_1 \zeta_2 u = \frac{\partial^2 u^*}{\partial t^2}, \quad (14.47d)$$

where N_1 and N_2 are defined by (14.38).

So, we can write the following variational formulation of the above problem:

Find u , u^* , \mathbf{v} , \mathbf{v}^* such that $u(., t) \in H_0^1(\Omega)$, $u^*(., t) \in H_0^1(\Omega)$, $\mathbf{v}(., t) \in [L^2(\Omega)]^2$, $\mathbf{v}^*(., t) \in [L^2(\Omega)]^2$ and

$$\frac{d}{dt} \int_{\Omega} \eta u \varphi d\mathbf{x} = - \int_{\Omega} \mathbf{v} \cdot \nabla \varphi d\mathbf{x} d\mathbf{x} \quad \forall \varphi \in H_0^1(\Omega), \quad (14.48a)$$

$$\frac{d}{dt} \int_{\Omega} \gamma^{-1} \mathbf{v} \cdot \psi d\mathbf{x} = \int_{\Omega} \nabla u \cdot \psi d\mathbf{x} \quad \forall \psi \in [L^2(\Omega)]^2, \quad (14.48b)$$

$$\begin{aligned} & \frac{d}{dt} \int_{\Omega} \mathbf{v} \cdot \psi d\mathbf{x} + \int_{\Omega} N_1 \mathbf{v} \cdot \psi d\mathbf{x} \\ &= \frac{d}{dt} \int_{\Omega} \mathbf{v}^* \cdot \psi d\mathbf{x} + \int_{\Omega} N_2 \mathbf{v}^* \cdot \psi d\mathbf{x} \quad \forall \psi \in [L^2(\Omega)]^2, \end{aligned} \quad (14.48c)$$

$$\begin{aligned} & \frac{d^2}{dt^2} \int_{\Omega} u \varphi d\mathbf{x} + \frac{d}{dt} \int_{\Omega} (\zeta_1 + \zeta_2) u \varphi d\mathbf{x} \\ &+ \int_{\Omega} \zeta_1 \zeta_2 u \varphi d\mathbf{x} = \frac{d^2}{dt^2} \int_{\Omega} u^* \varphi d\mathbf{x} \quad \forall \varphi \in H_0^1(\Omega). \end{aligned} \quad (14.48d)$$

After approximation in space by finite elements as in Sect. 13.4.1, we obtain:

$$D_{h,0} \frac{d\mathbf{U}^*}{dt} = -R_h \mathbf{V}, \quad (14.49a)$$

$$B_{h,0} \frac{d\mathbf{V}^*}{dt} = R_h^* \mathbf{U}, \quad (14.49b)$$

$$B_{h,0} \frac{d\mathbf{V}}{dt} + B_{h,1} \mathbf{V} = B_{h,0} \frac{d\mathbf{V}^*}{dt} + B_{h,2} \mathbf{V}^*, \quad (14.49c)$$

$$D_{h,1} \frac{d^2\mathbf{U}}{dt^2} + D_{h,2} \frac{d\mathbf{U}}{dt} + D_{h,3} \mathbf{U} = D_{h,1} \frac{d^2\mathbf{U}^*}{dt^2}, \quad (14.49d)$$

where $D_{h,j}$, $j = 0..3$ are diagonal matrices and $B_{h,j}$, $j = 0..2$ are 2×2 block-diagonal matrices.

Now, let us approximate (14.49c) by a centered second-order scheme in time. We have:

$$\begin{aligned} B_{h,0} \frac{\mathbf{V}^{n+1} - \mathbf{V}^n}{\Delta t} + B_{h,1} \frac{\mathbf{V}^{n+1} + \mathbf{V}^n}{2} &= B_{h,0} \frac{\mathbf{V}^{*n+1} - \mathbf{V}^{*n}}{\Delta t} \\ &\quad + B_{h,2} \frac{\mathbf{V}^{*n+1} + \mathbf{V}^{*n}}{2}, \end{aligned}$$

which can be rewritten as

$$\begin{aligned} B_{h,3} \mathbf{V}^{n+1} &= \left(\frac{1}{\Delta t} B_{h,0} - \frac{1}{2} B_{h,1} \right) \mathbf{V}^n \\ &\quad + B_{h,0} \frac{\mathbf{V}^{*n+1} - \mathbf{V}^{*n}}{\Delta t} + B_{h,2} \frac{\mathbf{V}^{*n+1} + \mathbf{V}^{*n}}{2}, \end{aligned} \quad (14.50)$$

where

$$B_{h,3} = \frac{1}{\Delta t} B_{h,0} + \frac{1}{2} B_{h,1}$$

is a 2×2 block-diagonal matrix.

Equations (14.49a)–(14.49d) and (14.50) show that the PML can be applied to quadrilaterals of any shape. The treatment of the additional equations is equivalent to the inversion of a block-diagonal matrix in the same way as the equations in the physical domain.

14.2.5 Approximation in Time

The approximation of all the time derivatives by second-order centered finite difference methods provides stable schemes. The use of higher-order schemes raises, however, some difficulties that we shall describe using the 2D wave equation.

Let us consider the system (14.48a)–(14.48d). The application of the modified equation approach to this system requires the computation of the third-order derivative in time in terms of derivatives in space, as described in Sect. 4.7.3. For these equations, we have for instance, when $\eta = \gamma = 1$:

$$\begin{aligned}\frac{\partial^3 u^*}{\partial t^3} &= \frac{\partial^2}{\partial t^2} \left(\frac{\partial u^*}{\partial t} \right) = \frac{\partial^2}{\partial t^2} (\nabla \cdot \mathbf{v}) \\ &= \frac{\partial^2}{\partial t^2} (-\nabla \cdot (N_1 \mathbf{v}) + \Delta u + \nabla \cdot (N_2 \mathbf{v}^*)).\end{aligned}$$

This first step of our computation shows that it will be very difficult to obtain a simple expression of $\partial^3 u^* / \partial t^3$. Moreover, this term will introduce mass matrices derived from the gradient and the divergence term which are neither diagonal nor block-diagonal matrices. So, the use of such an approach in time is too complicated and expensive in the case of the PML. On the other hand, symmetric schemes require us to formulate the whole system approximated in space as a first-order ODE system, which is also not obvious. Therefore, the best way of time-differencing with a PML is the use of centered second-order approximations with a smaller value of the time-step (half of the CFL seems to be enough) when the velocity does not vary too much. For some configurations with large variations of the velocity (as in the Foothills experiment below), one can even use the maximum value of the time-step. Similar problems are also encountered for ABC.

14.3 Numerical Illustrations

In this section, we give some numerical experiments which illustrate the modeling of unbounded domains and the other features of mass-lumped mixed formulations of wave equations.

14.3.1 PML for the 2D Elastics System

In the first experiment, we illustrate the efficiency and the accuracy for the most complex wave equation which is the elastics system. In this experiment, the Lamé's coefficients are $\lambda = 1/4$ and $\mu = 1$. ρ is equal to 1 and the source, which is a pulse of frequency 1/2 parallel to the x axis, is at the center of the domain. All around a square domain $\Omega =]0, 25[^2$, we add a layer Ω_e of thickness 4, which corresponds to 4/3 of the wavelength of the P -wave and twice the wavelength of the S -wave. Then, we solve the elastics system in Ω and the PML equations in Ω_e . In Fig. 14.1, we show the absorption of the waves. The bottom snapshot in this figure is the same as the fourth one, when the wave disappeared from the physical domain, with a scale divided by 100.

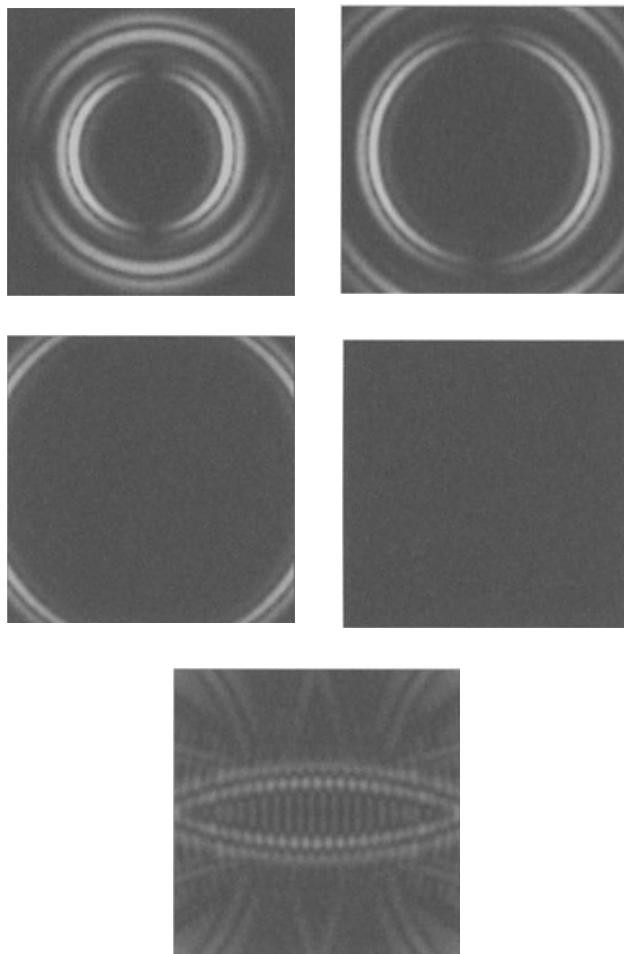


Fig. 14.1. Absorption of the waves in an elastic domain. The bottom snapshot corresponds to the one in which the wave disappeared with a scale divided by 100

14.3.2 Scattering by an Ogive of an Electromagnetic Wave

In this section, we give two experiments of scattering of an electromagnetic wave by a 2D ogive for a TM model². The physical domain $\Omega_\phi = [0, 30]^2$ is surrounded by a PML of thickness 1. The source is the pulse defined in (12.73) located at the point (1,10) with $R = 1.5$. We use a Q_3 approximation. The domain is meshed by 9948 quadrilaterals (1164 for the PML), which corresponds to 239 656 degrees of freedom for the electric field (27 984 for the PML) and 159 168 degrees of freedom for the magnetic field (18 624 for the PML). We have about three elements per wavelength in each direction.

² The speed of light was here normalized to 1.

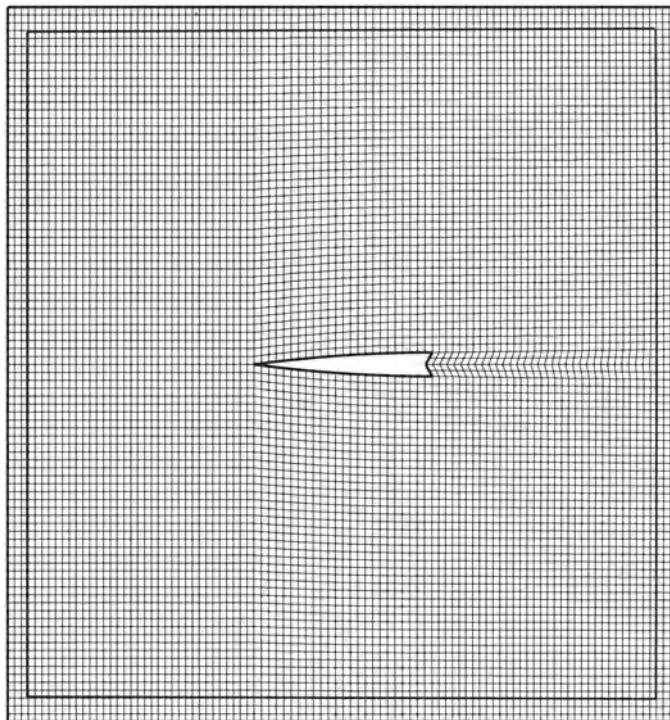


Fig. 14.2. The mesh around the ogive and its PML

In a first experiment (Fig. 14.3), we set $\varepsilon = \mu = 1$ in the whole domain. In the second experiment (Fig. 14.4), we have the same values in the domain except for the layer of elements around the ogive in which we set

$$\underline{\varepsilon} = \begin{pmatrix} 1 & 1/4 \\ 1/4 & 2 \end{pmatrix}.$$

The time-step was $1/30$, which corresponds to about half of the CFL. The CPU time³ is 4 mn 20 s.

14.3.3 A “Foothills” Experiment

The last experiment is in the field of geophysics [32]. It is a so-called “Foothills” model in which a wave propagates in a layered domain representing the subsoil of a hilly region. The equations here are those of the acoustics which are often used by geophysicists as an approximation of the elastics system. The data of the model (provided by P. Ricarte from IFP and

³ For a DEC AlphaStation 500, 1 processor 21164 (500 MHz), 256 MB, 4.3 GB.

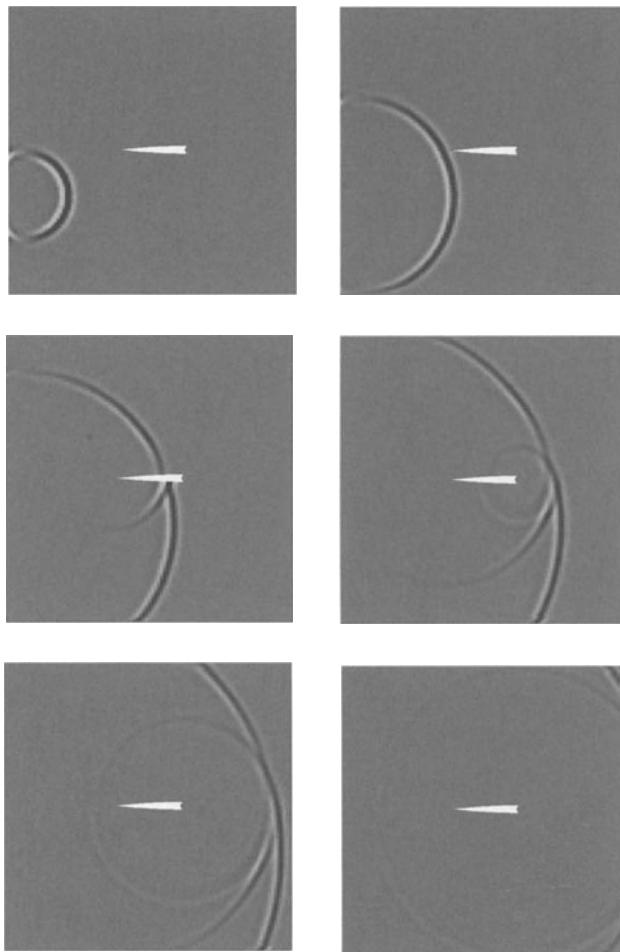


Fig. 14.3. Snapshots of the magnetic field for an isotropic medium. The (normalized) time varies from $t = 5$ (above left) to $t = 30$ (below right) in intervals equal to 5

J. Muller from ELF) are shown in Fig. 14.5. The length of the domain is 4200 m and its maximum height is a little less than 3000 m. The velocities vary from 1700 m/s (in the top layers) to 5500 m/s (in the bottom layers). The source (described by a Gaussian function in space of radius 42 m), located at the middle of the upper boundary, is a second-order Ricker function (second derivative of a Gaussian function) with a frequency equal to 20 Hz.

We mesh this domain by adapting the mesh to the velocity zones such that the edges of the elements are located at the interfaces. We have about two elements per wavelength in each direction. This leads to a mesh with 3141 elements plus 552 elements for the PML surrounding the domain except for

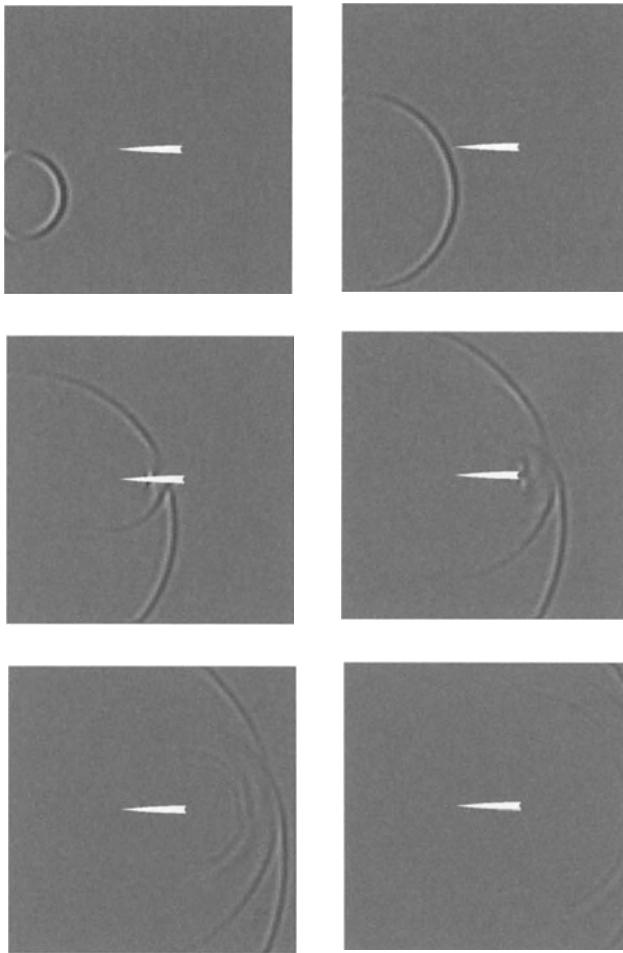


Fig. 14.4. Snapshots of the magnetic field for an ogive with an anisotropy zone around it. The (normalized) time varies from $t = 5$ (above left) to $t = 30$ (below right) in intervals equal to 5

the top boundary (Fig. 14.6). We solve the acoustics equation for $0 \leq t \leq 2$ s in its mixed formulation by using Q_5 (92 986 degrees of freedom in u , CPU: 8 mn) then Q_8 elements (237 409 degrees of freedom in u , CPU: 28 mn). Both approximations provide exactly the same solution, which should mean that we have obtained the converged solution. Solutions obtained on regular meshes (with 2 and 3 elements per wavelength for the smallest wavelength) with Q_5 and Q_8 coincide neither between themselves nor with the solution obtained with an adapted mesh. This phenomenon is probably due to the error generated by the crossing of the interfaces which is $O(h)$ for regular



Fig. 14.5. The different layers in the Foothills model

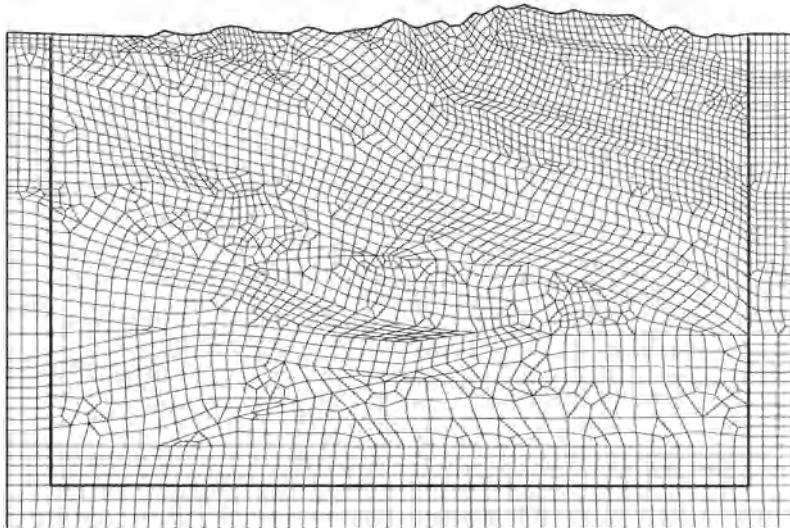


Fig. 14.6. The adapted mesh with its PML zone around it

meshes which do not follow the interfaces and $O(h^6)$ for the adapted mesh (as shown in Sect. 11.6).

The snapshots of the solution are given in Fig. 14.7 and a corresponding seismogram in Fig. 14.8.

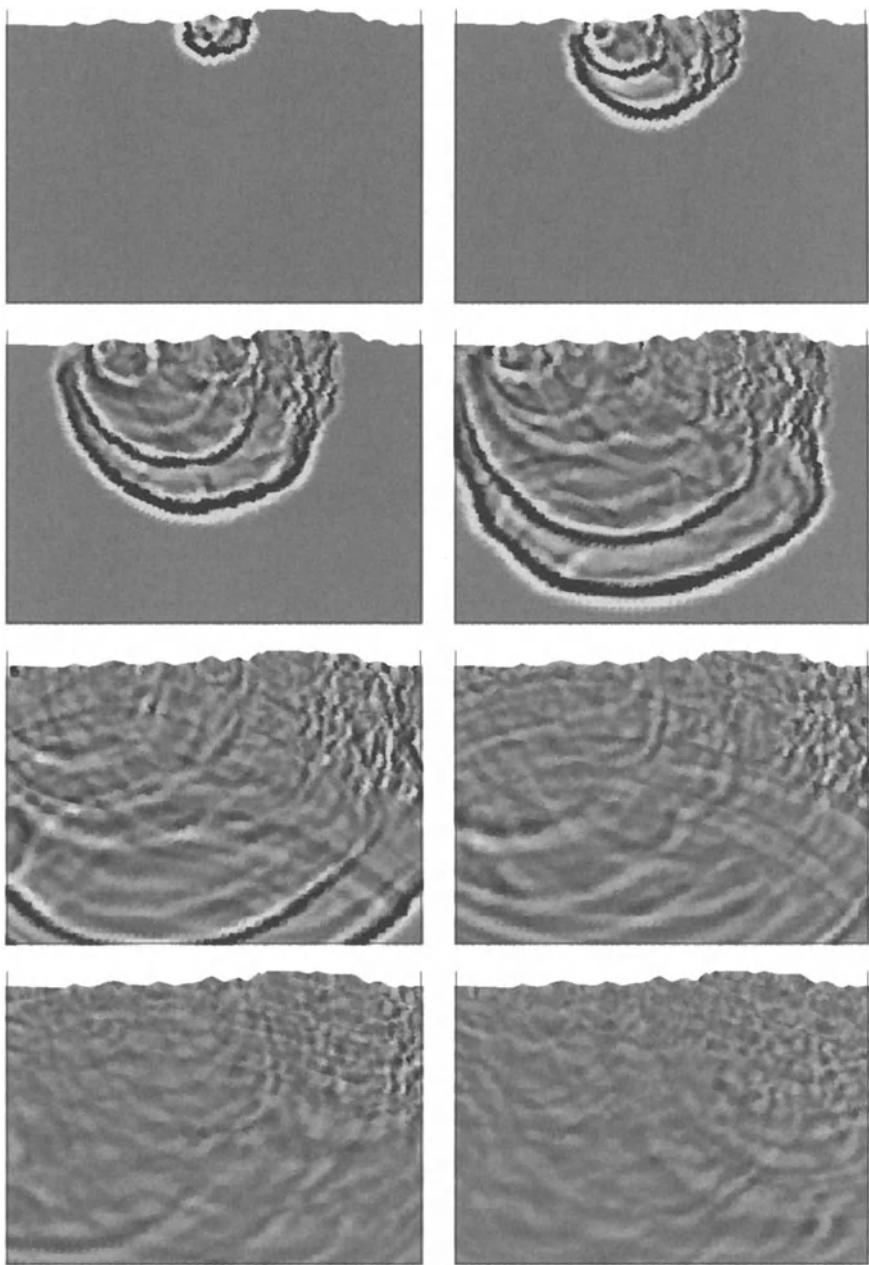


Fig. 14.7. Snapshots of the solution from $t = 0.2$ s (above left) to $t = 1.6$ s (below right) in intervals of 0.2 s

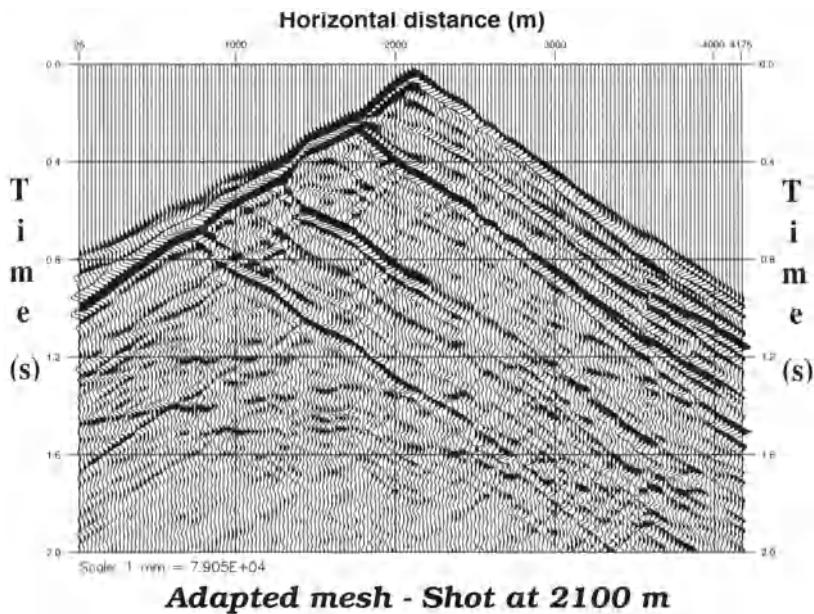


Fig. 14.8. The seismogram for a shot on the surface at $x = 2100$ m

14.4 Computational Issues

14.4.1 Tetrahedra or Hexahedra?

In Chaps. 12 and 13, we presented several mass-lumped tetrahedral (resp. triangular) and hexahedral (resp. quadrilateral) elements.

For continuous as well as for edge elements, the tensorial structure of hexahedra leads to orthogonality properties which enable us to gain storage and CPU time in a substantial way. On the other hand, the use of discontinuous approximations for the auxiliary variables and the local character of the stiffness matrices provide algorithms that are very easy to implement since one can treat each element separately. Unfortunately, all these features do not hold for tetrahedral elements whose shapes make orthogonality impossible. Of course, one could also write mixed formulations for tetrahedra (which seems fundamental for implementing the PML) in which the stiffness matrices would be locally defined but the lack of orthogonality would imply that all (or most of) the interactions between the basis functions are not equal to zero. This means that such formulations would significantly slow down the algorithms instead of speeding them up, as for hexahedral elements. Besides this, continuous tetrahedral elements are terribly expensive in terms of both storage and CPU time.

All these statements show that hexahedral elements are much more efficient than tetrahedral elements but... when one speaks with an industrial partner, an immediate objection is raised: Very few hexahedral mesh generators are available. This is true, but some are available and their number is increasing. If not, it would be worth devoting some work in this direction to be able to use high performance methods which could take into account the real needs of industry.

14.4.2 Finite Element or Finite Difference Methods?

For a long time (and even sometimes nowadays) it was claimed that using finite element methods for the wave equations was too expensive and required too much human investment. Therefore, finite difference methods, with all their imperfections, seemed more worthwhile.

This was true when people did not know how to mass-lump and when stiffness matrices required a huge storage in finite element methods (in truth, finite difference methods also need a substantial storage for complex physics). The methods developed in the previous chapters show that we can now construct finite element methods which are efficient in terms of storage and CPU time as well. No real comparisons were made between the two approaches but rough evaluations of the storage and of the number of operations show that, for a given accuracy, finite element methods can do much better than finite difference in cases of complex physics or geometry. In particular, the gain over finite difference methods is increased by the fact that, for finite difference, the space-step is dictated by the slowest velocity, which implies the use of small space-steps even in the large velocity regions. This drawback is particularly obvious in geophysics, where models (like the Foothills) contain a large number of layers in which the velocities can be multiplied up to 3 or 4 times. In finite element methods, the mesh can be adapted to the velocities (as in Fig. 14.6) and this induces a huge gain of storage and CPU time. This is even more obvious in 3D for which we saw that the gain of storage and CPU of mixed formulations is much more significant.

So, a simple change of mind is needed...

A. Appendix: Construction of a General $H(\mathbf{curl}, K)$ -Conforming Transform

A.1 A Local $H(\mathbf{curl}, \hat{K})$ -Conforming Isomorphism

A.1.1 Notation

Let us consider¹:

- two bounded open sets \hat{K} and K of \mathbb{R}^3 with smooth boundaries;
- $\mathbf{F} : \hat{K} \rightarrow K$ a C^1 -diffeomorphism;
- and also

$$\text{and } \begin{aligned} \hat{\psi} &: \hat{K} \rightarrow \mathbb{R}^3, \\ \psi &: K \rightarrow \mathbb{R}^3, \end{aligned} \quad (\text{A.1})$$

two vector fields related to each other by the following relation:

$$\psi \circ \mathbf{F} = \hat{\psi}.$$

We then define:

- The gradient of a vector field ψ , denoted $D\psi$, which is actually the matrix whose general term is

$$(D\psi)_{i,j} = \frac{\partial \psi_i}{\partial x_j}.$$

- The curl of a vector field ψ , denoted as $\nabla \times \psi$.

It will be more convenient to work with the matrix form of the curl that we recall below:

To any vector \mathbf{u} of \mathbb{R}^3 , we can associate the unique antisymmetric matrix $A(\mathbf{u})$ defined as follows

$$\forall \mathbf{v} \in \mathbb{R}^3, A(\mathbf{u})\mathbf{v} = \mathbf{u} \times \mathbf{v}.$$

The relation between \mathbf{u} and A defines the following bijective application:

¹ This annex is a part of an unpublished paper by A. Elmkes.

$$\begin{array}{ccc} T : & \mathbb{R}^3 & \rightarrow S_3^- , \\ \boldsymbol{u} & \rightarrow & T(\boldsymbol{u}) = A(\boldsymbol{u}), \end{array} \quad (\text{A.2})$$

where S_3^- is the set of antisymmetric 3×3 matrices. $\forall A \in S_3^-$, $\boldsymbol{u} = T^{-1}(A)$ is called the “axial vector” associated with A . $\nabla \times \boldsymbol{\psi}$ is then defined as the axial vector associated with the antisymmetric matrix $D\boldsymbol{\psi} - D\boldsymbol{\psi}^*$, which can be written as

$$T(\nabla \times \boldsymbol{\psi}) = D\boldsymbol{\psi} - D\boldsymbol{\psi}^*.$$

In order to simplify our computations, we shall use the index notation, i.e. we shall suppress the sign \sum as soon as the index of summation is repeated. For instance, $\sum_i u_i e_i$ will be denoted $u_i e_i$.

Remark:

If $\|\cdot\|$ is the usual norm of \mathbb{R}^3 and $\|\cdot\|_*$, the corresponding matrix norm, we have:

$$\|\boldsymbol{u}\| = \|T(\boldsymbol{u})\|_*.$$

A.1.2 A Local $H(\text{curl}, \hat{K})$ -Conforming Isomorphism

Our purpose is to construct an isomorphism from $H(\text{curl}, \hat{K})$ to $H(\text{curl}, K)$. For a triangular tetrahedral finite element, the appropriate one [63, 92, 93] is the isomorphism \mathcal{R} which associates with a function $\hat{\boldsymbol{\psi}}$ the function $\tilde{\boldsymbol{\psi}}$ defined by

$$\tilde{\boldsymbol{\psi}} \circ \mathbf{F} = DF^{*-1} \hat{\boldsymbol{\psi}}, \quad (\text{A.3})$$

where $\hat{\boldsymbol{\psi}} : \hat{K} \rightarrow \mathbb{R}^3$ and $\tilde{\boldsymbol{\psi}} : K \rightarrow \mathbb{R}^3$.

Our purpose is to show that this transform is still the correct one for any conform mapping \mathbf{F} .

A Relation Between Matrices.

Lemma 6. *With the above notation, we have the following relation:*

$$T(\nabla \times \hat{\boldsymbol{\psi}}) = DF^* T(\nabla \times \boldsymbol{\psi}) DF,$$

where DF^* is the transposed matrix of DF .

Proof. We have:

$$\forall \hat{\mathbf{x}} \in \hat{K}, \boldsymbol{\psi} \circ \mathbf{F}(\hat{\mathbf{x}}) = DF^*(\hat{\mathbf{x}}) \tilde{\boldsymbol{\psi}} \circ \mathbf{F}(\hat{\mathbf{x}}).$$

Let us set $i, k \in \{1, 3\}$.

The above relation can be written as

$$\forall i \in \{1, N\}, \psi_i \circ \mathbf{F}(\hat{\mathbf{x}}) = \frac{\partial F_j}{\partial \hat{x}_i} \tilde{\psi}_j \circ \mathbf{F}(\hat{\mathbf{x}}).$$

Now, by differentiating this equality with respect to \hat{x}_k , we obtain:

$$\frac{\partial(\psi_i \circ \mathbf{F})}{\partial \hat{x}_k} = \frac{\partial^2 F_j}{\partial \hat{x}_k \partial \hat{x}_j} \tilde{\psi}_j + \frac{\partial F_j}{\partial \hat{x}_i} \frac{\partial(\tilde{\psi}_j \circ \mathbf{F})}{\partial \hat{x}_k}.$$

In other words,

$$\frac{\partial \psi_i}{\partial x_l} \frac{\partial F_l}{\partial \hat{x}_k} = \frac{\partial^2 F_j}{\partial \hat{x}_k \partial \hat{x}_i} \tilde{\psi}_j + \frac{\partial F_j}{\partial \hat{x}_i} \frac{\partial \tilde{\psi}_j}{\partial x_l} \frac{\partial F_l}{\partial \hat{x}_k}.$$

So, if we set $A_{i,k} = \frac{\partial^2 F_j}{\partial \hat{x}_k \partial \hat{x}_i} \tilde{\psi}_j$, we obtain:

$$D\psi DF = A + DF^* D\tilde{\psi} DF,$$

so that

$$\begin{aligned} D\hat{\psi} - D\hat{\psi}^* &= D\psi DF - (D\psi DF)^*, \\ &= DF^* D\tilde{\psi} DF - DFD\tilde{\psi} DF^* + A - A^*, \\ &= DF^*(D\tilde{\psi} - D\tilde{\psi}^*)DF + A - A^*. \end{aligned} \tag{A.4}$$

Now, since \mathbf{F} is a C^1 mapping, we can write:

$$\frac{\partial^2 F_j}{\partial \hat{x}_k \partial \hat{x}_i} = \frac{\partial^2 F_j}{\partial \hat{x}_i \partial \hat{x}_k},$$

which shows that A is self-adjoint.

So, we have:

$$D\hat{\psi} - D\hat{\psi}^* = DF^*(D\tilde{\psi} - D\tilde{\psi}^*)DF,$$

which can be written as

$$T(\nabla \times \hat{\psi}) = DF^* T(\nabla \times \psi) DF.$$

A Relation Between Vectors.

Lemma 7. *If \mathbf{U} and $\hat{\mathbf{U}}$ are two vectors of \mathbb{R}^3 so that*

$$T(\hat{\mathbf{U}}) = B^* T(\mathbf{U}) B,$$

where B is a 3×3 invertible matrix, we have the following identity:

$$\hat{\mathbf{U}} = \det(B) B^{-1} \mathbf{U}.$$

Proof. Let us denote using $\{\mathbf{e}_i\}$ the canonical basis of \mathbb{R}^3 and using $\{A_i\}$ the basis of S_3^- defined by $A_i = T(\mathbf{e}_i)$.

By writing \mathbf{U} and $\hat{\mathbf{U}}$ in terms of $\{\mathbf{e}_i\}$, we obtain the following identities:

$$\begin{aligned}\mathbf{U} &= u_i \mathbf{e}_i, \\ \hat{\mathbf{U}} &= \hat{u}_i \mathbf{e}_i.\end{aligned}\tag{A.5}$$

Now, by applying T to the above identities, we have:

$$\begin{aligned}T(\mathbf{U}) &= u_i A_i, \\ T(\hat{\mathbf{U}}) &= \hat{u}_i A_i.\end{aligned}\tag{A.6}$$

Since we have

$$T(\hat{\mathbf{U}}) = B^* T(\mathbf{U}) B,$$

we can write

$$T(\hat{\mathbf{U}}) = u_i B^* A_i B.$$

On the other hand, since $\forall i, B^* A_i B \in S_3^-$, there exists $\{C_{i,j}\}$ so that:

$$\forall i, B^* A_i B = C_{i,j} A_j.$$

Then we obtain:

$$\begin{aligned}T(\hat{\mathbf{U}}) &= u_i B^* A_i B = u_i C_{i,j} A_j, \\ &= C_{i,j} u_i A_j = \hat{u}_j A_j.\end{aligned}\tag{A.7}$$

By identifying the terms, we have:

$$\forall j, \hat{u}_j = C_{i,j} u_i.$$

In other words,

$$\hat{\mathbf{U}} = C^* \mathbf{U}.$$

By some computations, aided by Maple, we can show that C is the comatrix of B , which implies that

$$C^* = \det(B) B^{-1}.$$

We finally obtain:

$$\hat{\mathbf{U}} = \det(B) B^{-1} \mathbf{U}.$$

Remark:

More generally, we can interpret this lemma as the inversion of the identity:

$$T(\hat{\mathbf{U}}) = B^* T(\mathbf{U}) B,$$

by writing

$$\forall \mathbf{U} \in \mathbb{R}^3, T^{-1}(B^* T(\mathbf{U}) B) = \det(B) B^{-1} \mathbf{U},$$

which provides, by applying T to the previous identity:

$$\forall \mathbf{U} \in \mathbb{R}^3, T(B^{-1}\mathbf{U}) = \det(B^{-1})B^*T(\mathbf{U})B.$$

Now, by applying the previous lemma to $\mathbf{U} = \nabla \times \psi$, $\hat{\mathbf{U}} = \nabla \times \hat{\psi}$ and $B = DF$, we deduce the following lemma.

Lemma 8. *With the same hypothesis as Lemma 6, we obtain*

$$\nabla \times \hat{\psi} = \det(DF)DF^{-1}\nabla \times \tilde{\psi}.$$

By using this lemma, we can prove the following proposition.

Proposition 6. *If we assume that $\det(DF) \geq 0$, then \mathcal{R} , defined in (A.3), is a bijective transform from $H(\mathbf{curl}, \hat{K})$ to $H(\mathbf{curl}, K)$.*

Proof. Let $\hat{\mathbf{q}}$ be a function of $H(\mathbf{curl}, \hat{K})$ and $\tilde{\mathbf{q}} : K \rightarrow \mathbb{R}^3$ defined by:

$$\tilde{\mathbf{q}} \circ \mathbf{F} = DF^{*-1}\hat{\mathbf{q}}.$$

- Obviously, $\tilde{\mathbf{q}} \in [L^2(K)]^3$.
- Let us show that $\nabla \times \tilde{\mathbf{q}} \in [L^2(K)]^3$.

In the distribution sense, we have on K :

$$\forall \tilde{\psi} \in [\mathcal{D}(K)]^3, \langle \nabla \times \tilde{\mathbf{q}}, \tilde{\psi} \rangle_{[\mathcal{D}'(K)]^3} = \int_K \tilde{\mathbf{q}} \cdot \nabla \times \tilde{\psi} \, dx.$$

By using the following change of variables

$$\begin{aligned} \mathbf{x} &= \mathbf{F}(\hat{\mathbf{x}}), \\ d\mathbf{x} &= \det(DF)d\hat{\mathbf{x}}, \end{aligned} \tag{A.8}$$

we obtain

$$\begin{aligned} \langle \nabla \times \tilde{\mathbf{q}}, \tilde{\psi} \rangle_{[\mathcal{D}(K)]^3} &= \int_K \tilde{\mathbf{q}} \cdot \nabla \times \tilde{\psi} \, dx, \\ &= \int_{\hat{K}} \tilde{\mathbf{q}} \circ \mathbf{F} \cdot \nabla \times \tilde{\psi} \circ \mathbf{F} \det(DF) \, d\hat{\mathbf{x}}. \end{aligned} \tag{A.9}$$

Now, by using Lemma 8, we have:

$$\nabla \times \hat{\psi} = \det(DF)DF^{-1}\nabla \times \tilde{\psi},$$

where $\tilde{\psi} = \mathcal{R}(\hat{\psi})$.

In other words,

$$\nabla \times \tilde{\psi} = \det(DF^{-1})DF\nabla \times \hat{\psi}.$$

So,

$$\begin{aligned} <\nabla \times \tilde{\mathbf{q}}, \tilde{\psi}>_{[\mathcal{D}'(K)]^3} &= \int_{\hat{K}} \tilde{\mathbf{q}} \circ \mathbf{F} \cdot \det(DF^{-1})DF\nabla \times \hat{\psi} \det(DF) d\hat{x}, \\ &= \int_{\hat{K}} \tilde{\mathbf{q}} \circ \mathbf{F} \cdot DF\nabla \times \hat{\psi} d\hat{x}, \\ &= \int_{\hat{K}} DF^* \tilde{\mathbf{q}} \cdot \nabla \times \hat{\psi} d\hat{x}, \\ &= \int_{\hat{K}} \hat{\mathbf{q}} \cdot \nabla \times \hat{\psi} d\hat{x}, \end{aligned} \tag{A.10}$$

which provides

$$<\nabla \times \tilde{\mathbf{q}}, \tilde{\psi}>_{[\mathcal{D}'(K)]^3} = <\nabla \times \hat{\mathbf{q}}, \hat{\psi}>_{[\mathcal{D}'(\hat{K})]^3}.$$

Now, since \mathcal{R} is an isomorphism from $[L^2(\hat{K})]^3$ to $[L^2(K)]^3$ and from $[\mathcal{D}(\hat{K})]^3$ to $[\mathcal{D}(K)]^3$, this proves that the distribution $\nabla \times \tilde{\mathbf{q}}$ actually belongs to $[L^2(K)]^3$.

A.2 A Global $H(\text{curl}, \Omega)$ -Conforming Isomorphism

A.2.1 Notation

We define:

- $\hat{\Omega}_1$ and $\hat{\Omega}_2$ to be two bounded open sets of \mathbb{R}^3 with smooth boundaries such that $\hat{\Omega}_1 \cup \hat{\Omega}_2 = \hat{\Gamma}$. Moreover, we denote $\hat{\Omega} = \hat{\Omega}_1 \cup \hat{\Omega}_2 \cup \hat{\Gamma}$.
- Ω_1 and Ω_2 to be two bounded open sets of \mathbb{R}^3 with smooth boundaries such that $\bar{\Omega}_1 \cup \bar{\Omega}_2 = \Gamma$. Moreover, we denote $\Omega = \Omega_1 \cup \Omega_2 \cup \Gamma$.
- and

$$\begin{aligned} \mathbf{F}_1 &: \hat{\Omega}_1 \rightarrow \Omega_1, \\ \mathbf{F}_2 &: \hat{\Omega}_2 \rightarrow \Omega_2, \end{aligned} \tag{A.11}$$

two C^1 -diffeomorphisms such that:

- $\mathbf{F}_{1|_{\hat{\Gamma}}} = \mathbf{F}_{2|_{\hat{\Gamma}}} = \mathbf{F}$.
- $\mathbf{F}_1(\hat{\Gamma}) = \mathbf{F}_2(\hat{\Gamma}) = \mathbf{F}(\hat{\Gamma}) = \Gamma$.

Let us now define the global isomorphism \mathcal{R} : Let $\hat{\mathbf{q}} \in H(\mathbf{curl}, \hat{\Omega})$. We denote:

$$\begin{aligned}\hat{\mathbf{q}}|_{\hat{\Omega}_1} &= \hat{\mathbf{q}}_1, \\ \hat{\mathbf{q}}|_{\hat{\Omega}_2} &= \hat{\mathbf{q}}_2.\end{aligned}\tag{A.12}$$

Then, we define $\mathbf{q} : \Omega \rightarrow \mathbb{R}^N$ by:

$$\begin{aligned}\mathbf{q}|_{\Omega_1} \circ \mathbf{F}_1 &= DF_1^{*-1} \hat{\mathbf{q}}_1, \\ \mathbf{q}|_{\Omega_2} \circ \mathbf{F}_2 &= DF_2^{*-1} \hat{\mathbf{q}}_2.\end{aligned}\tag{A.13}$$

In the same way, we denote:

$$\begin{aligned}\mathbf{q}|_{\Omega_1} &= \mathbf{q}_1, \\ \mathbf{q}|_{\Omega_2} &= \mathbf{q}_2.\end{aligned}\tag{A.14}$$

A.2.2 A Global $H(\mathbf{curl}, \Omega)$ -Conforming Isomorphism

Proposition 7. *If we assume that $\det(DF) \geq 0$, then \mathcal{R} , defined in (A.3), is a bijective transform from $H(\mathbf{curl}, \hat{\Omega})$ to $H(\mathbf{curl}, \Omega)$.*

Proof. The proof depends on the following two lemmas.

Lemma 9. *We maintain the above notation. Let us then define:*

- $\hat{\boldsymbol{\nu}}$ as a unit normal to $\hat{\Gamma}$.
- $\boldsymbol{\nu}$ as the unit normal to Γ with the same orientation as $\hat{\boldsymbol{\nu}}$.

Then, we have the following identity:

$$\boldsymbol{\nu} = DF_1^{*-1} \hat{\boldsymbol{\nu}} = DF_2^{*-1} \hat{\boldsymbol{\nu}}.$$

Proof. Let us assume that $\hat{\Gamma}$ is (locally) defined using the following equation:

$$\hat{h}(\hat{\mathbf{x}}) = 0.$$

The outward unit normal $\hat{\Gamma}$ is then given by:

$$\hat{\boldsymbol{\nu}} = \nabla \hat{h}.$$

In the same way, we can define Γ using:

$$h(\mathbf{x}) = \hat{h} \circ \mathbf{F}_1^{-1}(\hat{\mathbf{x}}) = \hat{h} \circ \mathbf{F}_2^{-1}(\hat{\mathbf{x}}) = 0.$$

After differentiating the previous equation, we obtain the expression of $\boldsymbol{\nu}$ in terms of $\hat{\boldsymbol{\nu}}$:

$$\boldsymbol{\nu} = DF_1^{*-1} \hat{\boldsymbol{\nu}} = DF_2^{*-1} \hat{\boldsymbol{\nu}}.$$

Lemma 10. *With the above notation, we have the following identities:*

$$\mathbf{q}_1 \times \boldsymbol{\nu} = \det(DF_1^{-1})DF_1\hat{\mathbf{q}}_1 \times \hat{\boldsymbol{\nu}}.$$

In the same way, we have:

$$\mathbf{q}_2 \times \boldsymbol{\nu} = \det(DF_2^{-1})DF_2\hat{\mathbf{q}}_2 \times \hat{\boldsymbol{\nu}}.$$

Proof. By definition, we know that

$$\begin{aligned}\mathbf{q}_1 \circ \mathbf{F}_1 &= DF_1^{*-1}\hat{\mathbf{q}}_1, \\ \mathbf{q}_2 \circ \mathbf{F}_2 &= DF_2^{*-1}\hat{\mathbf{q}}_2.\end{aligned}\tag{A.15}$$

On the other hand, by using Lemma 9, we obtain

$$\boldsymbol{\nu} = DF_1^{*-1}\hat{\boldsymbol{\nu}} = DF_2^{*-1}\hat{\boldsymbol{\nu}},$$

which implies that

$$\begin{aligned}\mathbf{q}_1 \circ \mathbf{F}_1 \times \boldsymbol{\nu} &= DF_1^{*-1}\hat{\mathbf{q}}_1 \times DF_1^{*-1}\hat{\boldsymbol{\nu}}, \\ \mathbf{q}_2 \circ \mathbf{F}_2 \times \boldsymbol{\nu} &= DF_2^{*-1}\hat{\mathbf{q}}_2 \times DF_2^{*-1}\hat{\boldsymbol{\nu}}.\end{aligned}\tag{A.16}$$

Now, we have:

$$DF_1^{*-1}\hat{\mathbf{q}}_1 \times DF_1^{*-1}\hat{\boldsymbol{\nu}} = T(DF_1^{*-1}\hat{\mathbf{q}}_1)DF_1^{*-1}\hat{\boldsymbol{\nu}}.$$

By using $B = \nabla F_1^*$ and $U = \hat{\mathbf{q}}_1$ in Lemma 7, we obtain:

$$T(DF_1^{*-1}\hat{\mathbf{q}}_1) = \det(DF_1^{-1})DF_1T(\hat{\mathbf{q}}_1)DF_1^*.$$

So, we have:

$$\begin{aligned}DF_1^{*-1}\hat{\mathbf{q}}_1 \times DF_1^{*-1}\hat{\boldsymbol{\nu}} &= \det(DF_1^{-1})DF_1T(\hat{\mathbf{q}}_1)DF_1^*DF_1^{*-1}\hat{\boldsymbol{\nu}}, \\ &= \det(DF_1^{-1})DF_1T(\hat{\mathbf{q}}_1)\hat{\boldsymbol{\nu}}, \\ &= \det(DF_1^{-1})DF_1\hat{\mathbf{q}}_1 \times \hat{\boldsymbol{\nu}}.\end{aligned}\tag{A.17}$$

We finally obtain:

$$DF_1^{*-1}\hat{\mathbf{q}}_1 \times DF_1^{*-1}\hat{\boldsymbol{\nu}} = \det(DF_1^{-1})DF_1\hat{\mathbf{q}}_1 \times \hat{\boldsymbol{\nu}},$$

and, in the same way:

$$DF_2^{*-1}\hat{\mathbf{q}}_2 \times DF_2^{*-1}\hat{\boldsymbol{\nu}} = \det(DF_2^{-1})DF_2\hat{\mathbf{q}}_2 \times \hat{\boldsymbol{\nu}}.$$

We can now prove Proposition 7.

- Obviously, \mathbf{q} belongs to $L^2(\Omega)$.

- Let us show that $\nabla \times \mathbf{q}$ also belongs to $L^2(\Omega)$.

We can equivalently show that:

$$\mathbf{q}_1 \circ \mathbf{F}_1 \times \boldsymbol{\nu} = \mathbf{q}_2 \circ \mathbf{F}_2 \times \boldsymbol{\nu} \text{ on } \hat{\Gamma}.$$

Following Lemma 10, $\mathbf{q}_1 \circ \mathbf{F}_1 \times \boldsymbol{\nu} = \mathbf{q}_2 \circ \mathbf{F}_2 \times \boldsymbol{\nu}$ on $\hat{\Gamma}$ if and only if:

$$\det(DF_1^{-1})DF_1\hat{\mathbf{q}}_1 \times \hat{\boldsymbol{\nu}} = \det(\nabla F_2^{-1})DF_2\hat{\mathbf{q}}_2 \times \hat{\boldsymbol{\nu}}.$$

Now, $\forall \hat{\mathbf{x}} \in \hat{\Gamma}$, $\hat{\mathbf{q}}_1(\hat{\mathbf{x}}) \times \hat{\boldsymbol{\nu}}(\hat{\mathbf{x}})$ is a vector tangent to the plane $\hat{\Gamma}$ at the point $\hat{\mathbf{x}}$ that we denote as $\hat{P}(\hat{\mathbf{x}})$. So is $\hat{\mathbf{q}}_2 \times \hat{\boldsymbol{\nu}}$. Since $\hat{\mathbf{q}} \in H(\mathbf{curl}, \hat{\Omega})$, we have in \hat{P} :

$$\hat{\mathbf{q}}_1 \times \hat{\boldsymbol{\nu}} = \hat{\mathbf{q}}_2 \times \hat{\boldsymbol{\nu}}.$$

On the other hand, since $F_1|_{\hat{\Gamma}} = F_2|_{\hat{\Gamma}} = F$ we have:

$$DF_1|_{\hat{P}} = DF_2|_{\hat{P}}.$$

In the same way, we have:

$$\det(DF_1|_{\hat{P}}) = \det(DF_2|_{\hat{P}}).$$

So, the identity

$$\hat{\mathbf{q}}_1 \times \hat{\boldsymbol{\nu}} = \hat{\mathbf{q}}_2 \times \hat{\boldsymbol{\nu}}$$

implies that

$$\det(DF_1^{-1})DF_1\hat{\mathbf{q}}_1 \times \hat{\boldsymbol{\nu}} = \det(DF_2^{-1})DF_2\hat{\mathbf{q}}_2 \times \hat{\boldsymbol{\nu}}.$$

In other words

$$\mathbf{q}_1 \times \boldsymbol{\nu} = \mathbf{q}_2 \times \boldsymbol{\nu},$$

which proves that $\mathbf{q} \in H(\mathbf{curl}, \Omega)$.

Remarks:

1. These properties were proved in the same way for the 2D case but we restricted this appendix to the 3D one, which is the more useful.
2. Of course, a similar study can be carried out for the $H(\mathbf{div})$ -conforming transform.

Bibliography

1. S. ABARBANEL, D. GOTTLIEB, *A mathematical analysis of the PML method*, J. Comput. Phys. **134** (2), pp. 357–363, 1997.
2. S. ABARBANEL, D. GOTTLIEB, J. S. HESTHAVEN, *Well-posed perfectly matched layers for advective acoustics* J. Comput. Phys. **154** (2), pp. 266–283, 1999.
3. J. D. ACHENBACH, *Wave propagation in elastic solids*, North-Holland, 1984.
4. R. M. ALFORD, K. R. KELLY, D. M. BOORE, *Accuracy of finite difference modeling of the acoustic wave equation*, Geophysics **39** (6) 834–842, 1974.
5. L. ANNÉ, *Schémas d'ordre élevé pour l'équation des ondes acoustiques en milieu hétérogène 3D*, Thesis, Université Paris VI, 1996.
6. L. ANNÉ, P. JOLY, Q. H. TRAN, *Construction and analysis of higher order finite difference schemes for the 1D wave equation*, Computational Geosciences **4**, pp. 207–249, 2000.
7. F. ASSOUS, P. CIARLET JR., J. SEGRÉ, *Numerical solution to the time-dependent Maxwell equations in two-dimensional singular domains: the singular complement method*, J. Comput. Phys. **161** (1), pp. 218–249, 2000.
8. F. ASSOUS, P. DEGOND, E. HEINTZE, P.-A. RAVIART, J. SEGRÉ, *On a finite-element method for solving the three-dimensional Maxwell equations*, J. Comput. Phys. **109** (2), pp. 222–237, 1993.
9. B. A. AULD, *Acoustic fields and waves in solids*, vol. 2, R.E. Krieger ed., 1990.
10. G. A. BAKER, *Error estimates for finite element methods for the second order hyperbolic equations*, SIAM J. Num. Anal. **13** (4), pp. 564–576, 1976.
11. G. A. BAKER, V. A. DOUGALIS, *The effect of quadrature errors on finite element approximations for the second order hyperbolic equations*, SIAM J. Num. Anal. **13** (4), pp. 577–598, 1976.
12. G. A. BAKER, P. GRAVES-MORRIS, *Padé approximants*. 2nd ed., Encyclopedia of Mathematics and Its Applications **59**, Cambridge University Press, Cambridge, 1996.
13. A. BAMBERGER, G. CHAVENT, P. LAILLY, *Étude de schémas numériques pour les équations de l'élastodynamique linéaire*, INRIA Report RR-0041, 1980.
14. M. BARBIÉRA, *Schémas aux différences finies d'ordre 4 pour les équations de l'élastodynamique linéaire*, Thesis, Université Paris IX-Dauphine, 1993.
15. M. BARBIÉRA, G. COHEN, *A scheme, fourth-order in space and time, for the 2-D linearized elastodynamics system*, Proc. 2nd Conf. Numerical and Mathematical Methods for Wave Propagation, SIAM, pp. 39–47, Newark, DE, June 1993.
16. A. BAYLISS, A. JORDAN, K. E. LEMESURIER, B. TURKEL, *A fourth order accurate finite difference scheme for the computation of elastic waves*, Bull. Seism. Soc. Am., 1987.
17. E. BÉCACHE, P. JOLY, C. TSOGKA, *Éléments finis mixtes et condensation de masse en élastodynamique linéaire. (I) Construction*, CRAS Math. **325**, série I, pp. 545–550, 1997.

18. A. BENDALI, L. HALPERN, *Conditions aux limites absorbantes pour le système de Maxwell dans le vide en dimension trois d'espace*, CRAS Math. **307**, série I, pp. 1011–1013, 1988.
19. J.-P. BÉRENGER, *A perfectly matched layer for the absorption of electromagnetic waves*, J. Comput. Phys. **114** (2), pp. 185–200, 1994.
20. J.-P. BÉRENGER, *Three-dimensional perfectly matched layer for the absorption of electromagnetic waves*, J. Comput. Phys. **127** (2), pp. 363–379, 1996.
21. C. CERJAN, D. KOSLOFF, R. KOSLOFF, M. RESHEF, *A nonreflecting boundary condition for discrete acoustic and elastic wave equations*, Geophysics **50**, pp. 705–708, 1985.
22. W. C. CHEW, W. H. WEEDON, *A 3D perfectly matched medium from modified Maxwell equations with stretched coordinates*, IEEE Microwave and Optic. Tech. Letters **7** (13), pp. 599–604, 1999.
23. M. J. S. CHIN-JOE-KONG, W. A. MULDER, M. VAN VELDHUIZEN, *Higher-order triangular and tetrahedral finite elements with mass lumping for solving the wave equation*, J. Eng. Math. **35** (4), pp. 405–426, 1999.
24. P. G. CIARLET, *The finite element method for elliptic problems*, North-Holland, 1987, Exercises 4.1.4 and following.
25. P. G. CIARLET, J.-L. LIONS, *Handbook of numerical analysis*, Vol. 2, North-Holland, 1991.
26. J.-P. CIONI, L. FÉZOUI, H. STÈVE, *A parallel time-domain Maxwell solver using upwind schemes and triangular meshes*, IMPACT in Computing Science and Engineering **5**, pp. 215–247, 1993.
27. G. CLAYTON, B. ENGQUIST, *Absorbing boundary conditions for acoustic and elastic wave equations*, Bull. Seism. Soc. Am. **71**, pp. 1529–1540, 1977.
28. G. COHEN, *A class of schemes, fourth order in space and time, for the 2D wave equation*, Proc. 6th IMACS Internat. Symp. on Computer Methods for Partial Differential Equations, Bethlehem, PA, USA, 23–27 June 1987.
29. G. COHEN, *Méthodes numériques d'ordre élevé pour les ondes en régime transitoire*, Collection didactique Vol. 12, 1ère section, INRIA ed., 1994.
30. G. COHEN, A. ELMKIES, *Eléments finis triangulaires P_2 avec condensation de masse pour l'équation des ondes*, INRIA Report RR-2418, 1994.
31. G. COHEN, S. FAUQUEUX, *Mixed finite elements with mass-lumping for the transient wave equation*, J. Comp. Acous. **8** (1), pp. 171–188, 2000.
32. G. COHEN, S. FAUQUEUX, *Efficient mixed finite elements for the acoustics equation*, Proc. 70th SEG Annual International Meeting and Exposition, Calgary, Aug. 2000.
33. G. COHEN, S. FAUQUEUX, *2D elastic modeling with efficient mixed finite elements*, Proc. 63th EAGE Meeting and Exposition, Amsterdam, June 2001.
34. G. COHEN, P. JOLY, *Fourth order schemes for the heterogeneous acoustics equation*, Comp. Meth. in Appl. Mech. and Engin. **80** (1–3), pp. 397–407, North-Holland, 1990.
35. G. COHEN, P. JOLY, *Construction and analysis of fourth-order finite difference schemes for the acoustic wave equation in inhomogeneous media*, SIAM J. Numer. Anal. **33** (4), pp. 1266–1302, 1996.
36. G. COHEN, P. JOLY, N. TORDJMAN, *Higher-order finite elements with mass lumping for the 1-D wave equation*. Finite Elements in Analysis and Design **17** (3-4), pp. 329–336, 1994.
37. G. COHEN, P. JOLY, J. E. ROBERTS, N. TORDJMAN, *Higher order triangular finite elements with mass lumping for the wave equation*. SIAM J. Numer. Anal., in press.
38. G. COHEN, P. MONK, *Gauss point mass lumping schemes for Maxwell's equations*, NMPDE Journal **14** (1), pp. 63–88, 1998.

39. G. COHEN, P. MONK, *Mur-Nédélec finite element schemes for Maxwell's equations*, Comp. Meth. in Appl. Mech. Eng. **169** (3–4), pp. 197–217, 1999.
40. F. COLLINO, P. MONK, *The perfectly matched layer in curvilinear coordinates*, SIAM J. Sci. Comput. **19** (6), pp. 2061–2090, 1998.
41. F. COLLINO, C. TSOGKA, *Application of the PML absorbing layer model to the linear elastodynamic problem in anisotropic heterogeneous media*, INRIA Report 0249-6399, 1998, to appear in Geophysics.
42. M. A. DABLAINE, *The application of high order differencing for the scalar wave equation*, Geophysics **51** (1), pp. 54–66, 1986.
43. R. DAUTRAY, J.-L. LIONS, *Mathematical analysis and numerical methods for science and technology*, Vol. 2, Springer-Verlag, 1988.
44. R. DAUTRAY, J.-L. LIONS, *Mathematical analysis and numerical methods for science and technology*, Vol. 5, Springer-Verlag, 1990.
45. P. J. DAVIS, P. RABINOWITZ, *Methods of numerical integration*, 2nd edn., Academic Press, 1984.
46. E. J. DEAN, R. GLOWINSKI, *A wave equation approach to the numerical solution of the Navier-Stokes equations for incompressible viscous flow*, CRAS Math. **325** (7), série I, pp. 783–791, 1997.
47. T. DEVÈZE, *Contribution à l'analyse, par différences finies, des équations de Maxwell dans le domaine temps*, Thesis, May 1992.
48. M. DUBINER, *Spectral methods on triangles and other domains*, J. Sci. Comput. **6** (4), pp. 345–390, 1991.
49. F. DUBOIS, *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** (4), pp. 1103–1142, 1990.
50. A. ELMKIES, *Sur les éléments finis d'arête pour la résolution des équations de Maxwell en milieu anisotrope et pour des maillages quelconques*, Thesis, Université de Paris-XI Orsay, 1998.
51. A. ELMKIES, P. JOLY, *Éléments finis d'arête et condensation de masse pour les équations de Maxwell: le cas 2D*, CRAS Math. **324**, série I, pp. 1287–1293, 1997.
52. A. ELMKIES, P. JOLY, *Éléments finis d'arête et condensation de masse pour les équations de Maxwell: le cas de la dimension 3*, CRAS Math. **325**, série I, pp. 1217–1222, 1997.
53. B. ENGQUIST, A. MAJDA, *Absorbing boundary conditions for the numerical simulation of waves*, Math. Comp. **31** (139), pp. 629–651, 1977.
54. C. A. ERINGEN, E. S. SUHUBI, *Elastodynamics*, Vols. 1 and 2, Academic Press, 1975.
55. S. FAUQUEUX, *Modélisation de la propagation d'ondes en milieu élastique par éléments finis mixtes avec condensation de masse*, Thesis, Université de Paris IX-Dauphine, in press.
56. R. P. FEYNMAN, R. B. LEIGHTON, M. SAND, *The Feynman lectures on physics*, Addison-Wesley, 1963.
57. B. FRAEIJJS DE VEUBEKE, *Variational principles and the patch test*, Int. J. Num. Meth. Eng. **8**, pp. 783–801, 1974.
58. C. W. GEAR, *Numerical initial value problems in ordinary differential equations*, Prentice Hall, 1971.
59. D. GIVOLI, *Non-reflecting boundary conditions*, J. Comp. Phys. **94** (1), pp. 1–29, 1991.
60. M. GOLDBERG, E. TADMOR, *New stability criteria for difference approximations of hyperbolic problems*, Lectures in Appl. Math. **22** (1), pp. 177–192, 1985.

61. B. GUSTAFSSON, H. O. KREISS, J. OLIGER, *Time dependent problems and difference methods*. Pure and Applied Mathematics. Wiley, New York, 1995.
62. E. HAIRER, S. P. NØRSETT, G. WANNER, *Solving ordinary differential equations*, Springer Series in Computational Mathematics, 1991.
63. Y. HAUGAZEAU, P. LACOSTE, *Lumping of the mass matrix for 1st-order mixed finite elements of $H(\text{curl})$* , Proc. 2nd Conf. on Numerical and Mathematical Methods for Wave Propagation, SIAM, pp. 39–47, Newark, DE, June 1993.
64. J.-P. HENNART, *Topics in finite element discretization of parabolic evolution problems*, Lecture Notes in Mathematics **909**, pp. 185–199, Springer-Verlag, 1982.
65. J.-P. HENNART, E. SAINZ, M. VILLEGAS, *On the efficient use of the finite element method in static neutron diffusion calculations*, Computational Methods in Nuclear Engineering **1**, pp. 3–87, 1979.
66. J. S. HESTHAVEN, *On the analysis and construction of perfectly matched layers for the linearized Euler equations*, J. Comput. Phys. **142** (1), pp. 129–147, 1998.
67. R. L. HIGDON, *Initial boundary value problems for hyperbolic systems*, SIAM Review **28**, pp. 115–135, 1977.
68. O. HOLBERG, *Large scale wave equation computations using higher-order difference operators*, Thesis, U. of Trondheim, 1987.
69. O. HOLBERG, *Computational aspects of the choice of operators and sampling interval for numerical differentiation in large-scale simulation of wave phenomena*, Geophysical Prospecting **35**, pp. 625–655, 1987.
70. R. HOLLAND, *Finite difference solution of initial value problems involving Maxwell's equations in isotropic media*, IEEE Trans. on Nucl. Sci. **30** (6), 1983.
71. T. HUGHES, *The finite element method: linear static and dynamic finite element analysis*, Prentice-Hall, 1987.
72. B. M. IRONS, *Un nouvel élément fini de coques générales- "Semiloof"*, Computing Methods in Applied Science and Engineering, Part 1, pp. 177–192, Lecture Notes in Computer Science, Springer, Berlin, 1974.
73. E. JONES, K. B. OLSEN, *Three-dimensional finite-difference modeling of non-linear ground motion*, Proc. CUREe Northridge Earthquake Research Conference, Los Angeles, August 20–22, 1997.
74. J. B. KELLER, F. ODEH, *Partial differential equations with periodic coefficients and Bloch waves in crystals*, J. Mathematical Physics **5**, pp. 1499–1504, 1964.
75. D. KOMATITSCH, J.-P. VILOTTE, *The Spectral element method: an efficient tool to simulate the seismic response of 2D and 3D geological structures*, Bull. Seis. Soc. Am. **88** (2), pp. 368–392, 1998.
76. H. O. KREISS, *Initial boundary value problems for hyperbolic systems*, Comm. Pure Appl. Math. **23**, pp. 277–298, 1970.
77. H. LAMB, *Hydrodynamics*, Cambridge University Press, 1974.
78. M. LASSAS, E. SOMERSALO, *On the existence and convergence of the solution of PML equations*, Computing **60** (3), pp. 229–241, 1998.
79. R. LEIS, *Initial boundary value problems in mathematical physics*, Wiley, New York, 1988.
80. A. R. LEVANDER, *Fourth-order velocity-stress finite difference scheme*, Proc. 57th SEG Annual International Meeting and Exposition, New Orleans, Oct. 1987.
81. J.-L. LIONS, E. MAGENES, *Problèmes aux limites non homogènes et applications*, Vol. 1, Dunod, 1968.

82. R. MADARIAGA, *Dynamics of an expanded circular fault*, Bull. Seismol. Soc. Am. **66**, pp. 639–666, 1976.
83. Y. MADAY, A.T. PATERA, *Spectral element methods for the incompressible Navier-Stokes equations*, State of the Art Survey in Computational Mechanics, ed. A. K. Noor, pp. 71–143, 1989.
84. Y. MADAY, E.M. RONQUIST, *Optimal error analysis of spectral methods with emphasis on non-constant coefficients and deformed geometries*, Comp. Meth. in Appl. Mech. Eng. **80**, pp. 91–115, 1990.
85. J. MÉTRAL, O. VACUS, *Caractère bien posé du problème de Cauchy pour le système de Bérenger*, CRAS Math. **328** (10), série I, pp. 847–852, 1999.
86. R. MITTRA, U. PEKEL, J. VEIHL, *A theoretical and numerical study of Berenger's perfectly matched layer (PML) concept for mesh truncation in time and frequency domains*, Approximations and numerical methods for the solution of Maxwell's equations, pp. 1–19, Oxford, 1995.
87. P. MOCZO, M. LUCKA, J. KRISTEK, M. KRISTEKOVA, *3D displacement finite differences and a combined memory optimization*, Bull. Seism. Soc. Am. **89**, pp. 69–79, 1999.
88. W. A. MULDER, *A comparison between higher-order finite elements and finite differences for solving the wave equation*, Numerical Methods in Engineering '96, Wiley, pp. 344–350, 1996.
89. W. A. MULDER, *Construction and application of higher-order mass-lumped finite elements for the wave equation*, J. Comput. Acoustics, in press.
90. G. MUR, *Absorbing boundary conditions for the finite-difference approximation of the time-domain electromagnetic-field equations*, IEEE Trans. on Electromagnetic Compatibility **21** (4), pp. 377–382, 1981.
91. G. MUR, A. T. DE HOOP, *A finite-element method for computing three-dimensional electromagnetic fields in inhomogeneous media*, IEEE Trans. Magnetics **MAG-21**, pp. 2188–2191, 1985.
92. J.-C. NÉDÉLEC, *Mixed finite elements in \mathbb{R}^3* , Numer. Math. **35** (3), pp. 315–341, 1980.
93. J.-C. NÉDÉLEC, *A new family of mixed finite elements in \mathbb{R}^3* , Numer. Math. **50** (1), pp. 57–81, 1986.
94. L. NICOLÉTIS, *Simulation numérique de la propagation d'ondes sismiques*, Thesis, U. Pierre and Marie Curie, Paris, 1981.
95. K. B. OLSEN, R. MADARIAGA, R. J. ARCHULETA, *Three-dimensional dynamic simulation of the 1992 Landers earthquake*, Science **278**, pp. 834–838, 1997.
96. A. RALSTON, P. RABINOWITZ, *A first course in numerical analysis*, 2nd edn., International Series in Pure and Applied Mathematics, McGraw-Hill, 1978.
97. P.-A. RAVIART, *The use of numerical integration in finite element methods for parabolic equations*, Topics in Numerical Analysis, Proc. Royal Irish Academy Conference on Numerical Analysis, J. J. H. Miller ed., Academic Press, pp. 233–264, 1973.
98. D. REDFERN, *The MAPLE handbook, MAPLE V release 4*, corr. second print, Springer-Verlag, 1996.
99. J. E. ROBERTS, J.-M. THOMAS, *Mixed and hybrid methods*, Handbook of Numerical Analysis, Vol. II, pp. 523–639, P. G. Ciarlet and J.-L. Lions eds., North-Holland, Amsterdam, 1991.
100. W. RUDIN, *Functional analysis*, McGraw-Hill, 1991.
101. L. SCHWARTZ, *Théorie des distributions*, Publications de l'Institut de Mathématique de l'Université de Strasbourg **IX–X**, new edition, Hermann, Paris 1966.
102. A. SEI, *Étude de schémas numériques pour des modèles de propagation d'ondes en milieu hétérogène*, Thesis, Université Paris IX-Dauphine, 1991.

103. G. SERIANI, E. PRIOLO, *Spectral element method for acoustic wave simulation in heterogeneous media*, Finite Elements in Analysis and Design **16** (3-4), pp. 337–348, 1994.
104. V. SHANKAR, W. HALL, A. MOHAMMADIAN, *A time-domain differential solver for electromagnetic problems*, Proc. IEEE **77** (5), pp. 709–721, 1989.
105. G. R. SHUBIN, J. B. BELL, *A modified equation approach to constructing fourth order methods for acoustic wave propagation*, SIAM J. Sci. Comput. **8** (2), pp. 135–151, 1987.
106. G. STRANG, G. J. FIX, *An analysis of the finite element method*, Prentice-Hall, 1973.
107. J. C. STRIKWERDA, *Finite difference schemes and partial differential equations*, Cole advanced books and software, Wadsworth & Brooks, 1989.
108. A. H. STROUD, *Approximate calculation of multiple integrals*, Prentice Hall, 1971.
109. A. TAFLOVE, M. E. BRODWIN, *Numerical solutions of steady state electromagnetic scattering problems using the time dependent Maxwell's equations*, IEEE M.T.T. **23** (8), 1975.
110. A. TAFLOVE, *Computational electrodynamics: The finite-difference time-domain method*, Artech House, Boston, 1995.
111. H. TAL-EZER, *Spectral methods in time for hyperbolic equations*, SIAM J. Numer. Anal. **23** (1), pp. 11–26, 1986.
112. C. K. W. TAM, L. AURIAULT, F. CAMBULI, *Perfectly matched layer as an absorbing boundary condition for the linearized Euler equations in open and ducted domains*, J. Comput. Phys. **144** (1), pp. 213–234, 1998.
113. M. E. TAYLOR, *Partial differential equations*, Vols. 1–3, Springer-Verlag, 1990.
114. J-M. THOMAS, *Sur l'analyse numérique de méthodes d'éléments finis hybrides et mixtes.*, Thèse d'État, Université de Paris VI, May 1977.
115. N. TORDJMAN, *Éléments finis d'ordre élevé avec condensation de masse pour l'équation des ondes*, Thesis, Université de Paris IX-Dauphine, Jan. 1995.
116. L.N. TREFETHEN, *Group velocity in finite difference schemes*, SIAM Rev. **24** (2), pp. 113–136, 1982.
117. J. TUOMELA, *On the construction of arbitrary order schemes for the many-dimensional wave equation*, BIT **36** (1), pp. 158–165, 1996.
118. R. VICHNEVETSKY, J. B. BOWLES, *Fourier analysis of numerical approximation of hyperbolic equation*, SIAM Series in Appl. Math., Philadelphia, 1982.
119. J. VIRIEUX, *P-SV wave propagation in heterogeneous media: Velocity-stress finite difference method*, Geophysics **51** (4), 1986.
120. K. YEE, *Numerical solutions of initial boundary value problems involving Maxwell's equations in isotropic media*, IEEE Trans. on Antennas and Propagation AP-16, pp. 302–307, 1966.
121. L. C. YOUNG, *An efficient finite element method for reservoir simulation*, Proc. 53rd Annual Fall Technical Conference and Exhibition of the Society of Petroleum Engineers of AIME, Houston, Texas, Oct. 1–3, 1978.
122. J. ZAHRADNIK, P. O'LEARY, J. SOCHACKI, *Finite-difference schemes for elastic waves based on the integration approach*, Geophysics **59**, pp. 928–937, 1994.
123. L. ZHAO, A.C. CANGELLARIS, *GT-PML: Generalized theory of perfectly matched layers and its application to reflectionless truncation of finite-difference time-domain grids*, IEEE Trans. Microwave Theory and Tech. **44** (1996), pp. 2555–2563.
124. O. C. ZIENKIEWICZ, R. L. TAYLOR, *The finite element method*, McGraw-Hill, 1991.

Index

- absorbing boundary condition (ABC)
 - 307
- acoustic impedance 146
- Arakawa's scheme 40, 48
- boundary condition
 - Dirichlet 11, 16, 133–134
 - displacement 12, 136
 - Neumann 11, 19, 133–135
 - perfectly conducting 17, 20
 - Silver-Müller 12, 21, 24, 136
 - traction 12, 136
- CFL 56, *see* stability conditions
- conforming transform
 - $H(\text{curl})$ 272, 331–339
 - $H(\text{div})$ 273
- damping layer 307
- Descarte's law 156, 157
- dispersion curves 111, 114–118, 120, 193, 194
- dispersion relation
 - continuous 26
 - discrete
 - finite difference 65–81
 - finite elements 179–186, 222–232
- dispersive scheme 103
- edge elements 18
 - hexahedral
 - first family 261–269
 - second family 273–283
 - quadrilateral
 - first family 269–271
 - second family 283–285
 - tetrahedral 289–292
 - triangular 286–289
- elastics system
 - boundary conditions 12–13
 - energy identity 24
 - finite difference approximation 60–62
 - boundary conditions 136
 - dispersion curves 118–120
 - dispersion relation 80–81
 - stability 97–99
 - finite element approximation 301–305
 - formulation 7–11
 - perfectly matched layer 315–317
 - plane wave analysis 29–30
 - variational formulation 22
- energy identity 22–24, 142–143
- FDM *see* finite difference method
- FEM *see* finite element method
- Fourier transform
 - in space 25, 66
 - in time 25, 66
- Gauss-Lobatto quadrature rules 173–177
- group velocity 101
- $H(\text{curl})$ 17
- $H(\text{div})$ 17
- image principle 135
- isoparametric finite elements 216
- isotropy curves 111, 118
- Jacobian 216
- Jacobian matrix 216
- leapfrog scheme 45, 129–130

- mass matrix 171
- mass-lumping
 - in 1D 171
 - in 2D 211–222
- Maxwell equations
 - boundary conditions 12
 - energy identity 23–24
 - finite difference approximation 58–60
 - boundary conditions 136
 - dispersion relation 78–80
 - stability 97
 - finite element approximation 261–293
 - formulation 4–7
 - functional spaces 17–18
 - perfectly matched layer 315
 - plane wave analysis 26–29
 - variational formulation 20–22
- mixed finite elements
 - hexahedral
 - first family 261–269
 - second family 273–283
 - quadrilateral
 - first family 269–271
 - second family 283–285
- modified equation approach 46–48, 56–57, 62–63, 130–132, 178–179
- numerical dispersion 102
- numerical dispersion coefficient 101
- numerical dissipation 84
- P-wave 10, 30
- perfectly matched layer 308
 - for the elastics system
 - in 2D 315–317
 - for the Maxwell equations
 - in 2D 315
 - for the wave equation
 - in 2D 311–315
 - in 3D 317–320
- phase velocity 101
- PML *see* perfectly matched layer
- positivity of an operator 84–85, 137–142
- pressure wave *see* P-wave
- pulsation 26
- Rayleigh wave 13
- reflection-transmission analysis
 - for continuous equations
- in 1D 145–146
- in 2D 156–158
- for finite difference methods
- in 1D 146–156
- in 2D 158–161
- for finite element methods
- in 1D 202–208
- in 2D 238–244
- S-wave 11, 30
- shear wave *see* S-wave
- Snell's law 156
- Sobolev spaces 15–17
- spectral elements
 - hexahedral 220–222
 - new formulation
 - elastics system 301–305
 - wave equation 294–301
 - non-conforming 257–259
 - quadrilateral 214–220
 - reference 211–214
 - tetrahedral 254–257
 - triangular 250–253
- stability
 - finite difference 85–99, 137–144
 - finite elements 186–192, 231–234, 271, 283, 288, 292
- stability conditions
 - finite difference 87
 - finite elements 188
- stiffness integral 18
- stiffness matrix 171
- stiffness term 18
- symbol of an operator 68
- symmetric schemes 48–50, 57
- TE *see* transverse-electric
- TM *see* transverse-magnetic
- transparent condition 307
- transverse-electric 6, 269–271, 283
- transverse-magnetic 6, 59, 269, 271, 283, 284, 286, 315, 323
- variational formulation 18–22
- vertex element 291
- von Neumann stability condition 84
- wave vector 26
- weak formulation 19
- Yee scheme 58–60
 - variational extensions 261–271



Location: <http://www.springer.de/phys/>

**You are one click away
from a world of physics information!**

Come and visit Springer's

Physics Online Library

Books

- Search the Springer website catalogue
- Subscribe to our free alerting service for new books
- Look through the book series profiles

You want to order?

Email to: orders@springer.de

Journals

- Get abstracts, ToC's free of charge to everyone
- Use our powerful search engine LINK Search
- Subscribe to our free alerting service LINK Alert
- Read full-text articles (available only to subscribers of the paper version of a journal)

You want to subscribe?

Email to: subscriptions@springer.de

You have a question on
an electronic product?

Electronic Media
• Get more information on our software and CD-ROMs
Email to: helpdesk-em@springer.de

..... ● Bookmark now:

**http://
www.springer.de/phys/**

Springer · Customer Service
Haberstr. 7 · 69126 Heidelberg, Germany
Tel: +49 (0) 6221 - 345 - 217/8
Fax: +49 (0) 6221 - 345 - 229 · e-mail: orders@springer.de
d&p - 6437.MNT/SFb



Springer