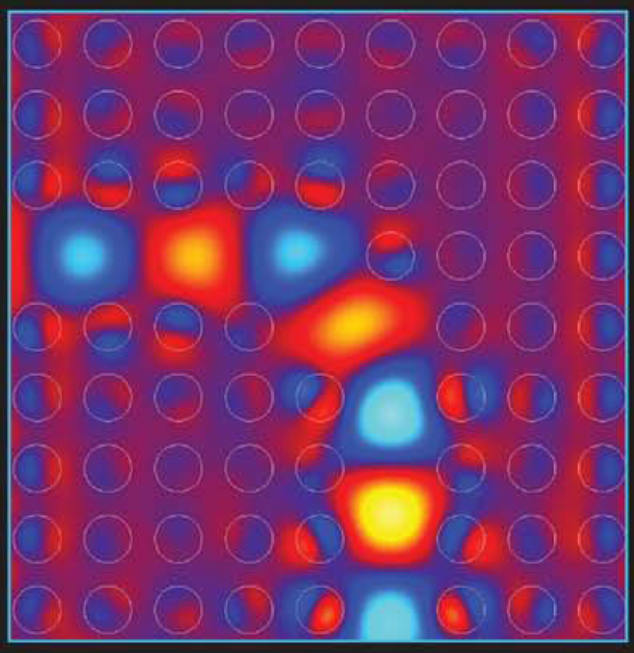
NANOSTRUCTURE SCIENCE AND TECHNOLOGY
Series Editor: David J. Lockwood

Computational Methods for Nanoscale Applications

Particles, Plasmons and Waves



Igor Tsukerman

Computational Methods for Nanoscale Applications

Nanostructure Science and Technology

Series Editor: David J. Lockwood, FRSC
National Research Council of Canada
Ottawa, Ontario, Canada

Current volumes in this series:

Functional Nanostructures: Processing, Characterization and Applications
Edited by Sudipta Seal

Light Scattering and Nanoscale Surface Roughness
Edited by Alexei A. Maradudin

Nanotechnology for Electronic Materials and Devices
Edited by Anatoli Korkin, Evgeni Gusev, and Jan K. Labanowski

Nanotechnology in Catalysis, Volume 3
Edited by Bing Zhou, Scott Han, Robert Raja, and Gabor A. Somorjai

Nanostructured Coatings
Edited by Albano Cavaleiro and Jeff T. De Hosson

Self-Organized Nanoscale Materials
Edited by Motonari Adachi and David J. Lockwood

Controlled Synthesis of Nanoparticles in Microheterogeneous Systems
Vincenzo Turco Liveri

Nanoscale Assembly Techniques
Edited by Wilhelm T.S. Huck

Ordered Porous Nanostructures and Applications
Edited by Ralf B. Wehrspohn

Surface Effects in Magnetic Nanoparticles
Dino Fiorani

Interfacial Nanochemistry: Molecular Science and Engineering at Liquid-Liquid Interfaces
Edited by Hitoshi Watarai

Nanoscale Structure and Assembly at Solid-Fluid Interfaces
Edited by Xiang Yang Liu and James J. De Yoreo

Introduction to Nanoscale Science and Technology
Edited by Massimiliano Di Ventra, Stephane Evoy, and James R. Heflin Jr.

Alternative Lithography: Unleashing the Potentials of Nanotechnology
Edited by Clivia M. Sotomayor Torres

Semiconductor Nanocrystals: From Basic Principles to Applications
Edited by Alexander L. Efros, David J. Lockwood, and Leonid Tsybeskov

Nanotechnology in Catalysis, Volumes 1 and 2
Edited by Bing Zhou, Sophie Hermans, and Gabor A. Somorjai

(Continued after index)

Igor Tsukerman

Computational Methods for Nanoscale Applications

Particles, Plasmons and Waves

Springer

Igor Tsukerman
Department of Electrical
and Computer Engineering
The University of Akron
Akron, OH 44325-3904
USA
igor@uakron.edu

Series Editor
David J. Lockwood
National Research Council of Canada
Ottawa, Ontario
Canada

ISBN: 978-0-387-74777-4 e-ISBN: 978-0-387-74778-1
DOI: 10.1007/978-0-387-74778-1

Library of Congress Control Number: 2007935245

© 2008 Springer Science+Business Media, LLC
All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.
The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Cover Illustration: Real part of the electric field phasor in the Fujisawa-Koshiba photonic waveguide bend.
From "Electromagnetic Applications of a New Finite-Difference Calculus", by Igor Tsukerman, IEEE Transactions on Magnetics, Vol. 41, No. 7, pp. 2206–2225, 2005.
© 2005 IEEE (by permission).

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

To the memory of my mother,
to my father,
and to the miracle of M.

Preface

> The purpose of this note . . . is to
> sort out my own thoughts . . .
> and to solicit ideas from others.
>
> Lloyd N. Trefethen
> Three mysteries of Gaussian elimination

Nobody reads prefaces. Therefore my preference would have been to write a short one that nobody will read rather than a long one that nobody will read. However, I ought to explain, as briefly as possible, the main motivation for writing the book and to thank – as fully and sincerely as possible – many people who have contributed to this writing in a variety of ways.

My motivation has selfish and unselfish components. The *un*selfish part is to present the elements of computational methods and nanoscale simulation to researchers, scientists and engineers who are not necessarily experts in computer simulation. I am hopeful, though, that parts of the book will also be of interest to experts, as further discussed in the Introduction and Conclusion.

The selfish part of my motivation is articulated in L. N. Trefethen's quote above. Whether or not I have succeeded in "sorting out my own thoughts" is not quite clear at the moment, but I would definitely welcome "ideas from others," as well as comments and constructive criticism.

* * *

I owe an enormous debt of gratitude to my parents for their incredible kindness and selflessness, and to my wife for her equally incredible tolerance of my character quirks and for her unwavering support under all circumstances. My son (who is a business major at The Ohio State University) proofread parts of the book, replaced commas with semicolons, single quotes with double quotes, and fixed my other egregious abuses of the English language.

Overall, my work on the book would have been an utterly pleasant experience had it not been interrupted by the sudden and heartbreaking death of my mother in the summer of 2006. I do wish to dedicate this book to her memory.

Acknowledgment and Thanks

Collaboration with Gary Friedman and his group, especially during my sabbatical in 2002–2003 at Drexel University, has influenced my research and the material of this book greatly. Gary's energy, enthusiasm and innovative ideas are always very stimulating.

During the same sabbatical year, I was fortunate to visit several research groups working on the simulation of colloids, polyelectrolytes, macro- and biomolecules. I am very grateful to all of them for their hospitality. I would particularly like to mention Christian Holm, Markus Deserno and Vladimir Lobaskin at the Max-Planck-Institut für Polymerforschung in Mainz, Germany; Rebecca Wade at the European Molecular Biology Laboratory in Heidelberg, and Thomas Simonson at the Laboratoire de Biologie Structurale in Strasbourg, France.

Alexei Sokolov's advanced techniques and experiments in optical sensors and microscopy with molecular-scale resolution had a strong impact on my students' and my work over the last several years. I thank Alexei for providing a great opportunity for collaborative work with his group at the Department of Polymer Science, the University of Akron.

In the course of the last two decades, I have benefited enormously from my communication with Alain Bossavit (Électricité de France and Laboratoire de Genie Electrique de Paris), from his very deep knowledge of all aspects of computational electromagnetism, and from his very detailed and thoughtful analysis of any difficult subject that would come up.

Isaak Mayergoyz of the University of Maryland at College Park has on many occasions shared his valuable insights with me. His knowledge of many areas of electromagnetism, physics and mathematics is very profound and often unmatched.

My communication with Jon Webb (McGill University, Montréal) has always been thought-provoking and informative. His astute observations and comments make complicated matters look clear and simple. I was very pleased that Professor Webb devoted part of his sabbatical leave to our joint research on Flexible Local Approximation MEthods (FLAME, Chapter 4).

Yuri Kizimovich (Plassotech Corp., California) and I have worked jointly on a variety of projects over the last 25 years. His original thinking and elegant solutions of practical problems have always been a great asset. Yury's help and long-term collaboration are greatly appreciated.

Even though over 20 years have already passed since the untimely death of my thesis advisor, Yu.V. Rakitskii, his students still remember very warmly

his relentless strive for excellence and quixotic attitude to scientific research. Rakitskii's main contribution was to numerical methods for stiff systems of differential equations. He was guided by the idea of incorporating, to the extent possible, analytical approximations into numerical methods. This approach is manifest in FLAME that I believe Rakitskii would have liked.

My sincere thanks go to

- Dmitry Golovaty (The University of Akron), for his help on many occasions and for interesting discussions.
- Viacheslav Dombrovski, a scientist of incomparable erudition, for many pearls of wisdom.
- Elena Ivanova and Sergey Voskoboynikov (Technical University of St. Petersburg, Russia), for their very, very diligent work on FLAME.
- Benjamin Yellen (Duke University), for many discussions, innovative ideas, and for his great contribution to the NSF-NIRT project on magnetic assembly of particles.
- Mark Stockman (Georgia State University), for sharing his very deep and broad knowledge and expertise in many areas of plasmonics and nanophotonics.
- J. Douglas Lavers (the University of Toronto), for his help, cooperation and continuing support over many years.
- Fritz Keilmann (the Max-Planck-Institut für Biochemie in Martinsried, Germany), for providing an excellent opportunity for collaboration on problems in infrared microscopy.
- Boris Shoykhet (Rockwell Automation), an excellent engineer, mathematician and finite element analyst, for many valuable discussions.
- Nicolae-Alexandru Nicorovici (University of Technology, Sydney, Australia), for his deep and detailed comments on "cloaking," metamaterials, and properties of photonic structures.
- H. Neal Bertram (UCSD – the University of California, San Diego), for his support. I have always admired Neal's remarkable optimism and enthusiasm that make communication with him so stimulating.
- Adalbert Konrad (the University of Toronto) and Nathan Ida (the University of Akron) for their help and support.
- Pierre Asselin (Seagate, Pittsburgh) for very interesting insights, particularly in connection with *a priori* error estimates in finite element analysis.
- Sheldon Schultz (UCSD) and David Smith (UCSD and Duke) for familiarizing me with plasmonic effects a decade ago.

I appreciate the help, support and opportunities provided by the International Compumag Society through a series of the International Compumag Conferences and through personal communication with its Board and members: Jan K Sykulski, Arnulf Kost, Kay Hameyer, François Henrotte, Oszkár Bíró, J.-P. Bastos, R.C. Mesquita, and others.

A substantial portion of the book forms a basis of the graduate course "Simulation of Nanoscale Systems" that I developed and taught at the

University of Akron, Ohio. I thank my colleagues at the Department of Electrical & Computer Engineering and two Department Chairs, Alexis De Abreu Garcia and Nathan Ida, for their support and encouragement.

My Ph.D. students have contributed immensely to the research, and their work is frequently referred to throughout the book. Alexander Plaks worked on adaptive multigrid methods and generalized finite element methods for electromagnetic applications. Leonid Proekt was instrumental in the development of generalized FEM, especially for the vectorial case, and of absorbing boundary conditions. Jianhua Dai has worked on generalized finite-difference methods. Frantisek Čajko developed schemes with flexible local approximation and carried out, with a great deal of intelligence and ingenuity, a variety of simulations in nano-photonics and nano-optics.

I gratefully acknowledge financial support by the National Science Foundation and the NSF-NIRT program, Rockwell Automation, 3ga Corporation and Baker Hughes Corporation.

NEC Europe (Sankt Augustin, Germany) provided not only financial support but also an excellent opportunity to work with Achim Basermann, an expert in high performance computing, on parallel implementation of the Generalized FEM. I thank Guy Lonsdale, Achim Basermann and Fabienne Cortial-Goutaudier for hosting me at the NEC on several occasions.

A number of workshops and tutorials at the University of Minnesota in Minneapolis[1] have been exceptionally interesting and educational for me. I sincerely thank the organizers: Douglas Arnold, Debra Lewis, Cheri Shakiban, Boris Shklovskii, Alexander Grosberg and others.

I am very grateful to Serge Prudhomme, the reviewer of this book, for many insightful comments, numerous corrections and suggestions, and especially for his careful and meticulous analysis of the chapters on finite difference and finite element methods.[2] The reviewer did not wish to remain anonymous, which greatly facilitated our communication and helped to improve the text. Further comments, suggestions and critique from the readers is very welcome and can be communicated to me directly or through the publisher.

Finally, I thank Springer's editors for their help, cooperation and patience.

[1] *Electrostatic Interactions and Biophysics*, April–May 2004, Theoretical Physics Institute.

Future Challenges in Multiscale Modeling and Simulation, November 2004; *New Paradigms in Computation*, March 2005; *Effective Theories for Materials and Macromolecules*, June 2005; *New Directions Short Course: Quantum Computation*, August 2005; *Negative Index Materials*, October 2006; *Classical and Quantum Approaches in Molecular Modeling*, July 2007 – all at the Institute for Mathematics and Its Applications, http://www.ima.umn.edu/

[2] Serge Prudhomme is with the Institute for Computational Engineering and Sciences (ICES), formerly known as TICAM, at the University of Texas at Austin.

Contents

Preface VII

1 Introduction 1
1.1 Why Deal with the Nanoscale? 1
1.2 Why Special Models for the Nanoscale? 3
1.3 How To Hone the Computational Tools 6
1.4 So What? 8

2 Finite-Difference Schemes 11
2.1 Introduction 11
2.2 A Primer on Time-Stepping Schemes 12
2.3 Exact Schemes 16
2.4 Some Classic Schemes for Initial Value Problems 18
2.4.1 The Runge–Kutta Methods 20
2.4.2 The Adams Methods 24
2.4.3 Stability of Linear Multistep Schemes 24
2.4.4 Methods for Stiff Systems 27
2.5 Schemes for Hamiltonian Systems 34
2.5.1 Introduction to Hamiltonian Dynamics 34
2.5.2 Symplectic Schemes for Hamiltonian Systems 37
2.6 Schemes for One-Dimensional Boundary Value Problems 39
2.6.1 The Taylor Derivation 39
2.6.2 Using Constraints to Derive Difference Schemes 40
2.6.3 Flux-Balance Schemes 42
2.6.4 Implementation of 1D Schemes for Boundary Value Problems 46
2.7 Schemes for Two-Dimensional Boundary Value Problems 47
2.7.1 Schemes Based on the Taylor Expansion 47
2.7.2 Flux-Balance Schemes 48
2.7.3 Implementation of 2D Schemes 50
2.7.4 The Collatz “Mehrstellen” Schemes in 2D 51

2.8 Schemes for Three-Dimensional Problems 55
2.8.1 An Overview 55
2.8.2 Schemes Based on the Taylor Expansion in 3D 55
2.8.3 Flux-Balance Schemes in 3D 56
2.8.4 Implementation of 3D Schemes 57
2.8.5 The Collatz "Mehrstellen" Schemes in 3D 58
2.9 Consistency and Convergence of Difference Schemes 59
2.10 Summary and Further Reading 64

3 The Finite Element Method 69
3.1 Everything is Variational 69
3.2 The Weak Formulation and the Galerkin Method 75
3.3 Variational Methods and Minimization 81
3.3.1 The Galerkin Solution Minimizes the Error 81
3.3.2 The Galerkin Solution and the Energy Functional 82
3.4 Essential and Natural Boundary Conditions 83
3.5 Mathematical Notes: Convergence, Lax–Milgram and Céa's Theorems 86
3.6 Local Approximation in the Finite Element Method 89
3.7 The Finite Element Method in One Dimension 91
3.7.1 First-Order Elements 91
3.7.2 Higher-Order Elements 102
3.8 The Finite Element Method in Two Dimensions 105
3.8.1 First-Order Elements 105
3.8.2 Higher-Order Triangular Elements 120
3.9 The Finite Element Method in Three Dimensions 122
3.10 Approximation Accuracy in FEM 123
3.11 An Overview of System Solvers 129
3.12 Electromagnetic Problems and Edge Elements 139
3.12.1 Why Edge Elements? 139
3.12.2 The Definition and Properties of Whitney-Nédélec Elements 142
3.12.3 Implementation Issues 145
3.12.4 Historical Notes on Edge Elements 146
3.12.5 Appendix: Several Common Families of Tetrahedral Edge Elements 147
3.13 Adaptive Mesh Refinement and Multigrid Methods 148
3.13.1 Introduction 148
3.13.2 Hierarchical Bases and Local Refinement 149
3.13.3 *A Posteriori* Error Estimates 151
3.13.4 Multigrid Algorithms 154
3.14 Special Topic: Element Shape and Approximation Accuracy . . 158
3.14.1 Introduction 158
3.14.2 Algebraic Sources of Shape-Dependent Errors: Eigenvalue and Singular Value Conditions 160

3.14.3 Geometric Implications of the Singular Value Condition 171
3.14.4 Condition Number and Approximation 179
3.14.5 Discussion of Algebraic and Geometric *a priori* Estimates 180
3.15 Special Topic: Generalized FEM 181
3.15.1 Description of the Method 181
3.15.2 Trade-offs 183
3.16 Summary and Further Reading 184
3.17 Appendix: Generalized Curl and Divergence 186

4 Flexible Local Approximation MEthods (FLAME) 189
4.1 A Preview 189
4.2 Perspectives on Generalized FD Schemes 191
4.2.1 Perspective #1: Basis Functions Not Limited to Polynomials 191
4.2.2 Perspective #2: Approximating the *Solution*, Not the Equation 192
4.2.3 Perspective #3: Multivalued Approximation 193
4.2.4 Perspective #4: Conformity vs. Flexibility 193
4.2.5 Why Flexible Approximation? 195
4.2.6 A Preliminary Example: the 1D Laplace Equation 197
4.3 Trefftz Schemes with Flexible Local Approximation 198
4.3.1 Overlapping Patches 198
4.3.2 Construction of the Schemes 200
4.3.3 The Treatment of Boundary Conditions 202
4.3.4 Trefftz–FLAME Schemes for Inhomogeneous and Nonlinear Equations 203
4.3.5 Consistency and Convergence of the Schemes 205
4.4 Trefftz–FLAME Schemes: Case Studies 206
4.4.1 1D Laplace, Helmholtz and Convection-Diffusion Equations 206
4.4.2 The 1D Heat Equation with Variable Material Parameter 207
4.4.3 The 2D and 3D Laplace Equation 208
4.4.4 The Fourth Order 9-point Mehrstellen Scheme for the Laplace Equation in 2D 209
4.4.5 The Fourth Order 19-point Mehrstellen Scheme for the Laplace Equation in 3D 210
4.4.6 The 1D Schrödinger Equation. FLAME Schemes by Variation of Parameters 210
4.4.7 Super-high-order FLAME Schemes for the 1D Schrödinger Equation 212
4.4.8 A Singular Equation 213
4.4.9 A Polarized Elliptic Particle 215
4.4.10 A Line Charge Near a Slanted Boundary 216
4.4.11 Scattering from a Dielectric Cylinder 217

4.5 Existing Methods Featuring Flexible or Nonstandard Approximation . . . 219
4.5.1 The Treatment of Singularities in Standard FEM . . . 221
4.5.2 Generalized FEM by Partition of Unity . . . 221
4.5.3 Homogenization Schemes Based on Variational Principles . . . 222
4.5.4 Discontinuous Galerkin Methods . . . 222
4.5.5 Homogenization Schemes in FDTD . . . 223
4.5.6 Meshless Methods . . . 224
4.5.7 Special Finite Element Methods . . . 225
4.5.8 Domain Decomposition . . . 226
4.5.9 Pseudospectral Methods . . . 226
4.5.10 Special FD Schemes . . . 227
4.6 Discussion . . . 228
4.7 Appendix: Variational FLAME . . . 231
4.7.1 References . . . 231
4.7.2 The Model Problem . . . 232
4.7.3 Construction of Variational FLAME . . . 232
4.7.4 Summary of the Variational-Difference Setup . . . 235
4.8 Appendix: Coefficients of the 9-Point Trefftz–FLAME Scheme for the Wave Equation in Free Space . . . 236
4.9 Appendix: the Fréchet Derivative . . . 237

5 Long-Range Interactions in Free Space . . . 239
5.1 Long-Range Particle Interactions in a Homogeneous Medium . . . 239
5.2 Real and Reciprocal Lattices . . . 242
5.3 Introduction to Ewald Summation . . . 243
5.3.1 A Boundary Value Problem for Charge Interactions . . . 246
5.3.2 A Re-formulation with “Clouds” of Charge . . . 248
5.3.3 The Potential of a Gaussian Cloud of Charge . . . 249
5.3.4 The Field of a Periodic System of Clouds . . . 251
5.3.5 The Ewald Formulas . . . 252
5.3.6 The Role of Parameters . . . 254
5.4 Grid-based Ewald Methods with FFT . . . 256
5.4.1 The Computational Work . . . 256
5.4.2 On Numerical Differentiation . . . 262
5.4.3 Particle–Mesh Ewald . . . 264
5.4.4 Smooth Particle–Mesh Ewald Methods . . . 267
5.4.5 Particle–Particle Particle–Mesh Ewald Methods . . . 269
5.4.6 The York–Yang Method . . . 271
5.4.7 Methods Without Fourier Transforms . . . 272
5.5 Summary and Further Reading . . . 274
5.6 Appendix: The Fourier Transform of “Periodized” Functions . . . 277
5.7 Appendix: An Infinite Sum of Complex Exponentials . . . 278

6 Long-Range Interactions in Heterogeneous Systems 281
6.1 Introduction 281
6.2 FLAME Schemes for Static Fields of Polarized Particles in 2D 285
6.2.1 Computation of Fields and Forces for Cylindrical Particles 289
6.2.2 A Numerical Example: Well-Separated Particles 291
6.2.3 A Numerical Example: Small Separations 294
6.3 Static Fields of Spherical Particles in a Homogeneous Dielectric 303
6.3.1 FLAME Basis and the Scheme 303
6.3.2 A Basic Example: Spherical Particle in Uniform Field . . 306
6.4 Introduction to the Poisson–Boltzmann Model 309
6.5 Limitations of the PBE Model 313
6.6 Numerical Methods for 3D Electrostatic Fields of Colloidal Particles 314
6.7 3D FLAME Schemes for Particles in Solvent 315
6.8 The Numerical Treatment of Nonlinearity 319
6.9 The DLVO Expression for Electrostatic Energy and Forces 321
6.10 Notes on Other Types of Force 324
6.11 Thermodynamic Potential, Free Energy and Forces 328
6.12 Comparison of FLAME and DLVO Results 332
6.13 Summary and Further Reading 337
6.14 Appendix: Thermodynamic Potential for Electrostatics in Solvents 338
6.15 Appendix: Generalized Functions (Distributions) 343

7 Applications in Nano-Photonics 349
7.1 Introduction 349
7.2 Maxwell's Equations 349
7.3 One-Dimensional Problems of Wave Propagation 353
7.3.1 The Wave Equation and Plane Waves 353
7.3.2 Signal Velocity and Group Velocity 355
7.3.3 Group Velocity and Energy Velocity 358
7.4 Analysis of Periodic Structures in 1D 360
7.5 Band Structure by Fourier Analysis (Plane Wave Expansion) in 1D 375
7.6 Characteristics of Bloch Waves 379
7.6.1 Fourier Harmonics of Bloch Waves 379
7.6.2 Fourier Harmonics and the Poynting Vector 380
7.6.3 Bloch Waves and Group Velocity 380
7.6.4 Energy Velocity for Bloch Waves 382
7.7 Two-Dimensional Problems of Wave Propagation 384
7.8 Photonic Bandgap in Two Dimensions 386
7.9 Band Structure Computation: PWE, FEM and FLAME 389
7.9.1 Solution by Plane Wave Expansion 389
7.9.2 The Role of Polarization 390

7.9.3 Accuracy of the Fourier Expansion 391
7.9.4 FEM for Photonic Bandgap Problems in 2D 393
7.9.5 A Numerical Example: Band Structure Using FEM 397
7.9.6 Flexible Local Approximation Schemes for Waves in Photonic Crystals 401
7.9.7 Band Structure Computation Using FLAME 405
7.10 Photonic Bandgap Calculation in Three Dimensions: Comparison with the 2D Case 411
7.10.1 Formulation of the Vector Problem 411
7.10.2 FEM for Photonic Bandgap Problems in 3D 415
7.10.3 Historical Notes on the Photonic Bandgap Problem 416
7.11 Negative Permittivity and Plasmonic Effects 417
7.11.1 Electrostatic Resonances for Spherical Particles 419
7.11.2 Plasmon Resonances: Electrostatic Approximation 421
7.11.3 Wave Analysis of Plasmonic Systems 423
7.11.4 Some Common Methods for Plasmon Simulation 423
7.11.5 Trefftz–FLAME Simulation of Plasmonic Particles 426
7.11.6 Finite Element Simulation of Plasmonic Particles 429
7.12 Plasmonic Enhancement in Scanning Near-Field Optical Microscopy 433
7.12.1 Breaking the Diffraction Limit 434
7.12.2 Apertureless and Dark-Field Microscopy 439
7.12.3 Simulation Examples for Apertureless SNOM 441
7.13 Backward Waves, Negative Refraction and Superlensing 446
7.13.1 Introduction and Historical Notes 446
7.13.2 Negative Permittivity and the "Perfect Lens" Problem . 451
7.13.3 Forward and Backward Plane Waves in a Homogeneous Isotropic Medium 456
7.13.4 Backward Waves in Mandelshtam's Chain of Oscillators 459
7.13.5 Backward Waves and Negative Refraction in Photonic Crystals 465
7.13.6 Are There Two Species of Negative Refraction? 471
7.14 Appendix: The Bloch Transform 477
7.15 Appendix: Eigenvalue Solvers 478

8 Conclusion: "Plenty of Room at the Bottom" for Computational Methods 487

References 489

Index 523

1

Introduction

Some years ago, a colleague of mine explained to me that a good presentation should address three key questions: 1) Why? (i.e. Why do it?) 2) How? (i.e. How do we do it?) and 3) So What?

The following sections answer these questions, and a few more.

1.1 Why Deal with the Nanoscale?

May you live in interesting times.

Eric Frank Russell, "U-Turn" (1950).

The complexity and variety of applications on the nanoscale are as great, or arguably greater, than on the macroscale. While a detailed account of nanoscale problems in a single book is impossible, one can make a general observation on the importance of the nanoscale: the properties of materials are strongly affected by their nanoscale structure. Over the last two decades, mankind has been gradually inventing and acquiring means to characterize and manipulate that structure. Many remarkable effects, physical phenomena, materials and devices have already been discovered or developed: nanocomposites, carbon nanotubes, nanowires and nanodots, nanoparticles of different types, photonic crystals, and so on.

On a more fundamental level, research in nanoscale physics may provide clues to the most profound mysteries of nature.

> "Where is the frontier of physics?", asks L.S. Schulman in the Preface to his book [Sch97]. "Some would say 10^{-33} cm, some 10^{-15} cm and some 10^{+28} cm. My vote is for 10^{-6} cm. Two of the greatest puzzles of our age have their origins at the interface between the macroscopic and microscopic worlds. The older mystery is the thermodynamic arrow of

> time, the way that (mostly) time-symmetric microscopic laws acquire a manifest asymmetry at larger scales. And then there's the superposition principle of quantum mechanics, a profound revolution of the twentieth century. When this principle is extrapolated to macroscopic scales, its predictions seem widely at odds with ordinary experience."

The second "puzzle" that Professor Schulman refers to is the apparent contradiction between the quantum-mechanical representation of micro-objects in a superposition of quantum states and a single unambiguous state that all of us really observe for macro-objects. Where and how exactly is this transition from the quantum world to the macro-world effected? The boundary between particle- or atomic-size quantum objects and macro-objects is on the nanoscale; that is where the "collapse of the quantum-mechanical wavefunction" from a superposition of states to one well-defined state would have to occur. Recent remarkable double-slit experiments by M. Arndt's Quantum Nanophysics group at the University of Vienna show no evidence of "collapse" of the wavefunction and prove the wave nature of large molecules with the mass of up to 1,632 units and size up to 2 nm (tetraphenylporphyrin $C_{44}H_{30}N_4$ and the fluorinated buckyball $C_{60}F_{48}$).[1] If further experiments with nanoscale objects are carried out, they will most likely confirm that the "collapse" of the wavefunction is not a fundamental physical law but only a metaphorical tool for describing the transition to the macroworld; still, such experiments will undoubtedly be captivating.

Getting back to more practical aspects of nanoscale research, I illustrate its promise with one example from Chapter 7 of this book. It is well known that visible light is electromagnetic waves with the wavelengths from approximately 400 nm (violet light) to ~700 nm (red light); green light is in the middle of this range. Thus there are approximately 2,000 wavelengths of green light per millimeter (or about 50,000 per inch). Propagation of light through a material is governed not only by the atomic-level properties but also, in many interesting and important ways, by the nanoscale/subwavelength structure of the material (i.e. the scale from 5–10 nm to a few hundred nanometers).

Consider ocean waves as an analogy. A wave will easily pass around a relatively small object, such as a buoy. However, if the wave hits a long line of buoys, interesting things will start to happen: an interference pattern may emerge behind the line. Furthermore, if the buoys are arranged in a two-dimensional array, possible wave patterns are richer still.

Substituting an electromagnetic wave of light (say, with wavelength $\lambda = 500$ nm) for the ocean wave and a lattice of dielectric cylindrical rods (say, 200 nm in diameter) for the two-dimensional array of buoys, we get what is known as a *photonic crystal*.[2] It is clear that the subwavelength structure

[1] M. Arndt *et al.*, Wave-particle duality of C60 molecules, *Nature* 401, 1999, pp. 680–682; http://physicsweb.org/articles/world/18/3/5.

[2] The analogy with electromagnetic waves would be closer mathematically but less intuitive if acoustic waves in the ocean were considered instead of surface waves.

of the crystal may bring about very interesting and unusual behavior of the wave.

Even more fascinating is the possibility of *controlling* the propagation of light in the material by a clever design of the subwavelength structure. "Cloaking" – making objects invisible by wrapping them in a carefully designed metamaterial – has become an area of serious research (J.B. Pendry *et al.* [PSS06]) and has already been demonstrated experimentally in the microwave region (D. Schurig *et al.* [SMJ+06]). Guided by such material, the rays of light would bend and pass around the object as if it were not there (G. Gbur [Gbu03], J.B. Pendry *et al.* [PSS06], U. Leonhardt [Leo06]). A note to the reader who wishes to hide behind this cloak: if you are invisible to the outside world, the outside world is invisible to you. This follows from the reciprocity principle in electromagnetism.[3]

Countless other equally fascinating nanoscale applications in numerous other areas could be given. Like it or not, we live in interesting times.

1.2 Why Special Models for the Nanoscale?

> A good model can advance fashion by ten years.
>
> Yves Saint Laurent

First, a general observation. A *simulation model* consists of a physical and mathematical formulation of the problem at hand and a computational method. The formulation tells us *what* to solve and the computational method tells us *how* to solve it. Frequently more than one formulation is possible, and almost always several computational techniques are available; hence there potentially are numerous combinations of formulations and methods. Ideally, one strives to find the best such combination(s) in terms of efficiency, accuracy, robustness, algorithmic simplicity, and so on.

It is not surprising that the *formulations* of nanoscale problems are indeed special. The scale is often too small for continuous-level macroscopic laws to be fully applicable; yet it is too large for a first-principles atomic simulation to be feasible. Computational compromises are reached in several different ways. In some cases, continuous parameters can be used with some caution and with suitable adjustments. One example is light scattering by small particles and the related "plasmonic" effects (Chapter 7), where the dielectric constant of metals or dielectrics can be adjusted to account for the size of the scatterers. In other situations, *multiscale* modeling is used, where a hierarchy of problems

[3] Perfect invisibility is impossible even theoretically, however. With some imperfection, the effect can theoretically be achieved only in a narrow range of wavelengths. The reason is that the special metamaterials must have dispersion – i.e. their electromagnetic properties must be frequency-dependent.

are solved and the information obtained on a finer level is passed on to the coarser ones and back. Multiscale often goes hand-in-hand with *multiphysics*: for example, molecular dynamics on the finest scale is combined with continuum mechanics on the macroscale. The Society for Industrial and Applied Mathematics (SIAM) now publishes a journal devoted entirely to this subject: *Multiscale Modeling and Simulation*, inaugurated in 2003.

The applications and problems in this book have some multiscale features but can still be dealt with on a single scale[4] – primarily the nanoscale. As an example: in colloidal simulation (Chapter 6) the molecular-scale degrees of freedom corresponding to microions in the solvent are "integrated out," the result being the Poisson–Boltzmann equation that applies on the scale of colloidal particles (approximately from 10 to 1000 nm). Still, simulation of optical tips (Section 7.12, p. 433) does have salient multiscale features.

Let us now discuss the computational side of nanoscale models. Computational analysis is a mature discipline combining science, engineering and elements of art. It includes general and powerful techniques such as finite difference, finite element, spectral or pseudospectral, integral equation and other methods; it has been applied to every physical problem and device imaginable.

Are these existing methods good enough for nanoscale problems? The answer can be anything from "yes" to "maybe" to "no," depending on the problem.

- When continuum models are still applicable, traditional methods work well. A relevant example is the simulation of light scattering by plasmon nanoparticles and of plasmon-enhanced components for ultra-sensitive optical sensors and near-field microscopes (Chapter 7). Despite the nanoscale features of the problem, equivalent material parameters (dielectric permittivity and magnetic permeability) can still be used, possibly with some adjustments. Consequently, commercial finite-element software is suitable for this type of modeling.
- When the system size is even smaller, as in macromolecular simulation, the use of equivalent material parameters is more questionable. In electrostatic models of protein molecules in solvents – an area of extensive and intensive research due to its enormous implications for biology and medicine – two main approaches coexist. In *implicit* models, the solvent is characterized by equivalent continuum parameters (dielectric permittivity and the Debye length). In the layer of the solvent immediately adjacent to the surface of the molecule, these equivalent parameters are dramatically different from their values in the bulk (A. Rubinstein & S. Sherman [RS04]). In contrast, *explicit* models directly include molecular dynamics of the solvent. This approach is in principle more accurate, as no approximation of the solvent by an equivalent medium is made, but the computational cost is extremely

[4] The Flexible Local Approximation MEthod (FLAME) of Chapter 4 can, however, be viewed as a two-scale method: the difference scheme is formed on a relatively coarse grid but incorporates information about the solution on a finer scale.

high due to a very large number of degrees of freedom corresponding to the molecules of the solvent. For more information on protein simulation, see T. Schlick's book [Sch02] and T. Simonson's review paper [Sim03] as a starting point.

- When the problem reduces to a system of ordinary differential equations, the computational analysis is on very solid ground – this is one of the most mature areas of numerical mathematics (Chapter 2). It is highly desirable to use numerical schemes that preserve the essential physical properties of the system. In Molecular Dynamics, such fundamental properties are the conservation of energy and momentum, and – more generally – *symplecticness* of the underlying Hamiltonian system (Section 2.5). Time-stepping schemes with analogous conservation properties are available and their advantages are now widely recognized (J.M. Sanz-Serna & M.P. Calvo [SSC94], Yu.B. Suris [Sur87, Sur96], R.D. Skeel *et al.* [RDS97]).
- Quantum mechanical effects require special computational treatment. The models are substantially different from those of continuum media for which the traditional methods (such as finite elements or finite differences) were originally designed and used. Nevertheless these traditional methods can be very effective at certain stages of quantum mechanical analysis. For example, classical finite-difference schemes (in particular, the Collatz "Mehrstellen" schemes, Chapter 2), have been successfully applied to the Kohn–Sham equation – the central procedure in Density Functional Theory. (This is the Schrödinger equation, with the potential expressed as a function of electron density.) For a detailed description, see E.L. Briggs *et al.* [BSB96] and T.L. Beck [Bec00]. Moreover, difference schemes can also be used to find the electrostatic potential from the Poisson equation with the electron density in the right hand side.
- Colloidal simulation considered in Chapter 6 is an interesting and special computational case. As explained in that chapter, classical methods of computation are not particularly well suited for this problem. Finite element meshes become too complex and impractical to generate even for a moderate number of particles in the model; standard finite-difference schemes require unreasonably fine grids to represent the boundaries of the particles accurately; the Fast Multipole Method does not work too well for inhomogeneous and/or nonlinear problems. A new finite-difference calculus of Flexible Local Approximation MEthods (FLAME) is a promising alternative (Chapter 4).

This list could easily be extended to include other examples, but the main point is clear: a vast assortment of computational methods, both traditional and new, are very helpful for the efficient simulation of nanoscale systems.

1.3 How To Hone the Computational Tools

> A computer makes as many mistakes in two seconds as 20 men working 20 years make.
>
> Murphy's Laws of Computing

Computer simulation is not an exact science. If it were, one would simply set a desired level of accuracy ϵ of the numerical solution and prove that a certain method achieves that level with the minimal number of operations $\Theta = \Theta(\epsilon)$. The reality is of course much more intricate. First, there are many possible measures of accuracy and many possible measures of the cost (keeping in mind that human time needed for the development of algorithms and software may be more valuable than the CPU time). Accuracy and cost both depend on the class and subclass of problems being solved. For example, numerical solution becomes substantially more complicated if discontinuities and edge or corner singularities of the field need to be represented accurately.

Second, it is usually close to impossible to guarantee, at the mathematical level of rigor, that the numerical solution obtained has a certain prescribed accuracy.[5] Third, in practice it is never possible to prove that any given method minimizes the number of arithmetic operations.

Fourth, there are modeling errors – approximations made in the formulation of the physical problem; these errors are a particular concern on the nanoscale, where direct and accurate experimental verification of the assumptions made is very difficult. Fifth, a host of other issues – from the algorithmic implementation of the chosen method to roundoff errors – are quite difficult to take into account. Parallelization of the algorithm and the computer code is another complicated matter.

With all this in mind, computer simulation turns out to be partially an art. There is always more than one way to solve a given problem numerically and, with enough time and resources, any reasonable approach is likely to produce a result eventually.

Still, it is obvious that not all approaches are equal. Although the accuracy and computational cost cannot be determined exactly, some qualitative measures are certainly available and are commonly used. The main characteristic is the *asymptotic* behavior of the number of operations and memory required for a given method as a function of some accuracy-related parameter. In mesh-based methods (finite elements, finite differences, Ewald summation,

[5] There is a notable exception in variational methods: rigorous *pointwise* error bounds can, for some classes of problems, be established using dual formulations (see p. 153 for more information). However, this requires numerical solution of a separate auxiliary problem for Green's function at each point where the error bound is sought.

etc.) the mesh size h or the number of nodes n usually act as such a parameter. The "big-oh" notation is standard; for example, the number of arithmetic operations θ being $\mathcal{O}(n^\gamma)$ as $n \to \infty$ means that $c_1 n^\gamma \leq \theta \leq c_2 n^\gamma$, where $c_{1,2}$ and γ are some positive constants independent of n. Computational methods with the operation count and memory $\mathcal{O}(n)$ are considered as asymptotically optimal; the doubling of the number of nodes (or some other such parameter) leads, roughly, to the doubling of the number of operations and memory size. For several classes of problems, there exist divide-and-conquer or hierarchical strategies with either optimal $\mathcal{O}(n)$ or slightly suboptimal $\mathcal{O}(n \log n)$ complexity. The most notable examples are Fast Fourier Transforms (FFT), Fast Multipole Methods, multigrid methods, and FFT-based Ewald summation.

Clearly, the numerical factors $c_{1,2}$ also affect the performance of the method. For real-life problems, they can be determined experimentally and their magnitude is not usually a serious concern. A notable exception is the Fast Multipole Method for multiparticle interactions; its operation count is close to optimal, $\mathcal{O}(n_\mathrm{p} \log n_\mathrm{p})$, where n_p is the number of particles, but the numerical prefactors are very large, so the method outperforms the brute-force approach ($\mathcal{O}(n_\mathrm{p}^2)$ pairwise particle interactions) only for a large number of particles, tens of thousands and beyond.

Given that the choice of a suitable method is partially an art, what is one to do? As a practical matter, the availability of good public domain and commercial software in many cases simplifies the decision. Examples of such software are

- Molecular Dynamics packages AMBER (Assisted Model Building with Energy Refinement, amber.scripps.edu); CHARMM/CHARMm (Chemistry at HARvard Macromolecular Mechanics, yuri.harvard.edu, accelrys.com/products/dstudio/index.html), NAMD (www.ks.uiuc.edu/Research/namd), GROMACS (gromacs.org), TINKER (dasher.wustl.edu/tinker), DL POLY (www.cse.scitech.ac.uk/ccg/software/DL_POLY/index.shtml).
- A finite difference Poisson-Boltzman solver DelPhi (honiglab.cpmc.columbia.edu).
- Finite Element software developed by ANSYS (ansys.com – comprehensive FE modeling, with multiphysics); by ANSOFT (ansoft.com – state-of-the-art FE package for electromagnetic design); by Comsol (comsol.com or femlab.com – the Comsol Multiphysics™package, also known as FEMLAB); and others.
- A software suite from Rsoft Group (rsoftdesign.com) for design of photonics components and optical networks.
- Electromagnetic time-domain simulation software from CST (Computer Simulation Technology, cst.com).

This list is certainly not exhaustive and, among other things, does not include software for *ab initio* electronic structure calculation, as this subject matter lies beyond the scope of the book.

The obvious drawback of using somebody else's software is that the user cannot extend its capabilities and apply it to problems for which it was not designed. Some tricks are occasionally possible (for example, equations in cylindrical coordinates can be converted to the Cartesian system by a mathematically equivalent transformation of material parameters), but by and large the user is out of luck if the code is proprietary and does not handle a given problem. For open-source software, users may in principle add their own modules to accomplish a required task, but, unless the revisions are superficial, this requires detailed knowledge of the code.

Whether the reader of this book is an intelligent user of existing software or a developer of his own algorithms and codes, the book will hopefully help him/her to understand how the underlying numerical methods work.

1.4 So What?

Avoid clichés like the plague!

William Safire's Rules for Writers

Multisyllabic clichés are probably the worst type, but I feel compelled to use one: nanoscale science and technology are *interdisciplinary*. The book is intended to be a bridge between two broad fields: computational methods, both traditional and new, on the one hand, and several nanoscale or molecular-scale applications on the other. It is my hope that the reader who has a background in physics, physical chemistry, electrical engineering or related subjects, and who is curious about the inner workings of computational methods, will find this book helpful for crossing the bridge between the disciplines. Likewise, experts in computational methods may be interested in browsing the application-related chapters.

At the same time, readers who wish to stay on their side of the "bridge" may also find some topics in the book to be of interest. An example of such a topic for numerical analysts is the FLAME schemes of Chapter 4; a novel feature of this approach is the systematic use of local approximation spaces *in the FD context*, with basis functions not limited to Taylor polynomials. Similarly, in the chapter on Finite Element analysis (Chapter 3), the theory of shape-related approximation errors is nonstandard and yields some interesting error estimates.

Since the prospective reader will not necessarily be an expert in any given subject of the book, I have tried, to the extent possible, to make the text accessible to researchers, graduate and even senior-level undergraduate students with a good general background in physics and mathematics. While part of the material is related to mathematical physics, the style of the book can be

characterized as *physical mathematics*[6] – "physical" explanation of the underlying mathematical concepts. I hope that this style will be tolerable to the mathematicians and beneficial to the reader with a background in physical sciences and engineering.

Sometimes, however, a more technical presentation is necessary. This is the case in the analysis of consistency errors and convergence of difference schemes in Chapter 2, Ewald summation in Chapter 5, and the derivation of FLAME basis functions for particle problems in Chapter 6. In many other instances, references to a rigorous mathematical treatment of the subject are provided.

I cannot stress enough that this book is very far from being a comprehensive treatise on nanoscale problems and applications. The selection of subjects is strongly influenced by my research interests and experience. Topics where I felt I could contribute some new ideas, methods and results were favored. Subjects that are covered nicely and thoroughly in the existing literature were not included. For example, material on Molecular Dynamics was, for the most part, left out because of the abundance of good literature on this subject.[7] However, one of the most challenging parts of Molecular Dynamics – the computation of long-range forces in a homogeneous medium – appears as a separate chapter in the book (Chapter 5). The novel features of this analysis are a rigorous treatment of "charge allocation" to grid and the application of finite-difference schemes, with the potential splitting, in real space.

Chapter 2 gives the necessary background on Finite Difference (FD) schemes; familiarity with numerical methods is helpful but not required for reading and understanding this chapter. In addition to the standard material on classical methods, their consistency and convergence, this chapter includes introduction to flexible approximation schemes, Collatz "Mehrstellen" schemes, and schemes for Hamiltonian systems.

Chapter 3 is a concise self-contained description of the Finite Element Method (FEM). No special prior knowledge of computational methods is required to read most of this chapter. Variational principles and their role are explained first, followed by a tutorial-style exposition of FEM in the simplest 1D case. Two- and three-dimensional scalar problems are considered in the subsequent sections of the chapter. A more advanced subject is edge elements that are crucial for vector field problems in electromagnetic analysis. Readers already familiar with FEM may be interested in the new treatment of approximation accuracy as a function of element shape; this is a special topic in Chapter 3.

---

[6] Not exactly the same as "engineering mathematics," a more utilitarian, user-oriented approach.

[7] J.M. Haile, *Molecular Dynamics Simulation: Elementary Methods*, Wiley-Interscience, 1997; D. Frenkel & B. Smit, *Understanding Molecular Simulation*, Academic Press, 2001; D.C. Rapaport, *The Art of Molecular Dynamics Simulation*, Cambridge University Press, 2004; T. Schlik [Sch02], and others.

Chapter 4 introduces the Finite Difference (FD) calculus of Flexible Local Approximation MEthods (FLAME). Local analytical solutions are incorporated into the schemes, which often leads to much higher accuracy than would be possible in classical FD. A large assortment of examples illustrating the usage of the method are presented.

Chapter 6 can be viewed as an extension of Chapter 5 to multiparticle problems in *heterogeneous* media. The simulation of such systems, due to its complexity, has received relatively little attention, and good methods are still lacking. Yet the applications are very broad – from colloidal suspensions to polymers and polyelectrolytes; in all of these cases, the media are inhomogeneous because the dielectric permittivities of the solute and solvent are usually quite different. Ewald methods can only be used if the solvent is modeled explicitly, by including polarization on the molecular level; this requires a very large number of degrees of freedom in the simulation. An alternative is to model the solvent implicitly by continuum parameters and use the FLAME schemes of Chapter 4. Application of these schemes to the computation of the electrostatic potential, field and forces in colloidal systems is described in Chapter 6.

Chapter 7 deals with applications in nano-photonics and nano-optics. It reviews the mathematical theory of Bloch modes, in connection with the propagation of electromagnetic waves in periodic structures; describes plane wave expansion, FEM and FLAME for photonic bandgap computation; provides a theoretical background for plasmon resonances and considers various numerical methods for plasmon-enhanced systems. Such systems include optical sensors with very high sensitivity, as well as scanning near-field optical microscopes with molecular-scale resolution, unprecedented in optics. Chapter 7 also touches upon negative refraction and nanolensing – areas of very intensive research and debate – and includes new material on the inhomogeneity of backward wave media.

2

Finite-Difference Schemes

2.1 Introduction

Due to its relative simplicity, Finite Difference (FD) analysis was historically the first numerical technique for boundary value problems in mathematical physics. The excellent review paper by V. Thomée [Tho01] traces the origin of FD to a 1928 paper by R. Courant, K. Friedrichs and H. Lewy, and to a 1930 paper by S. Gerschgorin. However, the Finite Element Method (FEM) that emerged in the 1960s proved to be substantially more powerful and flexible than FD. The modern techniques of hp-adaption, parallel multilevel preconditioning, domain decomposition have made FEM ever more powerful (Chapter 3). Nevertheless, FD remains a very valuable tool, especially for problems with relatively simple geometry.

This chapter starts with a gentle introduction to FD schemes and proceeds to a more detailed review. Sections 2.2–2.4 are addressed to readers with little or no background in finite-difference methods. Section 2.3, however, introduces a nontraditional perspective and may be of interest to more advanced readers as well. By approximating the *solution* of the problem rather than a generic smooth function, one can achieve much higher accuracy. This nontraditional perspective will be further developed in Chapter 4.

Section 2.4 gives an overview of classical FD schemes for Ordinary Differential Equations (ODE) and systems of ODE; Section 2.5 – an overview of Hamiltonian systems that are particularly important in molecular dynamics.

Sections 2.6–2.8 describe FD schemes for *boundary value problems* in one, two and three dimensions. Some ideas of this analysis, such as minimization of the consistency error for a constrained set of functions, are nonstandard.

Finally, Section 2.9 summarizes the most important results on consistency and convergence of FD schemes.

In addition to providing a general background on FD methods, this chapter is intended to set the stage for the generalized FD analysis with "Flexible Local Approximation" described in Chapter 4. The scope of the present chapter is limited, and for a more comprehensive treatment and analysis of

FD methods – in particular, elaborate time-stepping schemes for ordinary differential equations, schemes for gas and fluid dynamics, Finite-Difference Time-Domain (FDTD) methods in electromagnetics, etc. – I defer to many excellent more specialized monographs. Highly recommended are books by C.W. Gear [Gea71] (ODE, including stiff systems), U.M. Ascher & L.R. Petzold [AP98], K.E. Brenan *et al.* [KB96] (ODE, especially the treatment of differential-algebraic equations), S.K. Godunov & V.S. Ryabenkii [GR87a] (general theory of difference schemes and hyperbolic equations), J. Butcher [But87, But03] (time-stepping schemes and especially Runge–Kutta methods), T.J. Chung [Chu02] and S.V. Patankar [Pat80] (schemes for computational fluid dynamics), A. Taflove & S.C. Hagness [TH05] (FDTD).

2.2 A Primer on Time-Stepping Schemes

[1] The following example is the simplest possible illustration of key principles of finite-difference analysis. Suppose we wish to solve the ordinary differential equation

$$\frac{du}{dt} = \lambda u \quad \text{on } [0, t_{\max}], \quad u(0) = u_0, \quad \operatorname{Re}\lambda < 0 \tag{2.1}$$

numerically. The exact solution of this equation

$$u_{\text{exact}} = u_0 \exp(\lambda t) \tag{2.2}$$

obviously has infinitely many values at infinitely many points within the interval. In contrast, numerical algorithms have to operate with *finite* (discrete) sets of data. We therefore introduce a set of points (grid) $t_0 = 0, t_1, \ldots, t_{n-1}, t_n = t_{\max}$ over the given interval. For simplicity, let us assume that the grid size Δt is the same for all pairs of neighboring points: $t_{k+1} - t_k = \Delta t$, so that $t_k = k\Delta t$.

We now consider equation (2.1) at a moment of time $t = t_k$:

$$\frac{du}{dt}(t_k) = \lambda u(t_k) \tag{2.3}$$

The first derivative du/dx can be approximated on the grid in several different ways:

$$\begin{aligned}
\frac{du}{dt}(t_k) &= \frac{u(t_{k+1}) - u(t_k)}{\Delta t} + \mathcal{O}(\Delta t) \\
\frac{du}{dt}(t_k) &= \frac{u(t_k) - u(t_{k-1})}{\Delta t} + \mathcal{O}(\Delta t) \\
\frac{du}{dt}(t_k) &= \frac{u(t_{k+1}) - u(t_{k-1})}{2\Delta t} + \mathcal{O}((\Delta t)^2)
\end{aligned}$$

[1] I am grateful to Serge Prudhomme for very helpful suggestions and comments on the material of this section.

These equalities – each of which can be easily justified by Taylor expansion – lead to the algorithms known as forward Euler, backward Euler and central difference schemes, respectively:

$$\frac{u_{k+1} - u_k}{\lambda \Delta t} \; - \; u_k \; = \; 0 \tag{2.4}$$

or, equivalently,

$$u_{k+1} \; - \; (1 + \lambda \Delta t) u_k \; = \; 0 \quad \text{(forward Euler)} \tag{2.5}$$

$$\frac{u_k - u_{k-1}}{\lambda \Delta t} \; = \; u_k \tag{2.6}$$

or

$$(1 - \lambda \Delta t) u_k \; - \; u_{k-1} \; = \; 0 \quad \text{(backward Euler)} \tag{2.7}$$

$$\frac{u_{k+1} - u_{k-1}}{2\lambda \Delta t} \; = \; u_k \tag{2.8}$$

or

$$u_{k+1} \; - \; 2\lambda \Delta t u_k \; - \; u_{k-1} \; = \; 0 \quad \text{(central difference)} \tag{2.9}$$

where u_{k-1}, u_k and u_{k+1} are approximations to $u(t)$ at discrete times t_{k-1}, t_k and t_{k+1}, respectively. For convenience of analysis, the schemes above are written in the form that makes the dimensionless product $\lambda \Delta t$ explicit.

The (discrete) solution for the *forward Euler* scheme (2.4) can be easily found by *time-stepping*: start with the given initial value $u(0) = u_0$ and use the scheme to find the value of the solution at each subsequent step:

$$u_{k+1} \; = \; (1 + \lambda \Delta t) \, u_k \tag{2.10}$$

This difference scheme was obtained by approximating the original differential equation, and it is therefore natural to expect that the solution of the original equation will *approximately* satisfy the difference equation. This can be easily verified because in this simple example the exact solution is known. Let us substitute the exact solution (2.2) into the left hand side of the difference equation (2.4):

$$\epsilon_c \; = \; u_0 \left[\frac{\exp\left(\lambda(k+1)\Delta t\right) - \exp\left(k\lambda \Delta t\right)}{\lambda \Delta t} \; - \; \exp\left(k\lambda \Delta t\right) \right]$$

$$= \; u_0 \exp(k\lambda \Delta t) \left[\frac{\exp(\lambda \Delta t) - 1}{\lambda \Delta t} - 1 \right] \; = \; u_0 \exp(k\lambda \Delta t) \, \frac{\lambda \Delta t}{2} + \text{h.o.t.} \tag{2.11}$$

where the very last equality was obtained via the Taylor expansion for $\Delta t \to 0$, and "h.o.t." are higher order terms with respect to the time step Δt. Note that the exponential factor $\exp(k\lambda \Delta t)$ goes to unity if $\Delta t \to 0$ and the other parameters are fixed; however, if the moment of time $t = t_k$ is fixed, then this exponential is proportional to the value of the exact solution

Symbol ϵ_c stands for *consistency error* that is, by definition, obtained by substituting the exact solution into the difference scheme. The consistency error (2.11) is indeed "small" – it tends to zero as Δt tends to zero. More precisely, the error is of order one with respect to Δt. In general, the consistency error ϵ_c is said to be of order p with respect to Δt if

$$c_1 \Delta t^p \;\leq\; |\epsilon_c| \;\leq\; c_2 \Delta t^p \tag{2.12}$$

where $c_{1,2}$ are some positive constants independent of Δt. (In the case under consideration, $p = 1$.) A very common equivalent form of this statement is the "big-oh" notation:

$$|\epsilon_c| \;=\; \mathcal{O}((\Delta t)^p)$$

(see also Introduction, p. 7). While consistency error is a convenient and very important intermediate quantity, the ultimate measure of accuracy is the *solution* error, i.e. the deviation of the numerical solution from the exact one:

$$\epsilon_k \;=\; u_k \;-\; u_{\text{exact}}(t_k) \tag{2.13}$$

The connection between consistency and solution errors will be discussed in Section 2.9.

In our current example, we can evaluate the numerical error directly. The repeated "time-stepping" by the forward Euler scheme (2.10) yields the following numerical solution:

$$u_k \;=\; (1 + \lambda \Delta t)^k u_0 \;\equiv\; (1 - \xi)^k u_0 \tag{2.14}$$

where $\xi = -\lambda \Delta t$. (Note that Re $\xi > 0$, as Re λ is assumed negative.) The k-th time step corresponds to the time instant $t_k = k\Delta t$, and so in terms of time the numerical solution can then be rewritten as

$$u_k \;=\; [(1 - \xi)^{1/\xi}]^{-\lambda t_k}\, u_0 \tag{2.15}$$

From basic calculus, the expression in the square brackets tends to e^{-1} as $\xi \to 0$, and hence u_k tends to the exact solution (2.2) $u_0 \exp(\lambda t_k)$ as $\Delta t \to 0$. Thus in the limit of small time steps the forward Euler scheme works as expected.

However, in practice, when equations and systems much more complex than our example are solved, very small step sizes may lead to prohibitively high computational costs due to a large number of time steps involved. It is therefore important to examine the behavior of the numerical solution for any given positive value of the time step rather than only in the limit $\Delta t \to 0$. Three qualitatively different cases emerge from (2.14):

$$\begin{cases} |1 + \lambda \Delta t| < 1 \;\Leftrightarrow\; \Delta t < \Delta t_{\min}, \text{ numerical solution decays (as it should);} \\ |1 + \lambda \Delta t| > 1 \;\Leftrightarrow\; \Delta t > \Delta t_{\min}, \text{ numerical solution diverges;} \\ |1 + \lambda \Delta t| = 1 \;\Leftrightarrow\; \Delta t = \Delta t_{\min}, \text{ numerical solution oscillates.} \end{cases}$$

where

$$\Delta t_{\min} = -\frac{2\mathrm{Re}\,\lambda}{|\lambda|^2}, \quad \mathrm{Re}\,\lambda < 0 \tag{2.16}$$

$$\Delta t_{\min} = \frac{2}{|\lambda|}, \quad \lambda < 0 \;\; (\lambda \text{ real}) \tag{2.17}$$

For the purposes of this introduction, we shall call a difference scheme *stable* if, for a given initial condition, the numerical solution remains bounded for all time steps; otherwise the scheme is *unstable*.[2] It is clear that in the second and third case above the numerical solution is *qualitatively* incorrect. The forward Euler scheme is stable only for sufficiently small time steps – namely, for

$$\Delta t < \Delta t_{\min} \qquad \text{(stability condition for the forward Euler scheme)} \tag{2.18}$$

Schemes that are stable only for a certain range of values of the time step are called *conditionally stable*. Schemes that are stable for *any* positive time step are called *unconditionally stable*.

It is not an uncommon misconception to attribute the numerical instability to round-off errors. While round-off errors can exacerbate the situation, it is clear from (2.14) the instability will manifest itself even in exact arithmetic if the time step is not sufficiently small.

The *backward Euler* difference scheme (2.6) is substantially different in this regard. The numerical solution for that scheme is easily found to be

$$u_k = (1 - \lambda\Delta t)^{-k} u_0 \tag{2.19}$$

In contrast with the forward Euler method, for negative Re λ this solution is bounded (and decaying in time) *regardless of the step size* Δt. That is, the backward Euler scheme is unconditionally stable. However, there is a price to pay for this advantage: the scheme is an equation with respect to u_{k+1}. In the current example, solution of this equation is trivial (just divide by $1 - \lambda\Delta t$), but for nonlinear differential equations, and especially for (linear and nonlinear) *systems* of differential equations the computational cost of computing the solution at each time step may be high.

Difference schemes that require solution of a system of equations to find u_{k+1} are called *implicit*; otherwise the scheme is explicit. The forward Euler scheme is explicit, and the backward Euler scheme is implicit. The derivation of the consistency error for the backward Euler scheme is completely analogous to that of the forward Euler scheme, and the result is essentially the same, except for a sign difference:

$$\epsilon_c = -u_0 \exp(k\lambda\Delta t)\,\frac{\lambda\Delta t}{2} + \text{h.o.t.} \tag{2.20}$$

[2] More specialized definitions of stability can be given for various classes of schemes; see e.g. C.W. Gear [Gea71], J.C. Butcher [But03], E. Hairer *et al.* [HrW93] as well as the following sections of this chapter.

As in the forward Euler case, the exponential factor tends to unity as the time step goes to zero, but only if k and λ are fixed.

The very popular Crank–Nicolson scheme[3] can be viewed as an approximation of the original differential equation at time $t_{k+1/2} \equiv t_k + \Delta t/2$:

$$\frac{u_{k+1} - u_k}{\lambda \Delta t} - \frac{u_k + u_{k+1}}{2} = 0, \quad k = 0, 1, \ldots \tag{2.21}$$

Indeed, the left hand side of this equation is the central-difference approximation (completely analogous to (2.8), but with a twice smaller time step), while the right hand side approximates the value of $u(t_{k+1/2})$.

The time-stepping procedure for the Crank–Nicolson scheme is

$$\left(1 - \frac{\lambda \Delta t}{2}\right) u_{k+1} = \left(1 + \frac{\lambda \Delta t}{2}\right) u_k, \quad k = 0, 1, \ldots \tag{2.22}$$

and the numerical solution of the model problem is

$$u_k = \left(\frac{1 + \lambda \Delta t/2}{1 - \lambda \Delta t/2}\right)^k u_0 \tag{2.23}$$

Since the absolute value of the fraction here is less than one for all positive (even very large) time steps, the Crank–Nicolson scheme is unconditionally stable. Its consistency error is again found by substituting the exact solution (2.2) into the scheme (2.21). The result is

$$\epsilon_c = -u_0 \exp(k\lambda\Delta t) \frac{(\lambda \Delta t)^2}{12} + \text{h.o.t.} \tag{2.24}$$

The consistency error is seen to be of *second* order – as such, it is (for sufficiently small time steps) much smaller than the error of both Euler schemes.

2.3 Exact Schemes

As we have seen, the consistency error can be made smaller if one switches from Euler methods to the Crank–Nicolson scheme. Can the consistency error be reduced even further? One may try to "mix" the forward and backward

[3] Often misspelled as Crank-Nicholson. After John Crank (born 1916), British mathematical physicist, and Phyllis Nicolson (1917–1968), British physicist. http://www-groups.dcs.st-and.ac.uk/ history/Mathematicians/Nicolson.html http://www-groups.dcs.st-and.ac.uk/ history/Mathematicians/Crank.html The original paper is: J. Crank and P. Nicolson, A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type, *Proc. Cambridge Philos. Soc.*, vol. 43, pp. 50–67, 1947. [Re-published in: John Crank 80th birthday special issue of *Adv. Comput. Math.*, vol. 6, pp. 207–226, 1997.]

Euler schemes in a way similar to the Crank–Nicolson scheme, but by assigning some other weights θ and $(1-\theta)$, instead of $\frac{1}{2}$, to u_k and u_{k+1} in (2.21). However, it would soon transpire that the Crank–Nicolson scheme in fact has the smallest consistency error in this family of schemes, so nothing substantially new is gained by introducing the alternative weighting factors.

Nevertheless one can easily construct schemes whose consistency error cannot be beaten. Indeed, here is an example of such a scheme:

$$\frac{u_k}{u_{\text{exact}}(t_k)} - \frac{u_{k+1}}{u_{\text{exact}}(t_{k+1})} = 0 \tag{2.25}$$

More specifically for the equation under consideration

$$\frac{u_k}{\exp(-\lambda t_k)} - \frac{u_{k+1}}{\exp(-\lambda t_{k+1})} = 0 \tag{2.26}$$

Equivalently,

$$u_k - u_{k+1}\exp(\lambda\Delta t) = 0 \tag{2.27}$$

Obviously, by construction of the scheme, the analytical solution satisfies the difference equation *exactly* – that is, the consistency error of the scheme is *zero*. One cannot do any better than that!

The first reaction may be to dismiss this construction as cheating: the scheme makes use of the exact solution that in fact needs to be found. If the exact solution is known, the problem has been solved and no difference scheme is needed. If the solution is not known, the coefficients of this "exact" scheme are not available.

Yet the idea of "exact" schemes like (2.25) proves very useful. Even though the exact solution is usually not known, excellent approximations for it can frequently be found and used to construct a difference scheme. One key observation is that such approximations need not be global (i.e. valid throughout the computational domain). Since difference schemes are *local*, all that is needed is a good *local* approximation of the solution. Local approximations are much more easily obtainable than global ones. In fact, the Taylor series expansion that was implicitly used to construct the Euler and Crank–Nicolson schemes, and that will be more explicitly used in the following subsection, is just an example of a local approximation.

The construction of "exact" schemes represents a shift in perspective. The objective of Taylor-based schemes is to approximate the *differential operator* – for example, d/dt – with a suitable finite difference, and consequently the differential equation with the respective FD scheme. The objective of the "exact" schemes is to approximáte the *solution*.

Approximation of the differential operator is a very powerful tool, but it carries substantial redundancy: it is applicable to *all* sufficiently smooth functions to which the differential operator could be applied. By focusing on the *solution* only, rather than on a wide class of smooth functions, one can reduce or even eliminate this redundancy. As a result, the accuracy of the

numerical solution can be improved dramatically. This set of ideas will be explored in Chapter 4.

The following figures illustrate the accuracy of different one-step schemes for our simple model problem with parameter $\lambda = -10$. Fig. 2.1 shows the analytical and numerical solutions for time step $\Delta t = 0.05$. It is evident that the Crank–Nicolson scheme is substantially more accurate than the Euler schemes. The numerical errors are quantified in Fig. 2.2. As expected, the exact scheme gives the true solution up to the round-off error.

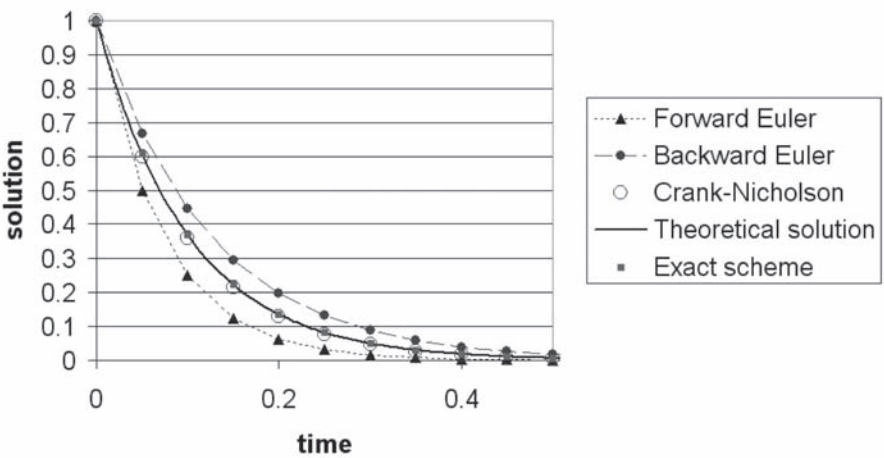


Fig. 2.1. Numerical solution for different one-step schemes. Time step $\Delta t = 0.05$. $\lambda = -10$.

For a larger time step $\Delta t = 0.25$, the forward Euler scheme exhibits instability (Fig. 2.3). The exact scheme still yields the analytical solution to machine precision. The backward Euler and Crank–Nicolson schemes are stable, but the numerical errors are higher than for the smaller time step.

R.E. Mickens [Mic94] derives "exact" schemes from a different perspective and extends them to a family of "nonstandard" schemes defined by a set of heuristic rules. We shall see in Chapter 4 that the "exact" schemes are a very natural particular case of a new finite-difference calculus – "Flexible Local Approximation MEthods" (FLAME).

2.4 Some Classic Schemes for Initial Value Problems

For completeness, this section presents a brief overview of a few popular time-stepping schemes for Ordinary Differential Equations (ODE).

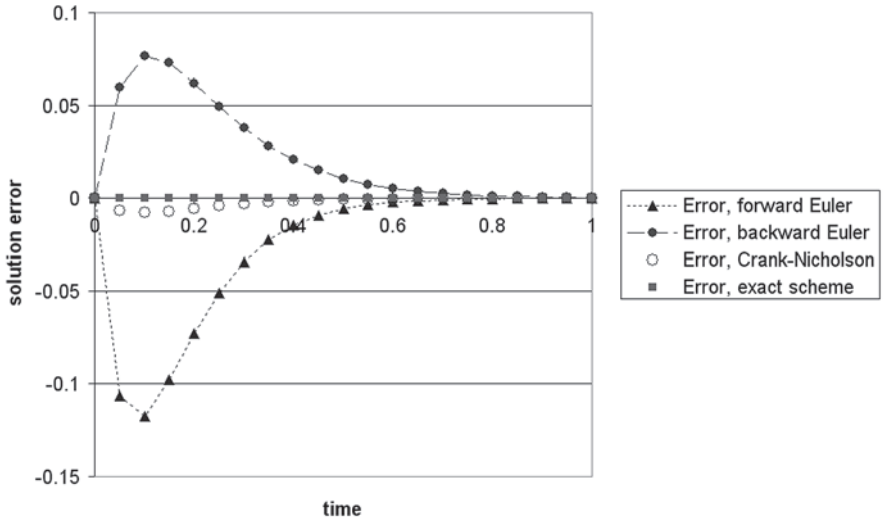


Fig. 2.2. Numerical errors for different one-step schemes. Time step $\Delta t = 0.05$. $\lambda = -10$.

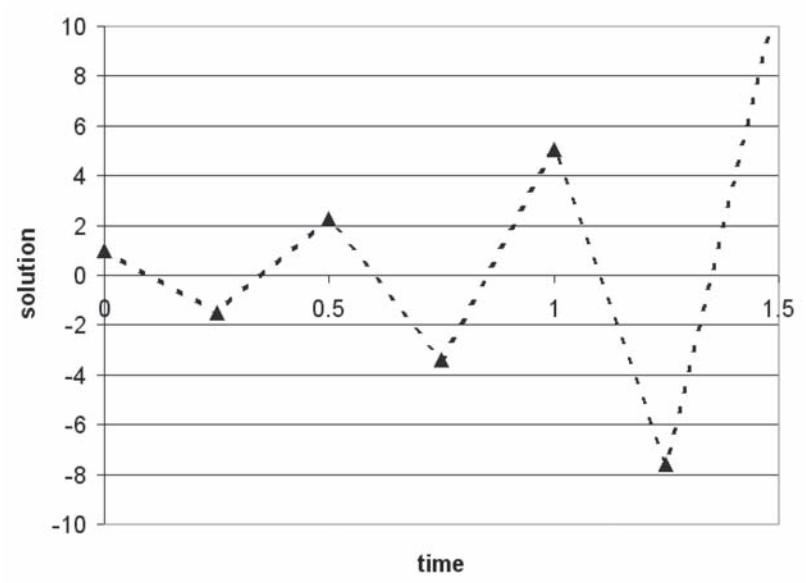


Fig. 2.3. Numerical solution for the forward Euler scheme. Time step $\Delta t = 0.25$. $\lambda = -10$.

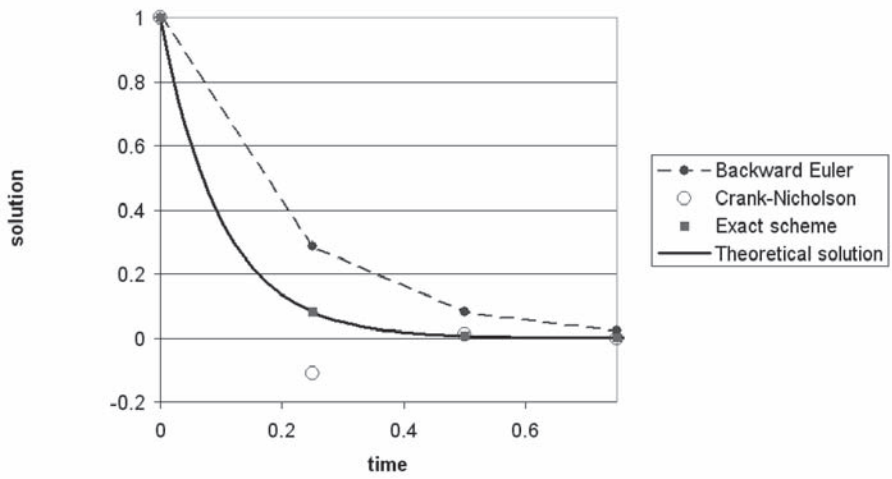


Fig. 2.4. Numerical solution for different one-step schemes. Time step $\Delta t = 0.25$. $\lambda = -10$.

2.4.1 The Runge–Kutta Methods

This introduction to Runge–Kutta (R-K) methods follows the elegant exposition by E. Hairer *et al.* [HrW93]. The main idea dates back to C. Runge's original paper of 1895.

The goal is to construct high order difference schemes for the ODE

$$y'(t) \;=\; f(t,y), \quad y(t_0) \;=\; y_0 \tag{2.28}$$

Our starting point is a simpler problem, with the right hand side independent of y:

$$y'(t) \;=\; f(t), \quad y(t_0) \;=\; y_0 \tag{2.29}$$

This problem not only has an analytical solution

$$y(t) \;=\; y_0 \;+\; \int_{t_0}^{t} f(\tau)d\tau \tag{2.30}$$

but also admits accurate approximations via numerical quadratures. For example, the midpoint rule gives

$$y_1 \;\equiv\; y(t_1) \;\approx\; y_0 \;+\; \Delta t_0 f\left(t_0 + \frac{\Delta t_0}{2}\right)$$

$$y_2 \;\equiv\; y(t_2) \;\approx\; y_1 \;+\; \Delta t_1 f\left(t_1 + \frac{\Delta t_1}{2}\right)$$

and so on. Here t_0, t_1, etc., are a discrete set of points in time, and the time steps $\Delta t_0 = t_1 - t_0$, $\Delta t_1 = t_2 - t_1$, etc., do not have to be equal.

It is straightforward to verify that this numerical quadrature (that doubles as a time-stepping scheme) has second order accuracy with respect to the maximum time step.

An analogous formula for taking the numerical solution of the original equation (2.28) from a generic point t in time to $t + \Delta t$ would be

$$y(t + \Delta t) \;\approx\; y(t) \;+\; \Delta t f\left(t + \frac{\Delta t}{2},\, y\left(t + \frac{\Delta t}{2}\right)\right) \tag{2.31}$$

The obstacle is that the value of y at the midpoint $t + \frac{\Delta t}{2}$ is not directly available. However, this value may be found approximately via the forward Euler scheme with the time step $\Delta t/2$:

$$y\left(t + \frac{\Delta t}{2}\right) \;\approx\; y(t) \;+\; \frac{\Delta t}{2} f\left(t, y(t)\right) \tag{2.32}$$

A valid difference scheme can now be produced by inserting this midpoint value into the numerical quadrature (2.31). The customary way of writing the overall procedure is as the following sequence:

$$k_1 \;=\; f(t, y) \tag{2.33}$$

$$k_2 \;=\; f\left(t + \frac{\Delta t}{2},\; y(t) + \frac{\Delta t}{2} k_1\right) \tag{2.34}$$

$$y(t + \Delta t) \;=\; y(t) \;+\; \Delta t\, k_2 \tag{2.35}$$

This is the simplest R-K method with two *stages* (k_1 is computed at the first stage and k_2 at the second). The generic form of an s-stage explicit R-K method is as follows [HrW93]:

$$\begin{aligned}
k_1 &= f(t_0, y_0) \\
k_2 &= f\left(t_0 + c_2 \Delta t,\, y_0 + \Delta t a_{21} k_1\right) \\
k_3 &= f\left(t_0 + c_3 \Delta t,\, y_0 + \Delta t\left(a_{31} k_1 + a_{32} k_2\right)\right) \\
&\cdots \\
k_s &= f\left(t_0 + c_s \Delta t,\, y_0 + \Delta t\left(a_{s1} k_1 + \cdots + a_{s,s-1} k_{s-1}\right)\right) \\
y(t + h) &= y_0 \;+\; \Delta t(b_1 k_1 + b_2 k_2 + \cdots + b_s k_s)
\end{aligned}$$

The procedure is indeed explicit, as the computation at each subsequent stage depends only on the values computed at the previous stages. The "input data" for the R-K method at any given time step consists only of one value y_0 at the beginning of this step and does not include any other previously computed values. Thus the R-K time step sizes can be chosen independently, which is very useful for adaptive algorithms. The multi-*stage* method should not be confused with multi-*step* schemes (such as e.g. the Adams methods, Section 2.4.2 below) where the input data at each discrete time point contains the values of y at several previous steps. Changing the time step in multistep methods may be cumbersome and may require "re-initialization" of the algorithm.

To write R-K schemes in a compact form, it is standard to collect all the coefficients a, b and c in J. Butcher's tableau:

$$
\begin{array}{c|cccccc}
0 & & & & & & \\
c_2 & a_{21} & & & & & \\
c_3 & a_{31} & a_{32} & & & & \\
\dots & \dots & \dots & \dots & \dots & & \\
\dots & \dots & \dots & \dots & \dots & \dots & \\
c_s & a_{s1} & a_{s2} & \dots & \dots & \dots & a_{s,s-1} \\
\hline
 & b_1 & b_2 & \dots & \dots & \dots & b_s
\end{array}
$$

One further intuitive observation is that the k parameters in the R-K method are values of function f at some intermediate points. As a rule, one wants these intermediate points to be close to the actual solution $y(t)$ of (2.28). Then, according to (2.28), the ks also approximate the time derivative of y over the current time step. Thus at the i-th stage of the procedure function f is evaluated, roughly speaking, at point $(t_0 + c_i\Delta t,\ y_0 + (a_{i1} + \cdots + a_{i,s-1})y'(t_0)\Delta t)$. From these considerations, condition

$$c_i \;=\; a_{i1} + \cdots + a_{i,s-1}, \quad i = 2, 3 \dots s$$

emerges as natural (although not, strictly speaking, necessary).

The number of stages is in general different from the order of the method (i.e. from the asymptotic order of the consistency error with respect to the time step), and one wishes to find the free parameters a, b and c that would maximize the order. For $s \geq 5$, no explicit s-stage R-K method of order s exists (E. Hairer *et al.* [HrW93], J.C. Butcher [But03]). However, a family of four-stage explicit R-K methods of fourth order are available [HrW93, But03]. The most popular of these methods are

$$
\begin{array}{c|cccc}
0 & & & & \\
1/2 & 1/2 & & & \\
1/2 & 0 & 1/2 & & \\
1 & 0 & 0 & 1 & \\
\hline
 & 1/6 & 2/6 & 2/6 & 1/6
\end{array}
$$

and

$$
\begin{array}{c|cccc}
0 & & & & \\
1/3 & 1/3 & & & \\
2/3 & -1/3 & 1 & & \\
1 & 1 & -1 & 1 & \\
\hline
 & 1/8 & 3/8 & 3/8 & 1/8
\end{array}
$$

Stability conditions for explicit Runge–Kutta schemes can be obtained along the following lines. For the model scalar equation (2.1)

$$\frac{dy}{dt} = \lambda y \quad \text{on } [0, t_{\max}], \quad u(0) = u_0 \tag{2.36}$$

the exact solution changes by the factor of $\exp(\lambda h)$ over one time step. If the R-K method is of order p, the respective factor in the growth of the numerical solution is the Taylor approximation

$$T(\xi) = \sum_{k=0}^{p} \frac{\xi^k}{k!}, \qquad \xi \equiv \lambda \Delta t$$

to this exponential factor. Stability regions then correspond to $|T(\xi)| < 1$ in the complex plane $\xi \equiv \lambda \Delta t$ (Fig. 2.5).

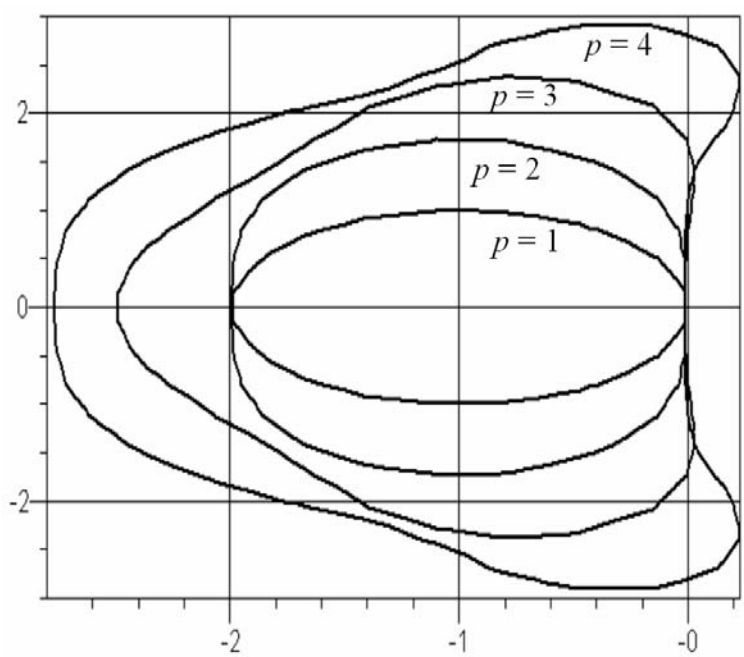


Fig. 2.5. Stability regions in the $\lambda\Delta t$-plane for explicit Runge–Kutta methods of orders one through four.

Further analysis of R-K methods can be found in monographs by J. Butcher [But03], E. Hairer *et al.* [HrW93], and C.W. Gear [Gea71].

2.4.2 The Adams Methods

Adams methods are a popular class of *multistep* schemes, where the solution values from several previous time steps are utilized to find the numerical solution at the subsequent step. This is accomplished by polynomial interpolation. The following brief summary is due primarily to E. Hairer *et al.* [HrW93].

Consider again the general ODE (2.28) (reproduced here for easy reference):

$$y'(t) \;=\; f(t,y), \quad y(t_0) \;=\; y_0 \tag{2.37}$$

Let the grid be uniform, $t_i = t_0 + i\Delta t$, and integrate the differential equation over one time step:

$$y(t_{n+1}) \;=\; y(t_n) \;+\; \int_{t_n}^{t_{n+1}} f(t, y(t))\, dt \tag{2.38}$$

The integrand is a function of the unknown solution and obviously is not directly available; however, it can be approximated by a polynomial $p(t)$ passing through k previous numerical solution values $(t_i, f(y_i))$. The numerical solution at time step $n+1$ is then found as

$$y_{n+1} \;=\; y_n \;+\; \int_{t_n}^{t_{n+1}} p(t)\, dt \tag{2.39}$$

Coefficients of $p(t)$ can be found explicitly (e.g. via backward differences), and the scheme is then obtained after inserting the expression for p into (2.39). This explicit calculation appears in all texts on numerical methods for ODE and is not included here.

Adams methods can also be used in the *Nordsieck form*, where instead of the values of function f at the previous time steps approximate Taylor coefficients for the solution are stored. These approximate coefficients form the Nordsieck vector $(y_n,\ \Delta t y'_n,\ \frac{\Delta t^2}{2} y''_n,\ \ldots,\ \frac{\Delta t^k}{k!}\, y_n^{(k)})$. This form makes it easier to change the time step size as needed.

2.4.3 Stability of Linear Multistep Schemes

It is clear from the introduction in Section 2.2 that stability characteristics of the difference scheme are of critical importance for the numerical solution. Stability depends on the intrinsic properties of the underlying differential equation (or a system of ODE), as well as on the difference scheme itself and the mesh size. This section highlights the key points in the stability analysis of linear multistep schemes; the results and conclusions will be used, in particular, in the next section (stiff systems).

Stability of linear multistep schemes is covered in all texts on FD schemes for ODE (e.g. C.W. Gear [Gea71], J. Butcher [But03], E. Hairer *et al.* [HrW93], U.M. Ascher & L.R. Petzold [AP98]). A comprehensive classification of types

of stability is given in the book by J.D. Lambert [Lam91]. This section, for the most part, follows Lambert's presentation.

Consider the test system of equations

$$y' \;=\; Ay, \quad y \in \mathbb{R}^n \tag{2.40}$$

where all eigenvalues of matrix A are for simplicity assumed to be distinct and to have strictly negative real parts, so that the system is stable. Further, let a linear k-step method be

$$\sum_{j=0}^{k} \alpha_j y_{+j} \;=\; \Delta t \sum_{j=0}^{k} \beta_j f_{+j} \tag{2.41}$$

where f is the right hand side of the system, h is (as usual) the mesh size, and index $+j$ indicates values at the j-th time step (the "current" step corresponding to $j = 0$). In our case, the right hand side $f = Ay$, and the multistep scheme becomes

$$\sum_{j=0}^{k} (\alpha_j I \;-\; \Delta t \beta_j A)\, y_{+j} \;=\; 0 \tag{2.42}$$

Since A is assumed to have distinct eigenvalues, it is diagonalizable, i.e.

$$Q^{-1} A Q \;=\; \Lambda \;\equiv\; \mathrm{diag}(\lambda_1, \ldots, \lambda_n) \tag{2.43}$$

where Q is a nonsingular matrix. The same transformation can then be applied to the whole scheme (2.42) by multiplying it with Q^{-1} on the left and introducing a variable change $y = Qz$. It is easy to see that, since the system matrix becomes diagonal upon this transformation, the system splits up into completely decoupled equations for each z_i, $i = 1, 2, \ldots, n$. With some abuse of notation now, dropping the index i for z_i and the respective eigenvalue λ_i, we get the scalar version of the scheme

$$\sum_{j=0}^{k} (\alpha_j \;-\; \Delta t \beta_j \lambda) z_{+j} \;=\; 0 \tag{2.44}$$

From the theory of difference equations it is well known that stability is governed by the roots[4] r_s ($s = 1,2, \ldots, k$) of the characteristic equation

$$\sum_{j=0}^{k} (\alpha_j \;-\; \Delta t \lambda \beta_j)\, r^j \;=\; 0 \tag{2.45}$$

Clearly, stability depends on the (dimensionless) parameter $h\lambda$.

The multistep method is said to be *absolutely stable* for given $\lambda \Delta t$ if all the roots r_s of the characteristic polynomial for this value of $\lambda \Delta t$ lie strictly inside the unit circle in the complex plane.

[4] Lambert's notation is used here.

The set of points $\lambda\Delta t$ in the $\lambda\Delta t$-plane for which the scheme is absolutely stable is called the *region of absolute stability*. For illustration, let us recall the simplest case – one-step schemes for the scalar equation $y' = \lambda y$:

$$\frac{y_{+1} - y_0}{\Delta t} = \lambda\left(\theta y_0 + (1-\theta)y_{+1}\right) \tag{2.46}$$

For $\theta = 0$ and 1, this is the implicit/explicit Euler method, respectively; for $\theta = 0.5$ it is the Crank–Nicolson (trapezoidal) scheme. The characteristic equation is obtained in a standard way, by formally substituting r^1 for y_{+1} and $r^0 = 1$ for y_0:

$$\frac{r-1}{\Delta t} = \lambda\left(\theta + (1-\theta)r\right) \tag{2.47}$$

The root is

$$r = \frac{1 + \lambda\theta\Delta t}{1 - \lambda(1-\theta)\Delta t} \tag{2.48}$$

For the explicit Euler scheme ($\theta = 1$)

$$r_{\text{expl.Euler}} = 1 + \lambda\Delta t \tag{2.49}$$

and so the region of absolute stability in the $\lambda\Delta t$-plane is the unit circle centered at -1 (Fig. 2.6).

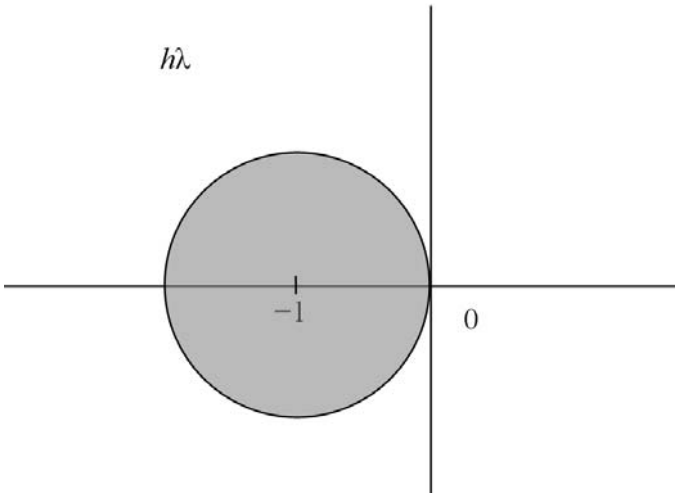


Fig. 2.6. Stability region of the explicit Euler method is the unit circle (shaded).

For the *implicit* Euler scheme ($\theta = 0$)

$$r_{\text{impl.Euler}} = \frac{1}{1 - \lambda\Delta t} \tag{2.50}$$

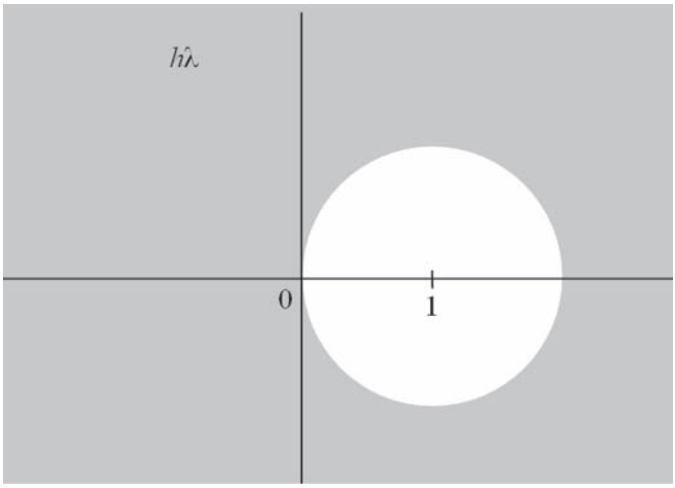


Fig. 2.7. Stability region of the implicit Euler method is the shaded area *outside* the unit circle.

the region of absolute stability is *outside* the unit circle centered at 1 (Fig. 2.7).

This stability region includes all negative values of $\lambda\Delta t$ – that is, for a negative λ, the scheme is stable for *any* (positive) time step. In addition, curiously enough, the scheme is stable in a vast area with *positive* $\lambda\Delta t$ – i.e. the numerical solution may decay exponentially when the exact one *grows* exponentially. This latter feature is somewhat undesirable but is typically of little significance, as in most cases the underlying differential equations describe stable systems with decaying solutions.

What about the Crank–Nicolson scheme? For $\theta = 0.5$ we have

$$r_{\text{Crank-Nicolson}} = \frac{1 + \lambda\Delta t/2}{1 - \lambda\Delta t/2} \tag{2.51}$$

and it is then straightforward to verify that the stability region is the half-plane $\lambda\Delta t < 0$ (Fig. 2.8).

The region of stability is clearly a key consideration for choosing a suitable class of schemes and the mesh size such that $h\lambda$ lies inside the region of stability.

2.4.4 Methods for Stiff Systems

One can identify two principal constraints on the choice of the time step in a numerical scheme for ODE. The first constraint has to do with the desired *approximation accuracy* (i.e. consistency error): if the solution varies smoothly and slowly in time, it can be approximated with sufficient accuracy even if the time step is large.

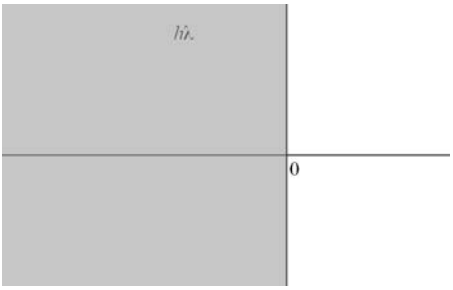


Fig. 2.8. Stability region of the Crank–Nicolson scheme is the left half-plane.

The second constraint is imposed by *stability* of the scheme. Let us recall, for example, that the stability condition for the simplest one-step scheme – the forward Euler method – is $\Delta t < 2/|\lambda|$ (2.18), (2.17) for real negative λ, in reference to the test equation (2.1)

$$\frac{dy}{dt} = \lambda y \quad \text{on } [0, t_{\max}], \quad u(0) = u_0 \tag{2.52}$$

More advanced explicit methods may have broader stability regions: see e.g. Fig. 2.5 for Runge–Kutta methods in Section 2.4.1. However, the improvement is not dramatic; for example, for the four-stage fourth-order Runge–Kutta method, the step size cannot exceed $\sim 2.785/|\lambda|$.

For a single scalar equation (2.52) with $\lambda < 0$ and a decaying exponential solution, the accuracy and stability restrictions on the time step size are commensurate. Indeed, accuracy calls for the step size on the order of the relaxation time $1/\lambda$ or less, which is well within the stability limit even for the simplest forward Euler scheme.

However, for *systems* of equations the stability constraint on the step size can be much more severe than the accuracy limit. Consider the following example:

$$\frac{dy_1}{dt} = \lambda_1 y_1; \qquad \lambda_1 = -1 \tag{2.53}$$

$$\frac{dy_2}{dt} = \lambda_2 y_2; \quad \lambda_2 = -1000 \tag{2.54}$$

The second component (y_2) dies out when $t \gg 1/|\lambda_2| = 10^{-3}$ and can then be neglected; beyond that point, the approximation accuracy would suggest the time step commensurate with the relaxation time of the first component, $1/|\lambda_1| = 1$. However, the stability condition $\Delta t \leq c/|\lambda|$ (where c depends on the method but is not much greater than 2–3 for most practical explicit schemes) has to hold for *both* λ and limits the time step to approximately $1/|\lambda_2| = 10^{-3}$.

In other words, the time step that would provide good approximation accuracy exceeds the stability limit by a factor of about 1000. A brute force

approach is to use a very small time step and accept the high computational cost as well as the tremendous redundancy in the numerical solution that will remain virtually unchanged over one time step.

An obvious possibility for a system with *decoupled* components is to solve the problem separately for each component. In the example above, one could time-step y_1 with $\Delta t_1 \sim 0.1$ for about 50 steps (after which y_1 will die out) and y_2 with $\Delta t_2 \sim 10^{-4}$ also for about 50 steps. However, decoupled systems are a luxury that one seldom has in practical problems. For example, the system of ODEs

$$z'(t) \;=\; Az; \;\; z(t) \in \mathbb{R}^2; \qquad A \;=\; \begin{pmatrix} 500.5 & -499.5 \\ -499.5 & 500.5 \end{pmatrix} \tag{2.55}$$

poses the same stability problem for explicit schemes as the previous example – simply because matrix A is obtained from the diagonal matrix $D = \mathrm{diag}(1, 1000)$ of the previous example by an orthogonal transformation $A = Q'DQ$, with

$$Q \;=\; \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$$

The "fast" and "slow" components, with their respective time scales, are now mixed up, but this is no longer immediately obvious. Recovering the two components is equivalent to solving a full eigenvalue-eigenvector problem for the system matrix, which can be done for small systems but is inefficient or even impossible for large ones. The situation is even more complicated for nonlinear problems and systems with time-varying coefficients.

A practical alternative lies in switching to *implicit* difference schemes. In return for excellent stability properties, one pays the price of having to solve for the unknown value of the numerical solution y_{n+1} at the next time step. This is in general a nonlinear equation (for a scalar ODE) or a nonlinear system of algebraic equations (for a system of ODEs, y being in that case a Euclidean vector).

Recall that for the ODE

$$y'(t) \;=\; f(t, y) \tag{2.56}$$

the simplest implicit scheme – the backward Euler method – is

$$y_{n+1} \;-\; y_n \;=\; \Delta t\, f(t_{n+1}, y_{n+1}) \tag{2.57}$$

A set of schemes that generalize the backward Euler algorithm to higher orders is due to C.W. Gear [Gea67, Gea71, HrW93] and is called "Backward Differentiation Formulae" (BDF). For illustration, let us derive the second order BDF scheme, the derivation of higher order schemes being analogous.

The second order scheme involves three grid points: $t_{-1} = t_0 - \Delta t$, t_0 and $t_{+1} = t_0 + \Delta t$; quantities related to the "current" time step t_0 will be labeled with index 0, quantities related to the previous and the next step will be

labeled with -1 and $+1$, respectively. The starting point is *almost* the same as for explicit Adams methods: an interpolation polynomial $p(t)$ (quadratic for the second order scheme) that passes through three points (t_{-1}, y_{-1}), (t_0, y_0) and (t_{+1}, y_{+1}). The values y_0 and y_{-1} of the solution at the current and previous steps are known. The value y_{+1} at the next step is an *un*known parameter, and a suitable condition is needed to evaluate it.

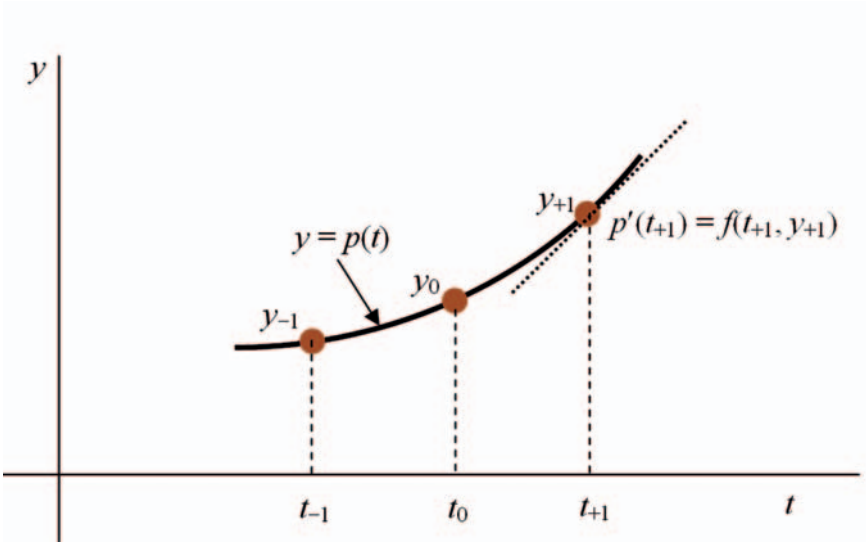


Fig. 2.9. Second-order BDF involves quadratic polynomial interpolation over three points: (t_{-1}, y_{-1}), (t_0, y_0) and (t_{+1}, y_{+1}).

In BDF, the following condition is imposed: the interpolating polynomial $p(t)$ must satisfy the underlying differential equation at time t_{+1}, i.e.

$$p'(t_{+1}) \;=\; f(t_{+1}, y_{+1}) \tag{2.58}$$

To find this interpolation polynomial and then the BDF scheme itself, let us for convenience move the origin of the coordinate system to the midpoint of the stencil and set $t_0 = 0$. Lagrange interpolation through the three points then gives

$$\begin{aligned} p(t) \;&=\; y_{-1}\,\frac{t(t-\Delta t)}{(-\Delta t)\cdot(-2\Delta t)} \;+\; y_0\,\frac{(t+\Delta t)(t-\Delta t)}{\Delta t\cdot(-\Delta t)} \;+\; y_{+1}\,\frac{(t+\Delta t)t}{2\Delta t\cdot\Delta t} \\ &=\; y_{-1}\,\frac{t(t-\Delta t)}{2\Delta t^2} \;-\; y_0\,\frac{(t+\Delta t)(t-\Delta t)}{\Delta t^2} \;+\; y_{+1}\,\frac{(t+\Delta t)t}{2\Delta t^2} \end{aligned} \tag{2.59}$$

The derivative of p (needed to impose condition (2.58) at the next step) is

$$p'(t) \;=\; \frac{y_{-1}}{2\Delta t^2}\,(2t-\Delta t) \;-\; \frac{y_0}{\Delta t^2}\,2t \;+\; \frac{y_{+1}}{2\Delta t^2}\,(2t+\Delta t) \tag{2.60}$$

Condition (2.58) is obtained by substituting $t = t_{+1}$:

$$p'(t_{+1}) \;=\; \frac{y_{-1}}{2\Delta t} \;-\; \frac{2y_0}{\Delta t} \;+\; \frac{3y_{+1}}{2\Delta t} \;=\; f(t_{+1}, y_{+1}) \tag{2.61}$$

or equivalently

$$\frac{3}{2}\,y_{+1} \;-\; 2y_0 \;+\; \frac{1}{2}\,y_{-1} \;=\; \Delta t\, f(t_{+1}, y_{+1}) \tag{2.62}$$

This is Gear's second order method. The scheme is implicit – it constitutes a (generally nonlinear) equation with respect to y_{+1} or, in the case of a vector problem ($y \in \mathbb{R}^n$), a system of equations. In practice, iterative linearization by the Newton–Raphson method is used and suitable linear system solvers are applied in the Newton–Raphson loop.

For reference, here is a list of BDF of orders k from one through six [HrW93]. The first order BDF scheme coincides with the implicit Euler method. BDF schemes of orders higher than six are unstable.

$$\begin{aligned}
y_{+1} - y_0 &= \Delta t\, f_{+1}\\
\frac{3}{2}\,y_{+1} - 2y_0 + \frac{1}{2}\,y_{-1} &= \Delta t\, f_{+1}\\
\frac{11}{6}\,y_{+1} - 3y_0 + \frac{3}{2}\,y_{-1} - \frac{1}{3}\,y_{-2} &= \Delta t\, f_{+1}\\
\frac{25}{12}\,y_{+1} - 4y_0 + 3\,y_{-1} - \frac{4}{3}\,y_{-2} + \frac{1}{4}\,y_{-3} &= \Delta t\, f_{+1}\\
\frac{137}{60}\,y_{+1} - 5y_0 + 5\,y_{-1} - \frac{10}{3}\,y_{-2} + \frac{5}{4}\,y_{-3} - \frac{1}{5}\,y_{-4} &= \Delta t\, f_{+1}\\
\frac{147}{60}\,y_{+1} - 6y_0 + \frac{15}{2}\,y_{-1} - \frac{20}{3}\,y_{-2} + \frac{15}{4}\,y_{-3} - \frac{6}{5}\,y_{-4} + \frac{1}{6}\,y_{-5} &= \Delta t\, f_{+1}
\end{aligned}$$

Since stability considerations are of paramount importance in the choice of difference schemes for stiff problems, an elaborate classification of schemes based on their stability properties – or more precisely, on their regions of absolute stability (see Section 2.4.3) – has been developed. The relevant material can be found in C.W. Gear's monograph [Gea71] and, in a more complete form, in J.D. Lambert's book [Lam91]. What follows is a brief summary of this stability classification.

A hierarchy of definitions of stability classes with progressively wider regions of stability are (Lambert's definitions are adopted):
A_0-stability $\Longleftarrow$ $A(0)$-stability $\Longleftarrow$ $A(\alpha)$-stability $\Longleftarrow$ stiff-stability $\Longleftarrow$ A-stability $\Longleftarrow$ L-stability

Definition 1. *A method is said to be A_0-stable if its region of absolute stability includes the (strictly) negative real semiaxis.*

Definition 2. *[Gea71], [Lam91] A method is said to be $A(\alpha)$-stable, $0 < \alpha < \pi/2$, if its region of absolute stability includes the "angular" domain $|\arg(\lambda\Delta t) - \pi| \leq \alpha$ in the $\lambda\Delta t$-plane (Fig. 2.10). A method is said to be $A(0)$-stable if it is $A(\alpha)$-stable for some $0 < \alpha < \pi/2$.*

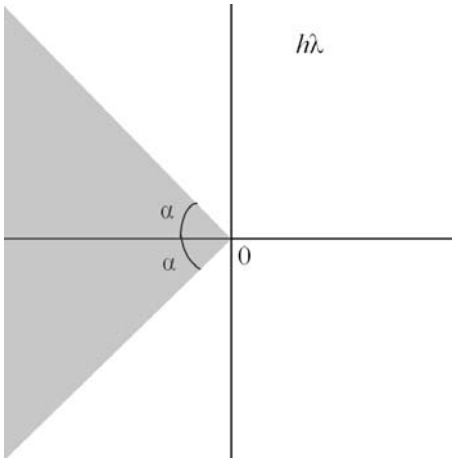


Fig. 2.10. $A(\alpha)$-stability region.

Definition 3. *[Gea71], [Lam91] A method is said to be A-stable if its region of absolute stability includes the half-plane Re $(\lambda\Delta t) < 0$.*

Definition 4. *A method is said to be stiffly-stable if its region of absolute stability includes the union of two domains (Fig. 2.11): (i) Re $(\lambda\Delta t) < -a$, and (ii) $-a \leq$ Re $(\lambda\Delta t) < 0$, $|\mathrm{Im}(\lambda\Delta t)| < c$, where a, c are positive real numbers.*

Thus stiff stability differs from A-stability in that slowly decaying but highly oscillatory solutions are irrelevant for stiff stability. The rationale is that for such solutions the time step is governed by *accuracy* requirements for the oscillatory components as much, or perhaps even more, than it is governed by stability requirements – hence this is not truly a stiff case.

Definition 5. *[Gea71, Lam91] A method is said to be L-stable if it is A-stable and, in addition, when applied to the scalar test equation $y' = \lambda y$, Re $\lambda < 0$, it yields $y_{n+1} = R(\lambda\Delta t)\, y_n$, with $|R(\lambda\Delta t)| \to 0$ as Re $\lambda\Delta t \to -\infty$.*

The notion of L-stability is motivated by the following test case. Consider one more time the Crank–Nicolson scheme applied to the model scalar equation $y' = \lambda y$:

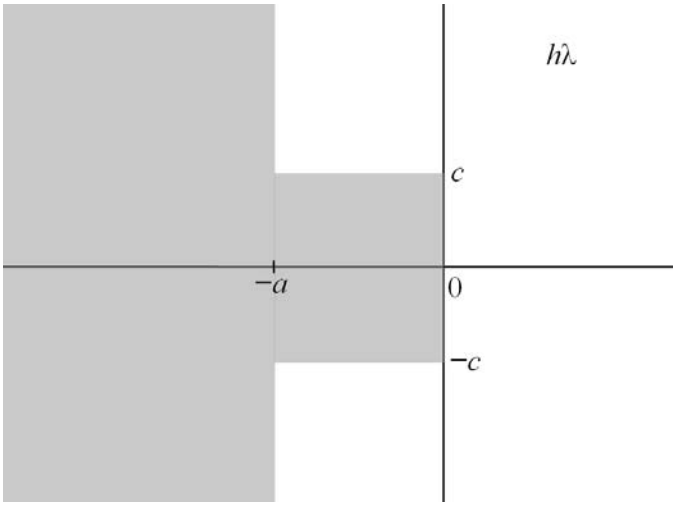


Fig. 2.11. Stiff-stability region.

$$\frac{y_{n+1} - y_n}{h} = \lambda \frac{y_{n+1} + y_n}{2} \tag{2.63}$$

The numerical solution is easily found to be

$$y_n = y_0 \left(\frac{1 + \lambda \Delta t/2}{1 - \lambda \Delta t/2} \right)^n \tag{2.64}$$

As already noted, the Crank–Nicolson scheme is absolutely stable for any $\lambda \Delta t$ with a negative real part. The solution above reflects this fact, as the expression in parentheses has the absolute value less than one for Re $\lambda \Delta t < 0$. Still, the numerical solution exhibits some undesirable behavior for "highly negative" values of λ, i.e. for $\lambda < 0$, $|\lambda| \Delta t \gg 1$. Indeed, in this case the actual solution decays very rapidly in time as $\exp(\lambda t)$, whereas the numerical solution decays very slowly but is highly oscillatory because the expression in parentheses in (2.64) is close to -1.

This is a case where the numerical solution disagrees with the exact one not just quantitatively but qualitatively. The problem is in fact much broader. If the difference scheme is not chosen judiciously, the character of the solution may be qualitatively incorrect (such as an oscillatory numerical solution vs. a rapidly decaying exact one). Further, important physical invariants (most notably energy or momentum) may not be conserved in the numerical solution, which may render the computated results nonphysical. This is important, in particular, in Molecular Dynamics, where energy conservation and, more generally, "symplecticness" of the underlying Hamiltonian system (Section 2.5) should be preserved.

With regard to stiff systems, an alternative solution strategy that does not involve difference schemes can sometimes be effective. The solution of a

linear system of ODE can be analytically expressed via matrix exponential $\exp(At)$ (see Appendix 2.10). Computing this exponential is by no means easy (many caveats are discussed in the excellent papers by C. Moler & C. Van Loan [ML78, ML03]); nevertheless the recursion relation $\exp(At) = (\exp(At/n))^n$ is helpful. The idea is that for n sufficiently large matrix At/n is "small enough" for its exponential to be computed relatively easily with sufficient accuracy; n is usually chosen as an integer power of two, so that the n-th power of the matrix can be computed by repeated squaring.

Two interesting motifs of this and the following section can now be noted:

- difference methods that ensure a qualitative/physical agreement between the numerical solutions and the exact ones;
- methods blending numerical and analytical approximations.

Many years ago, my advisor Iu.V. Rakitskii [Rak72, RUC79, RSY+85] was an active proponent of both themes. Nowadays, the qualitative similarity between discrete and continuous models is an important trend in mathematical studies and their applications. Undoubtedly, Rakitskii would have been happy to see the contribution of Yu.B. Suris, his former student, to the development of numerical methods preserving the physical invariants of Hamiltonian systems [Sur87]–[Sur96], as well as to discrete differential geometry (A.I. Bobenko & Yu.B. Suris [BSve]). Another "Rakitskii-style" development is the generalized finite-difference calculus of Flexible Local Approximation MEthods (FLAME, Chapter 4) that seamlessly incorporates local analytical approximations into difference schemes.

2.5 Schemes for Hamiltonian Systems

2.5.1 Introduction to Hamiltonian Dynamics

Note: no prior knowledge of Hamiltonian systems is necessary for reading this section.

As a starting example, consider a (classical) harmonic oscillator, such as a mass on a spring, described by the ODE

$$m\ddot{q} \;=\; -kq \tag{2.65}$$

(mass times acceleration equals force), where mass m and the spring constant k are known parameters and q is a coordinate. The general solution to this equation is

$$q(t) \;=\; q_0 \cos(\omega_0 t + \phi); \qquad \omega_0^2 \;=\; \frac{k}{m} \tag{2.66}$$

for some parameters q_0 and ϕ.

Even though the above expression in principle contains all the information about the solution, recasting the differential equation in a different form

brings a deeper perspective. The new insights are even more profound for multiparticle problems with multiple degrees of freedom.

The *Hamiltonian* of the oscillator – the energy function H expressed in terms of q and $\dot{q}$ – comprises the kinetic and potential terms:[5]

$$H = \frac{1}{2} m\dot{q}^2 + \frac{1}{2} kq^2 \tag{2.67}$$

We shall view H as a function of two variables: coordinate q and momentum $p = m\dot{q}$; in terms of these variables,

$$H(q,p) = \frac{p^2}{2m} + \frac{kq^2}{2} \tag{2.68}$$

The original second-order differential equation splits up into two first-order equations

$$\begin{cases} \dot{q} = m^{-1}p \\ \dot{p} = -kq \end{cases} \tag{2.69}$$

or in matrix-vector form

$$\dot{w} = Aw, \quad w = \begin{pmatrix} q \\ p \end{pmatrix}; \quad A = \begin{pmatrix} 0 & m^{-1}p \\ -k & 0 \end{pmatrix} \tag{2.70}$$

The right hand side of differential equations (2.69) is in fact directly related to the partial derivatives of $H(q,p)$:

$$\frac{\partial H(q,p)}{\partial p} = \frac{p}{m} \tag{2.71}$$

$$\frac{\partial H(q,p)}{\partial q} = kq \tag{2.72}$$

We thus arrive at the equations of Hamiltonian dynamics, with their elegant symmetry:

$$\begin{cases} \frac{\partial H(q,p)}{\partial p} = \dot{q} \\ \frac{\partial H(q,p)}{\partial q} = -\dot{p} \end{cases} \tag{2.73}$$

Energy conservation follows directly from these Hamiltonian equations by chain-rule differentiation:

$$\frac{\partial H}{\partial t} = \frac{\partial H}{\partial p}\dot{p} + \frac{\partial H}{\partial q}\dot{q} = \dot{q}\dot{p} - \dot{p}\dot{q} = 0$$

In the *phase plane* (q,p), constant energy levels correspond to ellipses

[5] More generally in mechanics, the Hamiltonian can be defined by its relationship with the Lagrangian of the system, and is indeed equal to the energy of the system if expressions for the generalized coordinates do not depend on time.

$$H(q,p) \;=\; \frac{p^2}{2m} \;+\; \frac{kq^2}{2} \;=\; \text{const} \tag{2.74}$$

For the Hamiltonian system, any particular solution $(q(t), p(t))$, viewed as a (moving) point in the phase plane, moves along the ellipse corresponding to the energy of the oscillator.

Further insight is gained by following the evolution of the $w = (q, p)$ points corresponding to a *collection* of oscillators (or the same oscillator observed repeatedly under different conditions). The initial coordinates and momenta of a family of oscillators are represented by a set of points in the phase plane. One may imagine that these points fill a certain geometric domain $\Omega(0)$ at $t = 0$ (shaded area in Fig. 2.12). With time, each of the points will follow its own elliptic trajectory, so that at any given moment of time t the initial domain $\Omega(0)$ will be transformed into some other domain $\Omega(t)$.

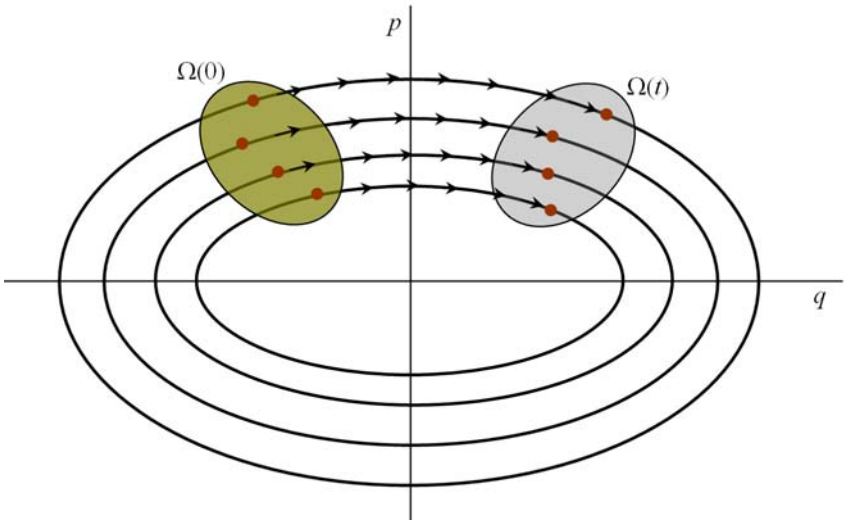


Fig. 2.12. The motion of a harmonic oscillator is represented in the (q, p) phase plane by a point moving around an ellipse. Domain $\Omega(0)$ contains a collection of such points (corresponding to an ensemble of oscillators or, equivalently, to a set of different initial conditions for one oscillator) at time $t = 0$. Domain $\Omega(t)$ contains the points corresponding to the same oscillators at some arbitrary moment of time t. The area of $\Omega(t)$ turns out not to depend on time.

By definition, it is the solutions of the Hamiltonian system that effect the mapping from $\Omega(0)$ to $\Omega(t)$. These solutions are given by matrix exponentials (see Appendix 2.10):

$$w(t) \;=\; \begin{pmatrix} q(t) \\ p(t) \end{pmatrix} \;=\; \exp(At) \begin{pmatrix} q(0) \\ p(0) \end{pmatrix} \tag{2.75}$$

The Jacobian of this mapping is the determinant of $\exp(At)$; as known from linear algebra, this determinant is equal to the product of eigenvalues $\lambda_{1,2}(\exp(At))$:

$$\det(\exp(At)) = \lambda_1(\exp(At))\,\lambda_2(\exp(At)) = \exp(\lambda_1(At))\exp(\lambda_2(At))$$
$$= \exp(\lambda_1(At) + \lambda_2(At)) = \exp(\mathrm{Tr}(At)) = 1 \tag{2.76}$$

(The eigenvalues of $\exp(At)$ are equal to the exponents of the eigenvalues of At; if this looks unfamiliar, see Appendix 2.10, p. 65).

Since the determinant of the transformation is unity, the evolution operator preserves the oriented area of $\Omega(t)$, *in addition to* energy conservation that was demonstrated earlier.

This result generalizes to higher-dimensional phase spaces in multiparticle systems. Such phase spaces comprise the generalized coordinates q_i and momenta p_i of N particles. If particle motion is three-dimensional, there are three degrees of freedom per particle[6] and hence $i = 1, 2, \ldots, 3N$; the dimension of the phase space is thus $6N$. The most direct analogy with area conservation is that the $6N$-dimensional phase volume is conserved under the evolution map [Arn89, HrW93, SSC94]. However, there is more. For any two-dimensional surface in the phase space, take its projections onto the individual phase planes (p_i, q_i) and sum up the oriented areas of these projections; this sum is conserved during the Hamiltonian evolution of the surface. Transformations that have this conservation property for the sum of the areas are called *symplectic*.

There is a very deep and elaborate mathematical theory of Hamiltonian phase flows on symplectic manifolds. A symplectic manifold is an even-dimensional differentiable manifold endowed with a closed nondegenerate differential 2-form; these notions, however, are not covered in this book. Further mathematical details are described in the monographs by V.I. Arnol'd [Arn89] and J.M. Sanz-Serna & M.P. Calvo [SSC94].

2.5.2 Symplectic Schemes for Hamiltonian Systems

This subsection gives a brief summary of FD schemes that preserve the symplectic property of Hamiltonian systems. The material comes from the paper by R.D. Skeel *et al.* [RDS97], from the results on Runge–Kutta schemes due to Yu.B. Suris [Sur87]–[Sur90] and J.M. Sanz-Serna [SSC94], and from the compendium of symplectic symmetric Runge–Kutta methods by W. Oevel & M. Sofroniou [OS97].

The governing system of ODEs in Newtonian mechanics and, in particular, molecular dynamics is

$$\ddot{r} = f(r), \quad r \in \mathbb{R}^n \tag{2.77}$$

[6] Disregarding the internal structure of particles and any degrees of freedom that may be associated with that.

where r is the position vector for a collection of n interacting particles and f is the normalized force vector (vector of forces divided by particle masses). It is assumed that the forces do not explicitly depend on time.

The simplest, and yet effective, difference scheme for this problem is known as the Störmer–Verlet method:[7]

$$\frac{r_{n+1} - 2r_n + r_{n-1}}{\Delta t^2} \;=\; f(r_n) \tag{2.78}$$

The left hand side of the Störmer scheme is a second-order (with respect to the time step Δt) approximation of $\ddot{r}$; this approximation is very common.

The velocity vector can be computed from the position vector by central differencing:

$$v_n \;=\; \frac{r_{n+1} - r_{n-1}}{2\Delta t} \tag{2.79}$$

Time-stepping for both vectors r and v simultaneously can be arranged in a "leapfrog" manner:

$$v_{n+1/2} \;=\; v_{n-1/2} \;+\; \Delta t\, f(r_n) \tag{2.80}$$

$$r_{n+1} \;=\; r_n \;+\; \Delta t\, v(n+1/2) \tag{2.81}$$

The leapfrog scheme (2.80), (2.81) is theoretically equivalent to the Störmer scheme (2.78), (2.79). The advantage of these schemes is that they are symplectic and at the same time explicit: no systems of equations need to be solved in the process of time-stepping. Several other symplectic integrators are considered by R.D. Skeel *et al.* [RDS97], but they are all implicit.

With regard to the Runge–Kutta methods, the Suris–Sanz-Serna condition of symplecticness is

$$b_i a_{ij} \;+\; b_j a_{ji} \;-\; b_i b_j \;=\; 0, \qquad i, j \;=\; 1, 2, \ldots s \tag{2.82}$$

where b_i, a_{ij} are the coefficients of an s-stage Runge–Kutta method defined on p. 21, except that here the scheme is no longer explicit – i.e. a_{ij} can be nonzero for any pair of indexes i, j.

W. Oevel & M. Sofroniou [OS97] give the following summary of symplectic Runge–Kutta schemes.

There is a unique one-stage symplectic method with the Butcher tableau

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$$

It represents the implicit scheme

$$r_{n+1} \;=\; r_n \;+\; \Delta t\, f\left(t_n + \frac{\Delta t}{2},\; \frac{1}{2}(r_n + r_{n+1})\right) \tag{2.83}$$

[7] Skeel *et al.* [RDS97] cite S. Toxvaerd's statement [Tox94] that "the first known published appearance [of this method] is due to Joseph Delambre (1791)".

The following two-stage method is also symplectic:

$$
\begin{array}{c|cc}
\frac{1}{2} \pm \frac{1}{2\sqrt{3}} & \frac{1}{4} & \frac{1}{4} \pm \frac{1}{2\sqrt{3}} \\
\frac{1}{2} \mp \frac{1}{2\sqrt{3}} & \frac{1}{4} \mp \frac{1}{2\sqrt{3}} & \frac{1}{4} \\
\hline
 & \frac{1}{2} & \frac{1}{2}
\end{array}
$$

W. Oevel & M. Sofroniou [OS97] list a number of other methods, up to six-stage ones; these methods were derived using symbolic algebra.

2.6 Schemes for One-Dimensional Boundary Value Problems

2.6.1 The Taylor Derivation

After a brief review of time-stepping schemes, we turn our attention to FD schemes for *boundary value problems*. Such schemes can be applied to various physical fields and potentials in one-dimension (this section), two and three dimensions (the following sections). The most common and straightforward way of generating FD schemes is by Taylor expansion. As the simplest example, consider the Poisson equation in 1D:

$$-\frac{d^2u}{dx^2} \;=\; f(x) \tag{2.84}$$

where $f(x)$ is a given function that in physical problems represents the distribution of sources. The minus sign in the right hand side is conventional in many physical problems (electrostatics, heat transfer, etc.).

Let us introduce a grid, for simplicity with a uniform spacing h, and consider a three-point stencil x_{k-1}, x_k, x_{k+1}, where $x_{k\pm1} = x_k \pm h$. We shall look for the difference scheme in the form

$$s_{-1}u_{k-1} \;+\; s_0u_k \;+\; s_{+1}u_{k+1} \;=\; f(x_k) \tag{2.85}$$

where the coefficients s (mnemonic for "scheme") are to be determined. These coefficients are chosen to approximate, with the highest possible order in terms of the grid size h, the Poisson equation (2.84). More specifically, let u^* be the exact solution of this equation, and let us write out the Taylor expansions of the values of u^* at the stencil nodes:

$$
\begin{aligned}
u^*_{k-1} \;&=\; u^*_k \;-\; hu^{*\prime}_k \;+\; \frac{1}{2}h^2u^{*\prime\prime}_k \;+\; \text{h.o.t.} \\
u^*_k \;&=\; u^*_k \\
u^*_{k+1} \;&=\; u^*_k \;+\; hu^{*\prime}_k \;+\; \frac{1}{2}h^2u^{*\prime\prime}_k \;+\; \text{h.o.t.}
\end{aligned}
$$

where the primes denote derivatives at the midpoint of the stencil, $x = x_k$, and "h.o.t." as before stands for "higher order terms". Substituting these Taylor expansions into the difference scheme (2.85) and collecting the powers of h, one obtains

$$(s_{-1}+s_0+s_{+1})\, u_k^* + (-s_{-1}+s_{+1})\, u^{*\prime}_k h + \frac{1}{2}(s_{-1}+s_{+1})\, u^{*\prime\prime}_k h^2 + \text{h.o.t.} \;=\; -u_k'' \tag{2.86}$$

where in the right hand side we took note of the fact that $f(x_k) = -u_k''$. The consistency error of the scheme is, by definition,

$$\begin{aligned} \epsilon_c \;&=\; (s_{-1} + s_0 + s_{+1})u_k^* \;+\; (-s_{-1} + s_{+1})\, u^{*\prime}_k h \\ &+\; \frac{1}{2}\left(s_{-1} + s_{+1} + \frac{2}{h^2}\right) u^{*\prime\prime}_k h^2 \;+\; \text{h.o.t.} \end{aligned} \tag{2.87}$$

The consistency error tends to zero as $h \to 0$ if and only if

$$\begin{aligned} s_{-1} + s_0 + s_{+1} \;&=\; 0 \\ -s_{-1} + s_{+1} \;&=\; 0 \\ s_{-1} + s_{+1} + 2/h^2 \;&=\; 0 \end{aligned}$$

from which the coefficients of the scheme are immediately found to be

$$s_{-1} \;=\; s_{+1} \;=\; -\,1/h^2; \quad s_0 = 2/h^2 \tag{2.88}$$

and the difference equation thus reads

$$\frac{-u_{k-1} \;+\; 2u_k \;-\; u_{k+1}}{h^2} \;=\; f(x_k) \tag{2.89}$$

It is easy to verify that this scheme is of second order with respect to h, i.e. its consistency error $\epsilon_c = \mathcal{O}(h^2)$. The Taylor analysis leading to this scheme is general, however, and can be extended to generate higher-order schemes, provided that the grid stencil is extended as well. As an exercise, the reader may verify that on a 5-point stencil of a uniform grid the scheme with coefficients $[1, -16, 30, -16, 1]/(12h^2)$ is of order four.

Practical implementation of FD schemes involves forming a system of equations for the nodal values of function u, imposing the boundary conditions, solving this system and processing the results. The implementation is described in Section 2.6.4.

2.6.2 Using Constraints to Derive Difference Schemes

In this subsection, a slightly different way of deriving difference schemes is presented. The idea is most easily illustrated in 1D but will prove to be fruitful in 2D and 3D, particularly for the development of the so-called "Mehrstellen" schemes (see Sections 2.7.4, 2.8.5).

For the 1D Poisson equation, we are looking for a three-point FD scheme of the form

$$s_{-1}u_{k-1} \;+\; s_0 u_k \;+\; s_{+1}u_{k+1} \;=\; s_f \tag{2.90}$$

Parameter s_f in the right hand side is not specified *a priori* and will be determined, along with $s_{\pm 1}$ and s_0, as a result of a formal procedure described below.

Let us again expand the exact solution u into the Taylor series around the midpoint x_k of the stencil:

$$u(x) \;=\; c_0 \;+\; c_1(x-x_k) \;+\; c_2(x-x_k)^2 \;+\; c_3(x-x_k)^3 \;+\; c_4(x-x_k)^4 \;+\; \text{h.o.t.} \tag{2.91}$$

The coefficients c_α are of course directly related to the derivatives of u at x_k but will initially be treated as undetermined parameters; later on, information available about them will be taken into account.

Consistency error of scheme (2.90) can be evaluated by substituting the Taylor expansion (2.91) into the scheme. Upon collecting similar terms for all coefficients c_α, we get

$$\begin{aligned}\epsilon_c \;=\; & -s_f \;+\; (s_{-1}+s_0+s_{+1})c_0 \;+\; (-s_{-1}+s_{+1})hc_1 \;+\; (s_{-1}+s_{+1})h^2c_2 \\ & +\; (-s_{-1}+s_{+1})h^3c_3 \;+\; (s_{-1}+s_{+1})h^4c_4 \;+\; \text{h.o.t.}\end{aligned} \tag{2.92}$$

If no information about the coefficients c_α were available, the best one could do to minimize the consistency error would be to set $s_f = 0$, $s_{-1}+s_0+s_{+1} = 0$, and $-s_{-1}+s_{+1} = 0$, which yields $u_{k-1} - 2u_k + u_{k+1} = 0$.

Not surprisingly, this scheme is not suitable for the Poisson equation with a nonzero right hand side: we have not yet made use of the fact that u satisfies this equation – that is, that the Taylor coefficients c_α are not arbitrary. In particular,

$$u''(x_k) \;=\; 2c_2 \;=\; -f(x_k) \tag{2.93}$$

This condition can be taken into account by using an idea that is, in a sense, dual to the method of Lagrange multipliers in constrained optimization. (Here we are in fact dealing with a special optimization problem – namely, minimization of the consistency error in the asymptotic sense.) In typical constrained optimization, restrictions are imposed on the *optimization parameters* being sought; in our case, these parameters are the coefficients s of the difference scheme. Note that constraints on optimization parameters, generally speaking, *inhibit* optimization.

In contrast, in our case the constraint applies to the parameters of the function being minimized. This narrows down the set of target functions and *facilitates* optimization. To incorporate the constraint on c_2 (2.93) into the minimization problem, one can introduce an analog of the Lagrange multiplier λ:

$$\epsilon_c \;=\; -s_f \;+\; (s_{-1}+s_0+s_{+1})c_0 \;+\; (-s_{-1}+s_{+1})hc_1 \;+\; (s_{-1}+s_{+1})h^2c_2$$

$$+ \ (-s_{-1} + s_{+1})h^3 c_3 \ + \ (s_{-1} + s_{+1})h^4 c_4 \ + \ \text{h.o.t.} \ - \ \lambda[2c_2 + f(x_k)]$$

or equivalently

$$\epsilon_c \ = \ (-s_f - \lambda f(x_k)) \ + \ (s_{-1} + s_0 + s_{+1})c_0 \ + \ (-s_{-1} + s_{+1})hc_1$$

$$+ \ (s_{-1}h^2 + s_{+1}h^2 - 2\lambda)c_2 \ + \ (-s_{-1} + s_{+1})h^3 c_3 \ + \ (s_{-1} + s_{+1})h^4 c_4 \ + \ \text{h.o.t.} \tag{2.94}$$

where λ is an arbitrary parameter that one is free to choose *in addition* to the coefficients of the scheme. As Sections 2.7.4 and 2.8.5 show, in 2D and 3D there are several such constraints and therefore several extra free parameters at our disposal.

Maximization of the order of the consistency error (2.94) yields the following conditions:

$$\begin{aligned} -s_f - \lambda f(x_k) \ &= \ 0 \\ s_{-1} + s_0 + s_{+1} \ &= \ 0 \\ -s_{-1} + s_{+1} \ &= \ 0 \\ s_{-1}h^2 + s_{+1}h^2 - 2\lambda \ &= \ 0 \end{aligned}$$

This gives, up to an arbitrary factor, $\lambda = 1$, $s_{\pm 1} = h^{-2}$, $s_0 = -2h^{-2}$, $s_f = -f(x_k)$, and the resultant difference scheme is

$$\frac{-u_{k-1} \ + \ 2u_k \ - \ u_{k+1}}{h^2} \ = \ f(x_k) \tag{2.95}$$

This new "Lagrange-like" derivation produces a well-known scheme in one dimension, but in 2D/3D the idea will prove to be more fruitful and will lead to "Mehrstellen" schemes introduced by L. Collatz [Col66].

2.6.3 Flux-Balance Schemes

The previous analysis was implicitly based on the assumption that the exact solution was sufficiently smooth to admit the Taylor approximation to a desired order. However, Taylor expansion typically breaks down in a number of important practical cases – particularly so in the vicinity of material interfaces. In 1D, this is exemplified by the following problem:

$$-\frac{d}{dx}\left(\lambda(x)\,\frac{du}{dx}\right) \ = \ f(x) \ \text{ on } \Omega \equiv [a,\, b], \quad u(a) \ = \ u_a, \ \ u(b) \ = \ u_b \tag{2.96}$$

where the boundary values u_a, u_b are given. In this equation, λ is the material parameter whose physical meaning varies depending on the problem: it is thermal conductivity in heat transfer, dielectric permittivity in electrostatics, magnetic permeability in magnetostatics (if the magnetic scalar potential is used), and so on. This parameter is usually discontinuous across interfaces of

different materials. In such cases, the solution satisfies the interface boundary conditions that in the 1D case are

$$u(x_0^-) \;=\; u(x_0^+); \qquad \lambda(x_0^-)\,\frac{du(x_0^-)}{dx} \;=\; \lambda(x_0^+)\,\frac{du(x_0^+)}{dx} \tag{2.97}$$

where x_0 is the discontinuity point for $\lambda(x)$, and the $-$ and $+$ labels correspond to the values immediately to the left and to the right of x_0, respectively.

The quantities $-\lambda(x)du/dx$ typically have the physical meaning of *fluxes*: for example, the heat flux (i.e. energy passed through point x per unit time) in heat transfer problems or the flux of charges (that is, electric current) in electric conduction, etc. The fundamental physical principle of energy or flux conservation can be employed to construct a difference scheme. For any chosen subdomain (often called "control volume" – in 1D, a segment), the outgoing energy flow (e.g. heat flux) is equal to the total capacity of sources (e.g. heat sources) within that subdomain. In electro- or magnetostatics, with the electric or magnetic scalar potential formulation, a similar principle of *flux balance* is used instead of energy balance.

For equation (2.96) energy or flux balance can mathematically be derived by integration. Indeed, let $\omega = [\alpha, \beta] \subset \Omega$.[8] Integrating the underlying equation (2.96) over ω, we obtain

$$\lambda(\alpha)\,\frac{du}{dx}(\alpha) \;-\; \lambda(\beta)\,\frac{du}{dx}(\beta) \;=\; \int_\alpha^\beta f(x)\,dx \tag{2.98}$$

which from the physical point of view is exactly the flux balance equation (outgoing flux from ω is equal to the total capacity of sources inside ω).

Fig. 2.13 illustrates the construction of the flux-balance scheme; α and β are chosen as the midpoints of intervals $[x_{k-1}, x_k]$ and $[x_k, x_{k+1}]$, respectively. The fluxes in the left hand side of the balance equation (2.98) are approximated by finite differences to yield

$$h^{-1}\left(\lambda(\alpha)\frac{u_k - u_{k-1}}{h} \;-\; \lambda(\beta)\frac{u_{k+1} - u_k}{h}\right) \;=\; h^{-1}\int_\alpha^\beta f(x)\,dx \tag{2.99}$$

If the central point x_k of the stencil is placed at the material discontinuity (as shown in Fig. 2.13), $\lambda(\alpha) \equiv \lambda_-$ and $\lambda(\beta) \equiv \lambda_+$. The factor h^{-1} is introduced to normalize the right hand side of this scheme to $\mathcal{O}(1)$ with respect to the mesh size (i.e. to keep the magnitude of the right hand side approximately constant as the mesh size decreases). The integral in the right hand side can be computed either analytically, if $f(x)$ admits that, or by some numerical quadrature – the simplest one being just $f(x_k)(\beta - \alpha)$. This flux-balance scheme has a solid foundation as a discrete energy conservation condition. From the mathematical viewpoint, this translates into favorable properties of the algebraic system of equations (to be considered in Section 2.6.4): matrix symmetry and, as a consequence, the discrete reciprocity principle.

[8] While symbol Ω refers to the whole computational domain, ω denotes its subdomain (typically "small" in some sense).

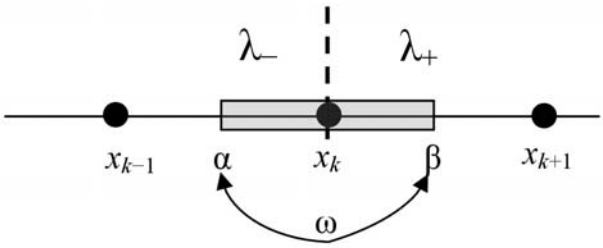


Fig. 2.13. A three-point flux balance scheme near a material interface in one dimension.

If the middle node of the stencil is *not* located exactly at the material boundary, the flux-balance scheme (2.99) is still usable, with $\lambda(\alpha)$ and $\lambda(\beta)$ being the values of λ in the material where the respective point α or β happens to lie. However, numerical accuracy deteriorates significantly. This can be shown analytically by substituting the exact solution into the flux-balance scheme and evaluating the consistency error.

Rather than performing this algebraic exercise, we simply consider a numerical illustration. Problem (2.96) is solved in the interval $[0, 1]$. The material boundary point is chosen to be an irrational number $a = 1/\sqrt{2}$, so that in the course of the numerical experiment it does not coincide with a grid node of any uniform grid. There are no sources (i.e. $f = 0$) and the Dirichlet conditions are $u(0) = 0$, $u(1) = 1$. The exact solution and the numerical solution with 10 grid nodes are shown in Fig. 2.14. The log-log plot of the relative error norm of the numerical solution vs. the number of grid nodes is given in Fig. 2.15. The dashed line in the figure is drawn for reference to identify the $\mathcal{O}(h)$ slope.

Comparison with this reference line reveals that the convergence rate is only $\mathcal{O}(h)$. Were the discontinuity point to coincide with a grid node, the scheme could easily be shown to be *exact* – in practice, the numerical solution would be obtained with machine precision. The farther the discontinuity point is from the nearest grid node (relative to the grid size), the higher the numerical error tends to be. This relative distance to the nearest node is plotted in Fig. 2.16 and does indeed correlate clearly with the numerical error in Fig. 2.15.

As in the case of Taylor-based schemes of the previous section, the flux-balance schemes prove to be a very natural particular case of "Trefftz–FLAME" schemes considered in Chapter 4; see in particular Section 4.4.2. Moreover, in contrast with standard schemes, in FLAME the location of material discontinuities relative to the grid nodes is almost irrelevant.

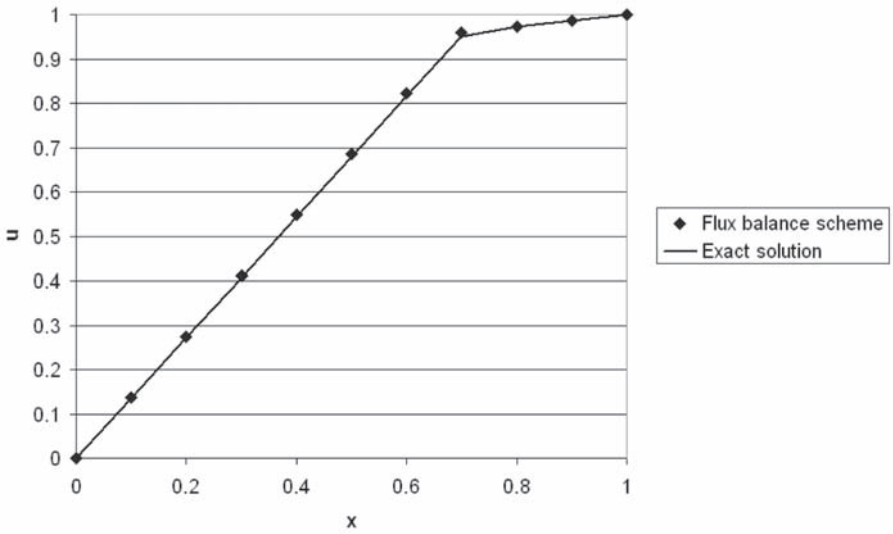


Fig. 2.14. Solution of the 1D problem with material discontinuity. $\lambda_- = 1$, $\lambda_+ = 10$.

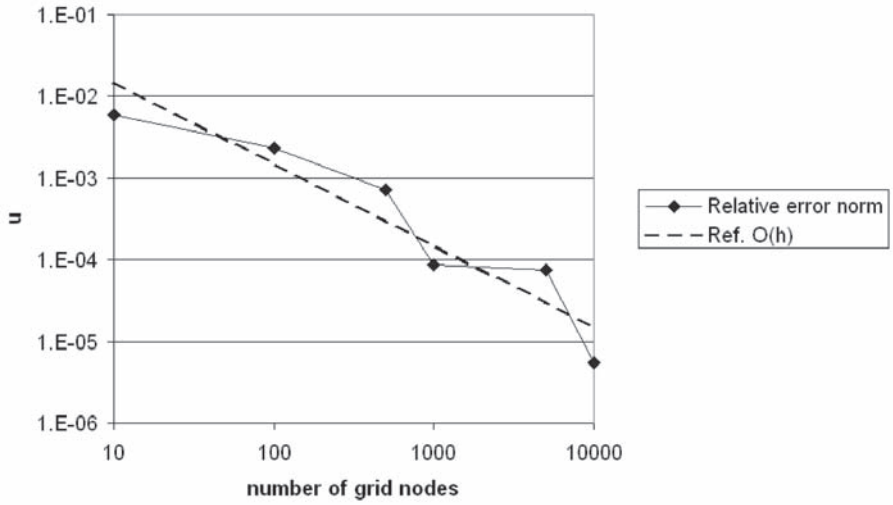


Fig. 2.15. Flux-balance scheme: errors vs. the number of grid points for the 1D problem with material discontinuity. $\lambda_- = 1$, $\lambda_+ = 10$.

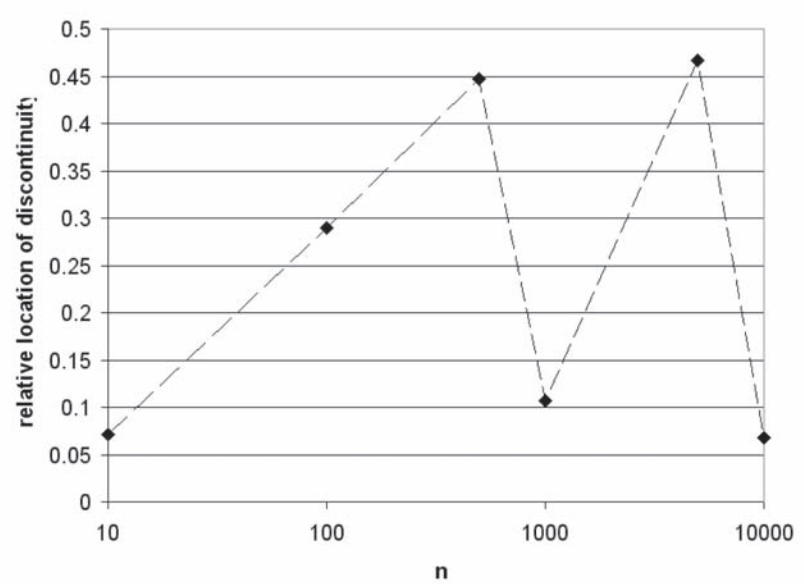


Fig. 2.16. Relative distance (as a fraction of the grid size) between the discontinuity point and the nearest grid node.

2.6.4 Implementation of 1D Schemes for Boundary Value Problems

Difference schemes like (2.89) or (2.99) constitute a *local* relationship between the values at the neighboring nodes of a particular stencil. Putting these local relationships together, one obtains a global system of equations.

With the grid nodes numbered consecutively from 1 to n,[9] the $n \times n$ matrix of this system is tridiagonal. Indeed, row k of this matrix corresponds to the difference equation – in our case, either (2.89) or (2.99) – that connects the unknown values of u at nodes $k-1$, k and $k+1$.

For example, the flux-balance scheme (2.99) leads to a matrix L with diagonal entries $L_{kk} = (\lambda^+ + \lambda^-)/h$ and the off-diagonal ones $L_{k-1,k} = -\lambda^-/h$, $L_{k,k+1} = -\lambda^+/h$, where as before λ^- and λ^+ are the values of material parameter λ at the midpoints of intervals $[x_{k-1}, x_k]$ and $[x_k, x_{k+1}]$, respectively.

These entries are modified at the end points of the interval to reflect the Dirichlet boundary conditions.[10] At the boundary nodes, the Dirichlet condition can be conveniently enforced by setting the corresponding diagonal

[9] Numbering from 0 to $n-1$ is often more convenient, and is the default in languages like C/C++. However, I have adopted the default numbering of Matlab and of the classic versions of FORTRAN.

[10] The implementation of Neumann and other boundary conditions is covered in all textbooks on FD schemes: L. Collatz [Col66], A.A. Samarskii [Sam01], J.C. Strikwerda [Str04], W.E. Milne [Mil70], and many others.

matrix entry to one, the other entries in its row to zero, and the respective entry in the right hand side to the given Dirichlet value of the solution.

In addition, if j is a Dirichlet boundary node and i is its neighbor, the $L_{ij}u_j$ term in the i-th difference equation is known and therefore gets moved (with the opposite sign) to the right hand side, while the (i, j) matrix entry is simultaneously set to zero. The same procedure is valid in two and three dimensions, except that in these cases a boundary node can have several neighbors.[11]

The system matrix L corresponding to this three-point scheme is tridiagonal, and the system can be easily solved by Gaussian elimination (A. George & J.W-H. Liu [GL81]) or its modifications (S.K. Godunov & V.S. Ryabenkii [GR87a]).

2.7 Schemes for Two-Dimensional Boundary Value Problems

2.7.1 Schemes Based on the Taylor Expansion

For illustration, let us again turn to the Poisson equation – this time in two dimensions:

$$-\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = f(x, y) \tag{2.100}$$

We introduce a Cartesian grid with grid sizes h_x, h_y and the number of grid subdivisions N_x, N_y in the x- and y-directions, respectively. To keep the notation simple, we consider the grid to be uniform along each axis; more generally, h_x could vary along the x-axis and h_y could vary along the y-axis, but the essence of the analysis would remain the same. Each node of the grid can be characterized in a natural way by two integer indices n_x and n_y corresponding to the x- and y-directions; $1 \le n_x \le N_x + 1$, $1 \le n_y \le N_y + 1$.

To generate a Taylor-based difference scheme for the Poisson equation (2.100), it is natural to approximate the x- and y- partial derivatives separately in exactly the same way as done in 1D. The resulting scheme for grid nodes not adjacent to the domain boundary is

$$\frac{-u_{n_x-1,n_y} + 2u_{n_x,n_y} - u_{n_x+1,n_y}}{h_x^2} + \frac{-u_{n_x,n_y-1} + 2u_{n_x,n_y} - u_{n_x,n_y+1}}{h_y^2} = f(x_n, y_n) \tag{2.101}$$

where x_n, y_n are the coordinates of the grid node (n_x, n_y). Note that difference scheme (2.101) involves the values of u on a 5-point grid stencil (three points in each coordinate direction, with the middle node shared, Fig. 2.17). As in

[11] The same is true in 1D for higher order schemes with more than three stencil nodes in the interior of the domain (more than two nodes in boundary stencils).

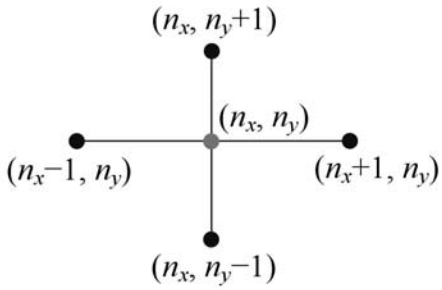


Fig. 2.17. A 5-point stencil for difference scheme (2.101) in 2D.

1D, scheme (2.101) is of second order, i.e. its consistency error is $\mathcal{O}(h^2)$, where $h = \max(h_x, h_y)$. By expanding the stencil, it is possible – again by complete analogy with the 1D case – to increase the order of the scheme. For example, on the stencil with nine nodes (five in each coordinate direction, with the middle node shared) a fourth order scheme can be obtained by combining two fourth order schemes in the x- and y-directions on their respective 5-point stencils. Other stencils can be used to construct higher-order schemes, and other ideas can be applied to this construction (see for example the Collatz "Mehrstellen" schemes on a 3×3 stencil in Section 2.7.4).

2.7.2 Flux-Balance Schemes

Let us now turn our attention to a more general 2D problem with a varying material parameter ϵ

$$-\nabla \cdot (\epsilon(x, y)\nabla u) \;=\; f(x, y) \tag{2.102}$$

where ϵ may depend on coordinates but not – in the linear case under consideration – on the solution u. Moreover, ϵ will be assumed piecewise smooth, with possible discontinuities only at material boundaries.[12]

At any material interface boundary, the following conditions hold:

$$\epsilon^- \frac{\partial u^-}{\partial n} \;=\; \epsilon^+ \frac{\partial u^+}{\partial n} \tag{2.103}$$

where "$-$" and "$+$" refer to the values on the two sides of the interface boundary and n is the normal to the boundary in a prescribed direction.

The integral form of the differential equation (2.102) is, by Gauss's Theorem,

[12] Throughout the book, "smoothness" is not characterized in a mathematically precise way. Rather, it is tacitly assumed that the level of smoothness is sufficient to justify all mathematical operations and analysis.

$$-\int_{\gamma} \epsilon(x,y)\,\frac{\partial u}{\partial n}\,d\gamma \;=\; \int_{\omega} f(x,y)\,d\omega \tag{2.104}$$

where ω is a subdomain of the computational domain Ω, γ is the boundary of ω, and n is the outward normal to that boundary.

The physical meaning of this integral equation is either energy conservation or flux balance, depending on the application. For example, in heat transfer this equation expresses the fact that the net flow of heat through the surface of volume ω is equal to the total amount of heat generated inside the volume by sources f. In electrostatics, (2.104) is an expression of Gauss's Law (the flux of the displacement vector $\mathbf{D}$ is equal to the total charge inside the volume).

The integral conservation principle (2.104) is valid for any subdomain ω. Flux-balance difference schemes are generated by applying this principle to a discrete set of subdomains ("control volumes") such as the shaded rectangle shown in Fig. 2.18. The grid nodes involved in the construction of the scheme are the same as in Fig. 2.17 and are not labeled to avoid overloading the picture. For this rectangular control volume, the surface flux integral in the

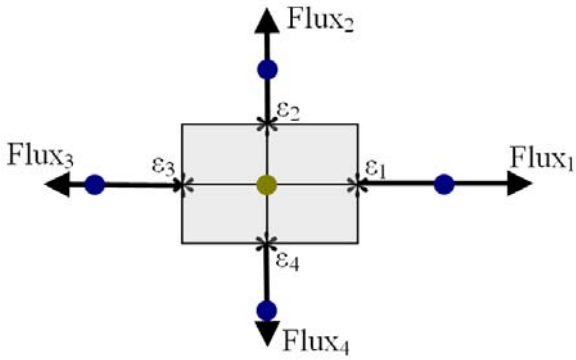


Fig. 2.18. Construction of the flux-balance scheme. The net flux out of the shaded control volume is equal to the total capacity of sources inside that volume.

balance equation (2.104) splits up into four fluxes through the edges of the rectangle. Each of these fluxes can be approximated by a finite difference; for example,

$$\text{Flux}_1 \;\approx\; \epsilon_1 h_y\,\frac{u_{n_x,n_y} - u_{n_x+1,n_y}}{h_x} \tag{2.105}$$

where ϵ_1 is the value of the material parameter at the edge midpoint marked with an asterisk in Fig. 2.18; the h_y factor is the length of the right edge of the shaded rectangle. (If the grid were not uniform, this edge length would be the average value of the two consecutive grid sizes.)

The complete difference scheme is obtained by summing up all four edge fluxes:

$$\epsilon_1 h_y \frac{u_{n_x,n_y} - u_{n_x+1,n_y}}{h_x} + \epsilon_2 h_x \frac{u_{n_x,n_y} - u_{n_x,n_y+1}}{h_y}$$
$$+ \epsilon_3 h_y \frac{u_{n_x,n_y} - u_{n_x-1,n_y}}{h_x} + \epsilon_4 h_x \frac{u_{n_x,n_y} - u_{n_x,n_y-1}}{h_y} = f(x_n, y_n)\, h_x h_y$$

The approximation of fluxes by finite differences hinges on the assumption of smoothness of the solution. At material interfaces, this assumption is violated, and accuracy deteriorates. The reason is that the Taylor expansion fails when the solution or its derivatives are discontinuous across boundaries. One can try to remedy that by generalizing the Taylor expansion and accounting for derivative jumps (A. Wiegmann & K.P. Bube [WB00]); however, this approach leads to unwieldy expressions. Another alternative is to replace the Taylor expansion with a linear combination of suitable basis functions that satisfy the discontinuous boundary conditions and therefore approximate the solution much more accurately. This idea is taken full advantage of in FLAME (Chapter 4).

2.7.3 Implementation of 2D Schemes

By applying a difference scheme on all suitable grid stencils, one obtains a system of equations relating the nodal values of the solution on the grid. To write this system in matrix form, one needs a *global* numbering of nodes from 1 to N, where $N = (N_x + 1)(N_y + 1)$. The numbering scheme is in principle arbitrary, but the most natural order is either row-wise or column-wise along the grid. In particular, for row-wise numbering, node (n_x, n_y) has the global number

$$n = (N_x + 1)(n_y - 1) + n_x - 1, \qquad 1 \le n \le N \tag{2.106}$$

With this numbering scheme, the global node numbers of the two neighbors of node $n = (n_x, n_y)$ in the same row are $n - 1$ and $n + 1$, while the two neighbors in the same column have global numbers $n + (N_x + 1)$ and $n - (N_x + 1)$, respectively. For nodes adjacent to the domain boundary, fictitious "neighbors" with node numbers that are nonpositive or greater than N are ignored.

It is then easy to observe that the 5-point stencil of the difference scheme leads to a five-diagonal system matrix, two of the subdiagonals corresponding to node–node connections in the same row, and the other two to connections in the same column. All other matrix entries are zero.

The Dirichlet boundary conditions are handled in a way similar to the 1D case. Namely, for a boundary node, the corresponding diagonal entry of the system matrix can be set to one (the other entries in the same row being zero), and the entry of the right hand side set to the required Dirichlet value. Moreover, if j is a boundary node and i is its non-boundary neighbor, the term $L_{ij}u_j$ in the difference scheme is known and is therefore moved to the right hand side (with the respective matrix entry (i, j) reset to zero).

There is a rich selection of computational methods for solving such linear systems of equations with large sparse matrices. Broadly speaking, these methods can be subdivided into direct and iterative solvers. *Direct solvers* are typically based on variants of Gaussian or Cholesky decomposition, with node renumbering and possibly block partitioning; see A. George & J.W-H. Liu [GL81, GLe] and Section 3.11 on p. 129. The second one is iterative methods – variants of conjugate gradient or more general Krylov-subspace iterations with preconditioners (R.S. Varga [Var00], Y. Saad [Saa03], D.K. Faddeev & V.N. Faddeeva [FF63], H.A. van der Vorst [vdV03a]) or, alternatively, domain decomposition and multigrid techniques (W. Hackbusch [Hac85], J. Xu [Xu92], A. Quarteroni & A. Valli [QV99]); see also Section 3.13.4.

2.7.4 The Collatz "Mehrstellen" Schemes in 2D

For the Poisson equation in 2D

$$-\nabla^2 u \;=\; f \tag{2.107}$$

consider now a 9-point grid stencil of 3×3 neighboring nodes. The node numbering is shown in Fig. 2.19.

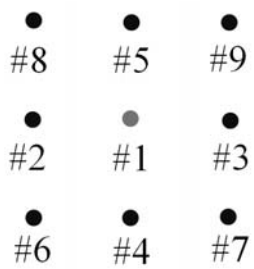


Fig. 2.19. The 9-point stencil with the local numbering of nodes as shown. The central node is numbered first, followed by the remaining nodes of the standard 5-point stencil, and then by the four corner nodes.

We set out to find a scheme

$$\sum\nolimits_{\alpha=1}^{9} s_\alpha u_\alpha \;=\; \sum\nolimits_{\alpha=1}^{9} w_\alpha f_\alpha \tag{2.108}$$

with coefficients $\{s_\alpha\}$, $\{w_\alpha\}$ ($\alpha = 1,2, \ldots, 9$) such that the consistency error has the highest order with respect to the mesh size. For simplicity, we shall now consider schemes with only one nonzero coefficient w corresponding to the central node (node #1) of the stencil. It is clear that w_1 in this case can be set to unity without any loss of generality, as the coefficients s still remain undetermined; thus

$$\sum_{\alpha=1}^{9} s_\alpha u_\alpha \;=\; f_1 \tag{2.109}$$

The consistency error of this scheme is, by definition,

$$\epsilon_c \;=\; \sum_{\alpha=1}^{9} s_\alpha u_\alpha^* \;-\; f_1 \;=\; \sum_{\alpha=1}^{9} s_\alpha u_\alpha^* \;+\; \nabla^2 u_1^* \tag{2.110}$$

where u^* is the exact solution of the Poisson equation and u_α^* is its value at node α. The goal is to minimize the consistency error in the asymptotic sense – i.e. to maximize its order with respect to h – by the optimal choice of the coefficients s_α of the difference scheme.

Suppose first that no *additional* information about u^* – other than it is a smooth function – is taken into consideration while evaluating consistency error (2.110). Then, expanding u^* into the Taylor series around the central point of the 9-point stencil, after straightforward algebra one concludes that only a second order scheme can be obtained – that is, asymptotically the same accuracy level as for the *five*-point stencil.

However, a scheme with higher accuracy can be constructed if additional information about u^* is taken into account. To fix ideas, let us consider the Laplace (rather than the Poisson) equation

$$\nabla^2 u^* \;=\; 0 \tag{2.111}$$

Differentiation of the Laplace equation with respect to x and y yields a few additional pieces of information:

$$\frac{\partial^3 u^*}{\partial x^3} \;+\; \frac{\partial^3 u^*}{\partial x^2 \partial y} \;=\; 0 \tag{2.112}$$

$$\frac{\partial^3 u^*}{\partial x \partial y^2} \;+\; \frac{\partial^3 u^*}{\partial y^3} \;=\; 0 \tag{2.113}$$

Another three equations of the same kind can be obtained by taking *second* derivatives of the Laplace equation, with respect to xx, xy, and yy. As the way these equations are produced is obvious, they are not explicitly written here to save space.

All these additional conditions on u^* impose constraints on the Taylor expansion of u^*. It is quite reasonable to seek a more accurate difference scheme if only *one* function (namely, u^*) is targeted, rather than a whole class of sufficiently smooth functions.

More specifically, let

$$\begin{aligned} u^*(x,y) \;=\; & c_0 \;+\; c_1 x \;+\; c_2 x^2 \;+\; c_3 x^3 \;+\; c_4 x^4 \;+\; c_5 y \\ & +\; c_6 xy \;+\; c_7 x^2 y \;+\; c_8 x^3 y \;+\; c_9 y^2 \;+\; c_{10} xy^2 \\ & +\; c_{11} x^2 y^2 \;+\; c_{12} y^3 \;+\; c_{13} xy^3 \;+\; c_{14} y^4 \;+\; \text{h.o.t.} \end{aligned} \tag{2.114}$$

where c_α ($\alpha = 1, 2, \ldots, 14$) are some coefficients (directly related, of course, to the partial derivatives of u^*). For convenience, the origin of the coordinate system has been moved to the midpoint of the 9-point stencil.

To evaluate and minimize the consistency error (2.110) of the difference scheme, we need the nodal values of the exact solution u^*. To this end, let us first rewrite expansion (2.114) in a more compact matrix-vector form:

$$u^*(x,y) \;=\; \underline{p}^T \underline{c} \tag{2.115}$$

where $\underline{p}^T$ is a row vector of 15 polynomials in x, y in the order of their appearance in expansion (2.114): $\underline{p}^T = [1, x, x^2, \ldots, xy^3, y^4]$; $\underline{c} \in \mathbb{R}^{15}$ is a column vector of expansion coefficients. The vector of nodal values of u^* on the stencil will be denoted with $\mathcal{N}u^*$ and is equal to

$$\mathcal{N}u^* \;=\; N\underline{c} \;+\; \text{h.o.t.} \tag{2.116}$$

The 9×15 matrix N comprises the 9 nodal values of the 15 polynomials on the stencil, i.e.

$$N_{\alpha\beta} \;=\; p_\beta(x_\alpha, y_\alpha) \tag{2.117}$$

Such matrices of nodal values will play a central role in the "Flexible Local Approximation MEthod" (FLAME) of Chapter 4.

Consistency error (2.110) *for the Laplace equation* then becomes

$$\epsilon_c^{\text{Laplace}} \;=\; \underline{s}^T N\underline{c} \;+\; \text{h.o.t.} \tag{2.118}$$

where $\underline{s} \in \mathbb{R}^9$ is a Euclidean vector of coefficients. If no information about the expansion coefficients c (i.e. about the partial derivatives of the solution) were available, the consistency error would have to be minimized for *all* vectors $\underline{c} \in \mathbb{R}^{15}$. In fact, however, u^* satisfies the Laplace equation, which imposes constraints on its second-order and higher-order derivatives. Therefore the target space for optimization is actually *narrower* than the full $\mathbb{R}^{15}$. If more constraints on the c coefficients are taken into account, higher accuracy of the difference scheme can be expected.

A "Lagrange-like" procedure (Section 2.6.2) for incorporating the constraints on u^* is in some sense dual to the standard technique of Lagrange multipliers: these multipliers are applied not to the optimization parameters but rather to the parameters of the target function u^*. Thus, we introduce five Lagrange-like multipliers λ_{1-5} to take into account five constraints on the c coefficients:

$$\begin{aligned}\epsilon_c^{\text{Laplace}} \;=\; & \underline{s}^T N\underline{c} \;-\; \lambda_1(c_2 + c_9) \;-\; \lambda_2(3c_3 + c_{10}) \;-\; \lambda_3(c_7 + 3c_{12}) \\ & -\; \lambda_4(6c_4 + c_{11}) \;-\; \lambda_5(6c_{14} + c_{11}) \;-\; \lambda_6(6c_8 + c_{13}) \;+\; \text{h.o.t.}\end{aligned} \tag{2.119}$$

For example, the constraint represented by λ_1 is just the Laplace equation itself (since $c_2 = \frac{1}{2}\frac{\partial^2 u^*}{\partial x^2}$, $c_9 = \frac{1}{2}\frac{\partial^2 u^*}{\partial y^2}$); the constraint represented by λ_2 is the derivative of the Laplace equation with respect to x (see (2.112)), and so on.

In matrix form, equation (2.119) becomes

$$\epsilon_c^{\text{Laplace}} \;=\; \underline{s}^T N\underline{c} \;-\; \underline{\lambda}^T Q\underline{c} \;+\; \text{h.o.t.} \tag{2.120}$$

where matrix Q corresponds to the λ-terms in (2.119). The same relationship can be rewritten in the block-matrix form

$$\epsilon_c \;=\; \left(\underline{s}^T\ \underline{\lambda}^T\right)\begin{pmatrix} N \\ -Q \end{pmatrix}\underline{c} \;+\; \text{h.o.t.} \tag{2.121}$$

As in the regular technique of Lagrange multipliers, the problem is now treated as *unconstrained*. The consistency error is reduced just to the higher order terms if

$$\begin{pmatrix} \underline{s} \\ \underline{\lambda} \end{pmatrix} \;\in\; \text{Null}\left(N^T; -Q^T\right) \tag{2.122}$$

assuming that this null space is nonempty.

The computation of matrices N and Q, as well as the null space above, is straightforward by symbolic algebra. As a result, the following coefficients are obtained for a stencil with mesh sizes $h_x = q_x h$, $h_y = q_y h$ in the x- and y-directions, respectively:

$$\begin{aligned} s_1 &= 20h^{-2} \\ s_{2,3} &= -2h^{-2}(5q_x^2 - q_y^2)/(q_y^2 + q_x^2) \\ s_{4,5} &= -2h^{-2}(5q_y^2 - q_x^2)/(q_y^2 + q_x^2) \\ s_{6-9} &= -h^{-2} \end{aligned}$$

.

If $q_x = q_y$ (i.e. $h_x = h_y$), the scheme simplifies:

$$\underline{s} \;=\; h^{-2}[20, -4, -4, -4, -4, -1, -1, -1, -1]$$

(20 corresponds to the central node, the -4's – to the mid-edge nodes, and the -1's – to the corner nodes).

This scheme was derived, from different considerations, by L. Collatz in the 1950's [Col66] and called a "Mehrstellenverfahren" scheme.[13] (See also A.A. Samarskii [Sam01] for yet another derivation.) It can be verified that this scheme is of order four in general but of order 6 in the special case of $h_x = h_y$. It will become clear in Sections 4.4.4 and 4.4.5 (pp. 209, 210) that the "Mehrstellen" schemes are a natural particular case of Flexible Local Approximation MEthods (FLAME) considered in Chapter 4.

More details about the "Mehrstellen" schemes and their application to the *Poisson* equation in 2D and 3D can be found in the same monographs by Collatz and Samarskii. The 3D case is also considered in Section 2.8.5, as it has important applications to long-range electrostatic forces in molecular dynamics (e.g. C. Sagui & T. Darden [SD99]) and in electronic structure calculation (E.L. Briggs *et al.* [BSB96]).

[13] In the English translation of the Collatz book, these methods are called "Hermitian".

2.8 Schemes for Three-Dimensional Problems

2.8.1 An Overview

The structure and subject matter of this section are very similar to those of the previous section on 2D schemes. To avoid unnecessary repetition, issues that are completely analogous in 2D and 3D will be reviewed briefly, but the differences between the 3D and 2D cases will be highlighted.

We again start with low-order Taylor-based schemes and then proceed to higher-order schemes, control volume/flux-balance schemes, and "Mehrstellen" schemes.

2.8.2 Schemes Based on the Taylor Expansion in 3D

The Poisson equation in 3D has the form

$$-\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}\right) = f(x, y, z) \tag{2.123}$$

Finite difference schemes can again be constructed on a Cartesian grid with the grid sizes h_x, h_y, h_z and the number of grid subdivisions N_x, N_y, N_z in the x-, y- and $z-$directions, respectively. Each node of the grid is characterized by three integer indices n_x n_y, n_z: $1 \le n_x \le N_x + 1$, $1 \le n_y \le N_y + 1$, $1 \le n_z \le N_z + 1$.

The simplest Taylor-based difference scheme for the Poisson equation is constructed by combining the approximations of the x-, y- and $z-$ partial derivatives:

$$\begin{aligned} &\frac{-u_{n_x-1,n_y,n_z} + 2u_{n_x,n_y,n_z} - u_{n_x+1,n_y,n_z}}{h_x^2}) \\ &+ \frac{-u_{n_x,n_y-1,n_z} + 2u_{n_x,n_y,n_z} - u_{n_x,n_y+1,n_z}}{h_y^2} \\ &+ \frac{-u_{n_x,n_y,n_z-1} + 2u_{n_x,n_y,n_z} - u_{n_x,n_y,n_z+1}}{h_z^2} = f(x_n, y_n, z_n) \end{aligned} \tag{2.124}$$

where x_n, y_n, z_n are the coordinates of the grid node (n_x, n_y, n_z). This difference scheme involves a *7-point* grid stencil (three points in each coordinate direction, with the middle node shared between them).

As in 1D and 2D, scheme (2.124) is of second order, i.e. its consistency error is $\mathcal{O}(h^2)$, where $h = \max(h_x, h_y, h_z)$. Higher-order schemes can be constructed in a natural way by combining the approximations of each partial derivative on its extended 1D stencil; for example, a 3D stencil with 13 nodes is obtained by combining three 5-point stencils in each coordinate direction, with the middle node shared. The resultant scheme is of fourth order. Another alternative is Collatz "Mehrstellen" schemes, in particular the fourth order scheme on a 19-point stencil considered in Section 2.8.5.

2.8.3 Flux-Balance Schemes in 3D

Consider now a 3D problem with a coordinate-dependent material parameter:

$$-\nabla \cdot (\epsilon(x, y, z)\nabla u) \;=\; f(x, y, z) \tag{2.125}$$

As before, ϵ will be assumed piecewise-smooth, with possible discontinuities only at material boundaries. The potential is continuous everywhere. The flux continuity conditions at material interfaces have the same form as in 2D:

$$\epsilon^- \frac{\partial u^-}{\partial n} \;=\; \epsilon^+ \frac{\partial u^+}{\partial n} \tag{2.126}$$

where "−" and "+" again refer to the values on the two sides of the interface boundary.

The integral form of the differential equation (2.125) is, by Gauss's Theorem

$$-\int_S \epsilon(x, y, z)\, \frac{\partial u}{\partial n}\, dS \;=\; \int_\omega f(x, y, z)\, d\omega \tag{2.127}$$

where ω is a subdomain of the computational domain Ω, S is the boundary surface of ω, and n is the normal to that boundary. As in 2D, the physical meaning of this integral condition is energy or flux balance, depending on the application.

A "control volume" ω to which the flux balance condition can be applied is (2.104) is shown in Fig. 2.20. The flux-balance scheme is completely analogous

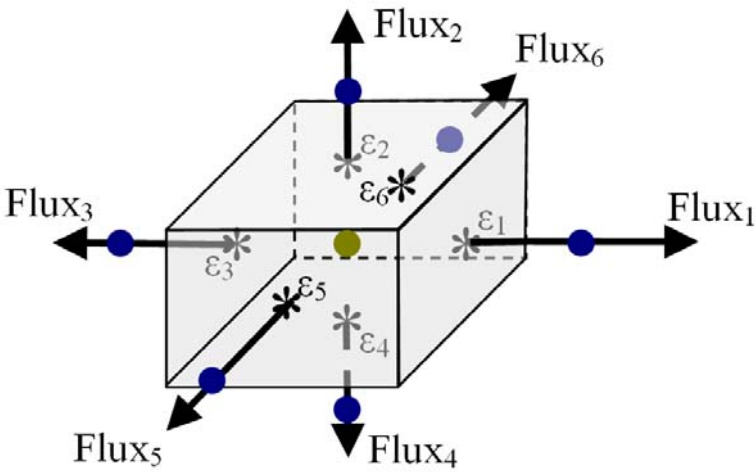


Fig. 2.20. Construction of the flux-balance scheme in three dimensions. The net flux out of the shaded control volume is equal to the total capacity of sources inside that volume. The grid nodes are shown as circles. For flux computation, the material parameters are taken at the midpoints of the faces.

to its 2D counterpart (see (2.106)):

$$\begin{aligned}
&\epsilon_1 h_y h_z \frac{u_{n_x,n_y,n_z} - u_{n_x+1,n_y,n_z}}{h_x} + \epsilon_2 h_x h_z \frac{u_{n_x,n_y,n_z} - u_{n_x,n_y+1,n_z}}{h_y} \\
&\quad + \epsilon_3 h_y h_z \frac{u_{n_x,n_y,n_z} - u_{n_x-1,n_y,n_z}}{h_x} + \epsilon_4 h_x h_z \frac{u_{n_x,n_y,n_z} - u_{n_x,n_y-1,n_z}}{h_y} \\
&\quad + \epsilon_5 h_x h_y \frac{u_{n_x,n_y,n_z} - u_{n_x,n_y,n_z+1}}{h_z} + \epsilon_6 h_x h_y \frac{u_{n_x,n_y,n_z} - u_{n_x,n_y,n_z-1}}{h_z} \\
&\quad = f(x_n, y_n, z_n)\, h_x h_y h_z
\end{aligned} \tag{2.128}$$

As in 2D, the accuracy of this scheme deteriorates in the vicinity of material interfaces, as the derivatives of the solution are discontinuous. Suitable basis functions satisfying the discontinuous boundary conditions are used in FLAME schemes (Chapter 4), which dramatically reduces the consistency error.

2.8.4 Implementation of 3D Schemes

Assuming for simplicity that the computational domain is a rectangular parallelepiped, one introduces a Cartesian grid with N_x, N_y and N_z subdivisions in the respective coordinate directions. The total number of nodes N_m in the mesh (including the boundary nodes) is $N_m = (N_x + 1)(N_y + 1)(N_z + 1)$. A natural node numbering is generated by letting, say, n_x change first, n_y second and n_z third, which assigns the global number

$$n = (N_x+1)(N_y+1)(n_z-1) + (N_x+1)(n_y-1) + n_x - 1, \qquad 1 \le n \le N \tag{2.129}$$

to node (n_x, n_y, n_z). When, say, a 7-point scheme is applied on all grid stencils, a 7-diagonal system matrix results. Two subdiagonals correspond to the connections of the central node (n_x, n_y, n_z) of the stencil to the neighboring nodes $(n_x \pm 1, n_y, n_z)$, another two subdiagonals to neighbors $(n_x, n_y \pm 1, n_z)$, and the remaining two subdiagonals to nodes $(n_x, n_y, n_z \pm 1)$. Boundary conditions are handled in a way completely analogous to the 2D case.

The selection of solvers for the resulting linear system of equations is in principle the same as in 2D, with direct and iterative methods being available. However, there is a practical difference. In two dimensions, thousands or tens of thousands of grid nodes are typically needed to achieve reasonable engineering accuracy; such problems can be easily solved with direct methods that are often more straightforward and robust than iterative algorithms. In 3D, the number of unknowns can easily reach hundreds of thousands or millions, in which case iterative methods may be the only option.[14]

[14] Even for the same number of unknowns in a 2D and a 3D problem, in the 3D case the number of nonzero entries in the system matrix is greater, the sparsity pattern of the matrix is different, and the 3D solver requires more memory and CPU time.

2.8.5 The Collatz "Mehrstellen" Schemes in 3D

The derivation and construction of the "Mehrstellen" schemes in 3D are based on the same ideas as in the 2D case, Section 2.7.4. For the Laplace equation, the "Mehrstellen" scheme can also be obtained as a direct and natural particular case of FLAME schemes in Chapter 4.

The 19-point stencil for a fourth order "Mehrstellen" scheme is obtained by discarding the eight corner nodes of a $3\times3\times3$ node cluster. The coefficients of the scheme for the Laplace equation on a uniform grid with $h_x = h_y = h_z$ are visualized in Fig. 2.21.

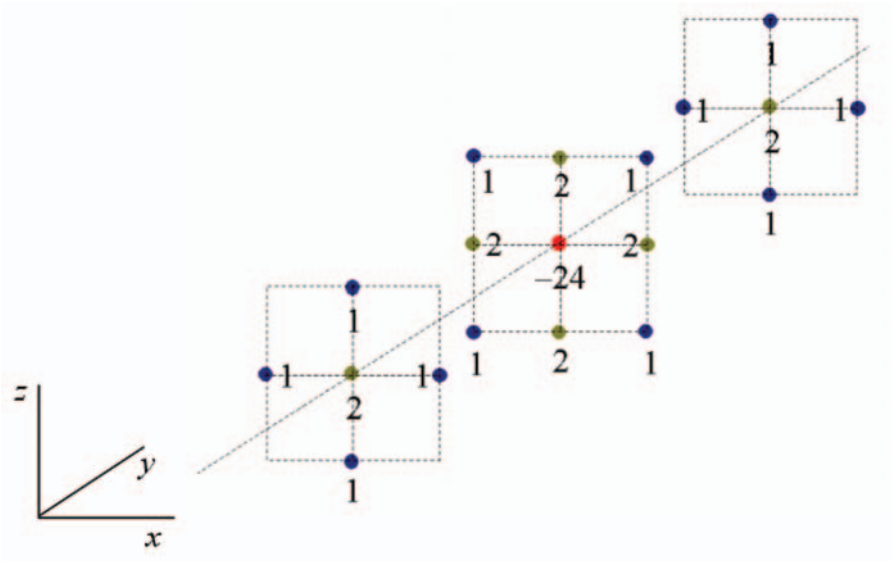


Fig. 2.21. For the Laplace equation, this fourth order "Mehrstellen"-Collatz scheme on the 19-point stencil is a direct particular case of Trefftz–FLAME. The grid sizes are equal in all three directions. For visual clarity, the stencil is shown as three slices along the y axis. (Reprinted by permission from [Tsu06] ©2006 Elsevier.)

In the more general case of unequal mesh sizes in the x-, y- and $z-$directions, the "Mehrstellen" scheme is derived in the monographs by L. Collatz and A.A. Samarskii. E.L. Briggs *et al.* [BSB96] list the coefficients of the scheme in a concise table form. The end result is as follows.

The coefficient corresponding to the central node of the stencil is $4/3\sum_\alpha h_\alpha^{-2}$ (where $\alpha = x,\ y,\ z$). The coefficients corresponding to the two immediate neighbors of the central node in the α direction are $-5/6h_\alpha^{-2} + 1/6\sum_\beta h_\beta^{-2}$ ($\beta = x,\ y$ or z). Finally, the coefficients corresponding to the nodes displaced by h_α and h_β in both α- and β-coordinate directions relative to the central node are $-1/12h_\alpha^{-2}\ -1/12h_\beta^{-2}$.

If the Poisson equation (2.123) rather than the Laplace equation is solved, with $f = f(x, y, z)$ a smooth function of coordinates, the right hand side of the 19-point Mehrstellen scheme is $f_h = \frac{1}{2} f_0 + \frac{1}{12} \sum_{\alpha=1}^{6} f_\alpha$, where f_0 is the value of f at the middle node of the stencil and f_α are the values of f at the six immediate neighbors of that middle node. Thus the computation of the right hand side involves the same 7-point stencil as for the standard second-order scheme for the Laplace equation, not the whole 19-point stencil. HODIE schemes by R.E. Lynch & J.R. Rice [LR80] generalize the Mehrstellen schemes and include additional points in the computation of the right hand side.

2.9 Consistency and Convergence of Difference Schemes

This section presents elements of convergence and accuracy theory of FD schemes. A more comprehensive and rigorous treatment is available in many monographs (e.g. L. Collatz [Col66], A.A. Samarskii [Sam01], J.C. Strikwerda [Str04], W.E. Milne [Mil70]).

Consider a differential equation in 1D, 2D or 3D

$$Lu = f \tag{2.130}$$

that we wish to approximate by a difference scheme

$$L_h^{(i)} \underline{u}_h^{(i)} = f_{hi} \tag{2.131}$$

on stencil (i) containing a given set of grid nodes. Here $\underline{u}_h^{(i)}$ is the Euclidean vector of the nodal values of the numerical solution *on the stencil*. Merging the difference schemes on all stencils into a global system of equations, one obtains

$$L_h \underline{u}_h = \underline{f}_h \tag{2.132}$$

where $\underline{u}_h$ and $\underline{f}_h$ are the numerical solution and the right hand side, respectively, viewed as Euclidean vectors of nodal values on the whole grid.

Exactly in what sense does (2.132) approximate the original differential equation (2.130)? A natural requirement is that the exact solution u^* of the differential equation should approximately satisfy the difference equation.

To write this condition rigorously, we need to substitute u^* into the difference scheme (2.132). Since this scheme operates on the *nodal values* of u^*, a notation for these nodal values is in order. We shall use the calligraphic letter $\mathcal{N}$ for this purpose: $\mathcal{N}u^*$ will mean the Euclidean vector of nodal values of u^* on the whole grid. Similarly, $\mathcal{N}^{(i)}u^*$ is the Euclidean vector of nodal values of u^* on a given stencil (i).

The consistency error *vector* $\underline{\epsilon}_c \equiv \{\epsilon_{ci}\}_{i=1}^n$ of scheme (2.132) is the residual obtained when the exact solution is substituted into the difference equation; that is,

$$L_h \mathcal{N} u^* = \underline{f}_h + \underline{\epsilon}_c \tag{2.133}$$

where as before the underscored symbols are Euclidean vectors. The consistency error (a number) is defined as a norm of the error vector:

$$\text{consistency error} \;\equiv\; \epsilon_c(h) \;=\; \|\underline{\epsilon}_c\|_k \;=\; \left\| L_h \mathcal{N} u^* \;-\; \underline{f}_h \right\|_k \tag{2.134}$$

where k is usually 1, 2, or ∞ (see Appendix 2.10 for definitions of these norms). There is, however, one caveat. According to definition (2.134), the meaningless scheme $h^{100} u_i = 0$ has consistency error of order 100 for *any* differential equation with a bounded solution. It is natural to interpret such high-order consistency just as an artifact of scaling and to apply a normalization condition across the board for all schemes. Specifically, we shall assume that the difference schemes are scaled in such a way that

$$c_1 f(r) \;\le\; \underline{f}_{hi} \;\le\; c_2 f(r), \quad \forall r \in \Omega^{(i)} \tag{2.135}$$

where $c_{1,2}$ do not depend on i and h.

We shall call a scheme consistent if, with scaling (2.135), the consistency error tends to zero as $h \to 0$:

$$\epsilon_c \;=\; \left\| L_h^{(i)} \left(\mathcal{N}^{(i)} u^* \;-\; \underline{u}_{hi} \right) \right\|_k \;\to\; 0 \;\text{ as } h \to 0 \tag{2.136}$$

Consistency is usually relatively easy to establish. For example, the Taylor expansions in Section 2.6.1 show that the consistency error of the three-point scheme for the Poisson equation in 1D is $\mathcal{O}(h^2)$; see (2.87)–(2.89). This scheme is therefore consistent.

Unfortunately, consistency by itself does not guarantee convergence. To see why, let us compare the difference equations satisfied by the numerical solution and the exact solution, respectively:

$$L_h \mathcal{N} u^* \;=\; \underline{f}_h \;+\; \underline{\epsilon}_c \tag{2.137}$$

$$L_h \underline{u}_h \;=\; \underline{f}_h \tag{2.138}$$

These are equations (2.132) and (2.133) written together for convenience. Clearly, systems of equations for the exact solution u^* (more precisely, its nodal values $\mathcal{N} u^*$) and for the numerical solution $\underline{u}_h$ have slightly different right hand sides. Consistency error ϵ_c is a measure of the *residual* of the difference equation, which is different from the accuracy *of the numerical solution* of this equation.

Does the small difference $\underline{\epsilon}_c$ in the right hand sides of (2.137) and (2.138) translate into a comparably small difference in the *solutions* themselves? If yes, the scheme is called *stable*. A formal definition of stability is as follows:

$$\epsilon_h \;\equiv\; \|\underline{\epsilon}_h\|_k \;\equiv\; \|\underline{u}_h \;-\; \mathcal{N} u^*\|_k \;\le\; C \|\underline{\epsilon}_c\|_k \tag{2.139}$$

where the factor C may depend on the exact solution u^* but not on the mesh size h.

Stability constant C is linked to the properties of the inverse operator L_h^{-1}. Indeed, subtracting (2.137) from (2.138), one obtains an expression for the error vector:

$$\underline{\epsilon}_h \;\equiv\; \underline{u}_h \,-\, \mathcal{N}u^* \;=\; L_h^{-1}\underline{\epsilon}_c \tag{2.140}$$

(assuming that L_h is nonsingular). Hence the numerical error can be estimated as

$$\epsilon_h \;\equiv\; \|\underline{\epsilon}_h\|_k \;\equiv\; \|\underline{u}_h \,-\, \mathcal{N}u^*\|_k \;\leq\; \|L_h^{-1}\|_k\,\|\underline{\epsilon}_c\|_k \tag{2.141}$$

where the matrix norm for L_h is induced by the vector norm, i.e., for a generic square matrix A,

$$\|A\|_k \;=\; \max_{x\neq 0}\frac{\|Ax\|_k}{\|x\|_k}$$

(see Appendix 2.10).

In summary, convergence of the scheme follows from consistency *and* stability. This result is known as the *Lax–Richtmyer Equivalence Theorem* (see e.g. J.C. Strikwerda [Str04]).

To find the consistency error of a scheme, one needs to substitute the exact solution into it and evaluate the residual (e.g. using Taylor expansions). This is a relatively straightforward procedure. In contrast, stability (and, by implication, convergence) are in general much more difficult to establish.

For conventional difference schemes and the Poisson equation, convergence is proved in standard texts (e.g. W.E. Milne [Mil70] or J.C. Strikwerda [Str04]). This convergence result in fact applies to a more general class of *monotone* schemes.

Definition 6. *A difference operator L_h (and the respective $N_m \times N_m$ matrix) is called* monotone *if $L_h x \geq 0$ for vector $x \in \mathbb{R}^{N_m}$ implies $x \geq 0$, where vector inequalities are understood entry-wise.*

In other words, if L_h is monotone and $L_h x$ has all nonnegative entries, vector x must have all nonnegative entries as well. Algebraic conditions related to monotonicity are reviewed at the end of this subsection.

To analyze convergence of monotone schemes, the following Lemma will be needed.

Lemma 1. *If the scheme is scaled according to (2.135) and the consistency condition (2.134) holds, there exists a reference nodal vector $\underline{u}_{1h}$ such that*

$$\underline{u}_{1h} \;\leq\; U_1 \quad\text{and}\quad L_h\underline{u}_{1h} \geq \sigma_1 > 0, \tag{2.142}$$

with numbers U_1 and σ_1 independent of h. (All vector inequalities are understood entry-wise.)

Remark 1. (Notation.) Subscript 1 is meant to show that, as seen from the proof below, the auxiliary potential u_{1h} may be related to the solution of the differential equation with the unit right hand side.

Proof. The reference potential u_{1h} can be found explicitly by considering the auxiliary problem

$$Lu_1 \;=\; 1 \tag{2.143}$$

with the same boundary conditions as the original problem. Condition (2.136) applied to the nodal values of u_1 implies that for sufficiently small h the consistency error will fall below $\frac{1}{2}\,c_1$, where c_1 is the parameter in (2.135):

$$\left|\underline{s}^{(i)T}\mathcal{N}^{(i)}u_1 \;-\; f_{hi}\right| \;\le\; \frac{1}{2}\,c_1$$

Therefore, since $f = 1$ in (2.135),

$$|\underline{s}^{(i)T}\mathcal{N}^{(i)}u_1| \;\ge\; |f_{hi}| \;-\; \left|f_{hi} \;-\; \underline{s}^{(i)T}\mathcal{N}^{(i)}u_1\right| \;\ge\; c_1 \;-\; \frac{1}{2}\,c_1 \;=\; \frac{1}{2}\,c_1 \tag{2.144}$$

(the vector inequality is understood entry-wise). Thus one can set $\underline{u}_{1h} = L_h\mathcal{N}u_1$, with $\sigma_1 = \frac{1}{2}\,c_1$ and $U_1 = \|u_1\|_\infty$. □

Theorem 1. *Let the following conditions hold for difference scheme (2.132):*

1. *Consistency in the sense of (2.136), (2.135).*
2. $$\text{Monotonicity}: \text{ if } \; L_h\underline{x} \;\ge\; 0, \quad \text{then } \underline{x} \;\ge\; 0 \tag{2.145}$$

Then the numerical solution converges in the nodal norm, and

$$\|\underline{u}_h - \mathcal{N}u^*\|_\infty \;\le\; \epsilon_c\,U_1/\sigma_1 \tag{2.146}$$

where σ_1 is the parameter in (2.142).

Proof. Let $\underline{\epsilon}_h = \underline{u}_h - \mathcal{N}u^*$. By consistency,

$$L_h\,\underline{\epsilon}_h \;\le\; \epsilon_c \;\le\; \epsilon_c\,L_h\,\underline{u}_{1h}/\sigma_1 \;=\; L_h(\epsilon_c\,\underline{u}_{1h}/\sigma_1)$$

where (2.142) was used. Hence due to monotonicity

$$\epsilon_h \;\le\; \epsilon_c\,\underline{u}_{1h}/\sigma_1 \tag{2.147}$$

It then also follows that

$$\epsilon_h \;\ge\; -\,\epsilon_c\,\underline{u}_{1h}/\sigma_1 \tag{2.148}$$

Indeed, if that were not true, one would have $(-\epsilon_h) \;\le\; \epsilon_c\,\underline{u}_{1h}/\sigma_1$, which would contradict the error estimate (2.147) for the system with $(-\underline{f})$ instead of $\underline{f}$ in the right hand side. □

We now summarize sufficient and/or necessary algebraic conditions for monotonicity. Of particular interest is the relationship of monotonicity to diagonal dominance, as the latter is trivial to check for any given scheme.

The summary is based on the monograph of R.S. Varga [Var00] and the reference book of V. Voevodin & Yu.A. Kuznetsov [VK84]. The mathematical facts are cited without proof.

Proposition 1. *A square matrix A is monotone if and only if it is nonsingular and $A^{-1} \geq 0$.*

[As a reminder, all matrix and vector inequalities in this section are understood entry-wise.]

Definition 7. *A square matrix A is called an M-matrix if it is nonsingular $a_{ij} \leq 0$ for all $i \neq j$ and $A^{-1} \geq 0$.*

Thus an M-matrix, in addition to being monotone, has nonpositive off-diagonal entries.

Proposition 2. *All diagonal elements of an M-matrix are positive.*

Proposition 3. *Let a square matrix A have nonpositive off-diagonal entries. Then the following conditions are equivalent:*

1. *A is an M-matrix.*
2. *There exists a positive vector w such that $A^{-1}w$ is also positive.*
3. *Re $\lambda > 0$ for any eigenvalue λ of A.*

(See [VK84] §36.15 for additional equivalent statements.)

Notably, the second condition above allows one to demonstrate monotonicity by exhibiting just *one* special vector satisfying this condition, which is simpler than verifying this condition for *all* vectors as stipulated in the definition of monotonicity.

Even more practical is the connection with diagonal dominance [VK84].

Proposition 4. *Let a square matrix A have nonpositive off-diagonal entries. If this matrix has strong diagonal dominance, it is an M-matrix.*

Proposition 5. *Let an* irreducible *square matrix A have nonpositive off-diagonal entries. If this matrix has weak diagonal dominance, it is an M-matrix. Moreover, all entries of A^{-1} are then (strictly) positive.*

A matrix is called irreducible if it cannot be transformed to a block-triangular form by permuting its rows and columns. The definition of weak diagonal dominance for a matrix A is

$$|A_{ii}| \;\geq\; \sum\nolimits_j |A_{ij}| \tag{2.149}$$

in each row i. The condition of strong diagonal dominance is obtained by changing the inequality sign to strict.

Thus diagonal dominance of matrix L_h of the difference scheme is a sufficient condition for monotonicity if the off-diagonal entries of L_h are nonpositive. As a measure of the relative magnitude of the diagonal elements, one can use

$$q \;=\; \frac{\min_i |L_{h,\,ii}|}{\sum_j |L_{h,ij}|} \tag{2.150}$$

with matrix L_h being weakly diagonally dominant for $q = 0.5$ and diagonal for $q = 1$. Diagonal dominance is a strong condition that unfortunately does not hold in general.

2.10 Summary and Further Reading

This chapter is an introduction to the theory and practical usage of finite difference schemes. Classical FD schemes are constructed by the Taylor expansion over grid stencils; this was illustrated in Sections 2.1–2.2 and parts of Sections 2.6–2.8. The chapter also touched upon classical schemes (Runge–Kutta, Adams and others) for ordinary differential equations and special schemes that preserve physical invariants of Hamiltonian systems.

Somewhat more special are the Collatz "Mehrstellen" schemes for the Poisson equation. These schemes (9-point in 2D and 19-point in 3D) are described in Sections 2.7.4 and 2.8.5. Higher approximation accuracy is achieved, in essence, by approximating the *solution* of the Poisson equation rather than a generic smooth function. We shall return to this idea in Chapter 4 and will observe that the Mehrstellen schemes are, at least for the Laplace equation, a natural particular case of "Flexible Local Approximation MEthods" (FLAME) considered in that chapter. In fact, in FLAME the classic FD schemes and the Collatz Mehrstellen schemes stem from one single principle and one single definition of the scheme.

Very important are the schemes based on flux or energy balance for a control volume; see Sections 2.6.3, 2.7.2, and 2.8.3. Such schemes are known to be quite robust, which is particularly important for problems with inhomogeneous media and material interfaces. The robustness can be attributed to the underlying solid physical principles (conservation laws).

For further general reading on FD schemes, the interested reader may consider the monographs by L. Collatz [Col66], J.C. Strikwerda [Str04], A.A. Samarskii [Sam01].

A comprehensive source of information not just on FD schemes but also on numerical methods for ordinary and partial differential equations in general is the book by A. Iserles [Ise96]. It covers one-step and multistep schemes for ODE, Runge–Kutta methods, schemes for stiff systems, FD schemes for the Poisson equation, the Finite Element Method, algebraic system solvers, multigrid and other fast solution methods, diffusion and hyperbolic equations.

For readers interested in schemes for fluid dynamics, S.V. Patankar's text [Pat80] may serve as an introduction. A more advanced book by T.J. Chung [Chu02] covers not only finite-difference, but also finite-volume and finite element methods for fluid flow. Also well-known and highly recommended are two monographs by R.J. LeVeque: one on schemes for advection-diffusion equations, with the emphasis on conservation laws [LeV96], and another one with a comprehensive treatment of hyperbolic problems [LeV02a]. The book by H.-G. Roos *et al.* [HGR96], while focusing (as the title suggests) on the mathematical treatment of singularly perturbed convection-diffusion problems, is also an excellent source of information on finite-difference schemes in general.

For theoretical analysis and computational methods for fluid dynamics *on the microscale*, see books by G. Karniadakis *et al.* [KBA01] and by J.A. Pelesko & D.H. Bernstein [PB02].

Several monographs and review papers are devoted to schemes for electromagnetic applications. The literature on Finite-Difference Time-Domain (FDTD) schemes for electromagnetic wave propagation is especially extensive; see http://www.fdtd.org. The most well-known FDTD monograph is by A. Taflove & S.C. Hagness [TH05]. The book by A.F. Peterson *et al.* [PRM98] covers, in addition to FD schemes in both time and frequency domain, integral equation techniques and the Finite Element Method for computational electromagnetics.

Appendix: Frequently Used Vector and Matrix Norms

The following vector and matrix norms are used most frequently.

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \tag{2.151}$$

$$\|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^{n} |A_{ij}| \tag{2.152}$$

$$\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{\frac{1}{2}} \tag{2.153}$$

$$\|A\|_2 = \max_{1\le i\le n} \lambda_i^{\frac{1}{2}}(A^*A) \tag{2.154}$$

where A^* is the Hermitian conjugate (= the conjugate transpose) of matrix A, and λ_i are the eigenvalues.

$$\|x\|_\infty = \max_{1\le i\le n} |x_i| \tag{2.155}$$

$$\|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |A_{ij}| \tag{2.156}$$

See linear algebra textbooks, e.g. Y. Saad [Saa03], R.A. Horn & C.R. Johnson [HJ90], F.R. Gantmakher [Gan59, Gan88] for further analysis and proofs.

Appendix: Matrix Exponential

It is not uncommon for an operation over some given class of objects to be defined in two (or more) different ways that for this class are equivalent. Yet one of these ways could have a broader range of applicability and can hence be used to generalize the definition of the operation.

This is exactly the case for the exponential operation. One way to define $\exp x$ is via simple arithmetic operations – first for x integer via repeated multiplications, then for x rational via roots, and then for all real x.[15] While

[15] The rigorous mathematical theory – based on either Dedekind's cuts or Cauchy sequences – is, however, quite involved; see e.g. W. Rudin [Rud76].

this definition works well for real numbers, its direct generalization to, say, complex numbers is not straightforward (because of the ambiguity of roots), and generalization to more complicated objects like matrices is even less clear.

At the same time, the exponential function admits an alternative definition via the Taylor series

$$\exp x \;=\; \sum_{n=0}^{\infty} \frac{x^n}{n!} \tag{2.157}$$

that converges absolutely for all x. This definition is directly applicable not only to complex numbers but to matrices and operators. Matrix exponential can be *defined* as

$$\exp A \;=\; \sum_{n=0}^{\infty} \frac{A^n}{n!} \tag{2.158}$$

where A is an arbitrary square matrix (real or complex). This infinite series converges for any matrix, and $\exp(A)$ defined this way can be shown to have many of the usual properties of the exponential function – most notably,

$$\exp((\alpha+\beta)A) \;=\; \exp(\alpha A) \;+\; \exp(\beta A), \quad \forall\alpha\in\mathbb{C}, \;\; \forall\beta\in\mathbb{C} \tag{2.159}$$

If A and B are two *commuting* square matrices of the same size, $AB = BA$, then

$$\exp(A+B) \;=\; \exp A \exp B, \qquad \text{if } AB \;=\; BA \tag{2.160}$$

Unfortunately, for non-commuting matrices this property is not generally true.

For a system of ordinary differential equations written in matrix-vector form as

$$\frac{dy(t)}{dt} \;=\; Ay, \quad y\in\mathbb{R}^n \tag{2.161}$$

the solution can be expressed via matrix exponential in a very simple way:

$$y(t) \;=\; \exp(At)\, y_0 \tag{2.162}$$

Note that if matrices A and $\tilde{A}$ are related via a similarity transform

$$A \;=\; S^{-1}\tilde{A}S \tag{2.163}$$

then

$$A^2 \;=\; S^{-1}\tilde{A}SS^{-1}\tilde{A}S \;=\; S^{-1}\tilde{A}^2S$$

and $A^3 \;=\; S^{-1}\tilde{A}^3S$, etc. – i.e. powers of A and $\tilde{A}$ are related via the same similarity transform. Substituting this into the Taylor series (2.158) for matrix exponential, one obtains

$$\exp A \;=\; S^{-1}\exp\tilde{A}\,S \tag{2.164}$$

This is particularly useful if matrix A is diagonalizable; then $\tilde{A}$ can be made diagonal and contains the eigenvalues of A, and $\exp(\tilde{A})$ is a diagonal matrix containing the exponents of these eigenvalues.[16]

[16] Matrices with distinct eigenvalues are diagonalizable; so are symmetric matrices.

Since matrix exponential is intimately connected with such difficult problems as full eigenvalue analysis and solution of general ODE systems, it is not surprising that the computation of $\exp(A)$ is itself highly complex in general. The curious reader may find it interesting to see the "nineteen dubious ways to compute the exponential of a matrix" (C. Moler & C. Van Loan, [ML78], [ML03]; see also W.A. Harris *et al.* [WAHFS01].

3

The Finite Element Method

3.1 Everything is Variational

The Finite Element Method (FEM) belongs to the broad class of variational methods, and so it is natural to start this chapter with an introduction and overview of such methods. This section emphasizes the importance of the variational approach to computation: it can be claimed – with only a small bit of exaggeration – that all numerical methods are variational.

To understand why, let us consider the Poisson equation in one, two or three dimensions as a model problem:

$$\mathcal{L}u \equiv -\nabla^2 u = \rho \quad \text{in } \Omega \tag{3.1}$$

This equation describes, for example, the distribution of the electrostatic potential u corresponding to volume charge density ρ if the dielectric permittivity is normalized to unity.

Solution u is sought in a functional space $V(\Omega)$ containing functions with a certain level of smoothness and satisfying some prescribed conditions on the boundary of domain Ω; let us assume zero Dirichlet conditions for definiteness. For purposes of this introduction, the precise mathematical details about the level of smoothness of the right hand side ρ and the boundary of the 2D or 3D domain Ω are not critical, and I mention them only as a footnote.[1] It is important to appreciate that solution u has infinitely many "degrees of freedom" – in mathematical terms, it lies in an infinite-dimensional functional space. In

[1] The domain is usually assumed to have a Lipschitz-continuous boundary; $f \in L_2(\Omega)$, $u \in H^2(\Omega)$, where L_2 and H^2 are the Lebesgue and Sobolev spaces standard in mathematical analysis. The requirements on the smoothness of u are relaxed in the weak formulation of the problem considered later in this chapter. Henri Léon Lebesgue (1875–1941) – a French mathematician who developed measure and integration theory. Sergei L'vovich Sobolev (1908–1989) – a Russian mathematician, renowned for his work in mathematical analysis (Sobolev spaces, weak solutions and generalized functions).

contrast, any *numerical* solution can only have a finite number of parameters. A general and natural form of such a solution is a linear combination of a finite number n of linearly independent approximating functions $\psi_\alpha \in V(\Omega)$:

$$u_{\text{num}} = \sum_{\alpha=1}^{n} c_\alpha \psi_\alpha \tag{3.2}$$

where c_α are some coefficients (in the example, real; for other problems, these coefficients could be complex). We may have in mind a set of polynomial functions as a possible example of ψ_α ($\psi_1 = 1$, $\psi_2 = x$, $\psi_3 = y$, $\psi_4 = xy$, $\psi_5 = x^2$, etc., in 2D). One important constraint, however, is that these functions must satisfy the Dirichlet boundary conditions, and so only a subset of polynomials will qualify. One of the distinguishing features of finite element analysis is a special procedure for defining *piecewise*-polynomial approximating functions. This procedure will be discussed in more detail in subsequent sections.

The key question now is: what are the "best" parameters c_α that would produce the most accurate numerical solution (3.2)? Obviously, we first need to define "best". It would be ideal to have a zero residual

$$R \equiv \mathcal{L} u_{\text{num}} - \rho \tag{3.3}$$

in which case the numerical solution would in fact be exact. That being in general impossible, the constraints on R need to be relaxed. While R may not be identically zero, let us require that there be a set of "measures of fitness" of the solution – numbers $f_\beta(R)$ – that are zero:

$$f_\beta(R) = 0, \qquad \beta = 1, 2, \ldots, n \tag{3.4}$$

It is natural to have the number of these measures, i.e. the number of conditions (3.4), equal to the number of undetermined coefficients c_α in expansion (3.2).

In mathematical terms, the numbers f_β are *functionals*: each of them acts on a function (in this case, R) and produces a number $f_\beta(R)$. The functionals can be real or complex, depending on the problem.

To summarize: the numerical solution is sought as a linear combination of n approximating functions, with n unknown coefficients; to determine these coefficients, one imposes n conditions (3.4). As it is difficult to deal with nonlinear constraints, the functionals f_β are almost invariably chosen as linear.

Example 1. Consider the 1D Poisson equation with the right hand side $\rho(x) = \cos x$ over the interval $[-\pi/2, \pi/2]$:

$$-\frac{d^2 u}{dx^2} = \cos x, \qquad u\left(-\frac{\pi}{2}\right) = u\left(\frac{\pi}{2}\right) = 0 \tag{3.5}$$

The obvious exact solution is $u^*(x) = \cos x$. Let us find a numerical solution using the ideas outlined above.

Let the approximating functions ψ_α be polynomials in x. To keep the calculation as simple as possible, the number of approximating functions in this example will be limited to two only. Linear polynomials (except for the one identically equal to zero) do not satisfy the zero Dirichlet boundary conditions and hence are not included in the approximating set. As the solution must be an even function of x, a sensible (but certainly not unique) choice of the approximating functions is

$$\psi_1 = \left(x - \frac{\pi}{2}\right)\left(x + \frac{\pi}{2}\right), \qquad \psi_2 = \left(x - \frac{\pi}{2}\right)^2\left(x + \frac{\pi}{2}\right)^2 \tag{3.6}$$

The numerical solution is thus

$$u_{\text{num}} \;=\; \underline{u}_1\psi_1 \;+\; \underline{u}_2\psi_2 \tag{3.7}$$

Here $\underline{u}$ is a Euclidean coefficient vector in $\mathbb{R}^2$ with components $\underline{u}_{1,2}$. Euclidean vectors are underlined to distinguish them from functions of spatial variables.

The residual (3.3) then is

$$R \;=\; -\,\underline{u}_1\psi_1'' \;-\; \underline{u}_2\psi_2'' \;-\; \cos x \tag{3.8}$$

As a possible example of "fitness measures" of the solution, consider two functionals that are defined as the values of R at points $x = 0$ and $x = \pi/4$:[2]

$$f_1(R) \;=\; R(0); \qquad f_2(R) \;=\; R\left(\frac{\pi}{4}\right) \tag{3.9}$$

With this choice of the test functionals, residual R, while not zero everywhere (which would be ideal but ordinarily not achievable), is forced by conditions (3.4) to be zero at least at points $x = 0$ and $x = \pi/4$. Furthermore, due to the symmetry of the problem, R will automatically be zero at $x = -\pi/4$ as well; this extra point comes as a bonus in this example. Finally, the residual is zero at the boundary points because both exact and numerical solutions satisfy the same Dirichlet boundary condition by construction.

The reader may recognize functionals (3.9) as *Dirac delta* functions $\delta(x)$ and $\delta(x - \pi/4)$, respectively. The use of Dirac deltas as test functionals in variational methods is known as *collocation*; the value of the residual is forced to be zero at a certain number of "collocation points" – in this example, two: $x = 0$ and $x = \pi/4$.

The two functionals (3.9), applied to residual (3.8), produce a system of two equations with two unknowns $\underline{u}_{1,2}$:

$$-\underline{u}_1\psi_1''(0) \;-\; \underline{u}_2\psi_2''(0) \;-\; \cos 0 \;=\; 0$$

[2] It is clear that these functionals are linear. Indeed, to any linear combination of two different Rs there corresponds a similar linear combination of their pointwise values.

$$-\underline{u}_1\psi_1''\left(\frac{\pi}{4}\right) \;-\; \underline{u}_2\psi_2''\left(\frac{\pi}{4}\right) \;-\; \cos\frac{\pi}{4} \;=\; 0$$

In matrix-vector form, this system is

$$L\underline{u} \;=\; \underline{\rho}, \quad L \;=\; -\begin{pmatrix} \psi_1''(0) & \psi_2''(0) \\ \psi_1''(\frac{\pi}{4}) & \psi_2''(\frac{\pi}{4}) \end{pmatrix}; \quad \underline{\rho} \;=\; \begin{pmatrix} \cos 0 \\ \cos\frac{\pi}{4} \end{pmatrix} \;=\; \begin{pmatrix} 1 \\ \frac{\sqrt{2}}{2} \end{pmatrix} \tag{3.10}$$

It is not difficult to see that for an arbitrary set of approximating functions ψ and test functionals f the entry $L_{\alpha\beta}$ of this matrix is $f_\alpha(\psi_\beta)$. In the present example, with the approximating functions chosen as (3.6), matrix L is easily calculated to be

$$L \;\approx\; \begin{pmatrix} -2 & 9.869604 \\ -2 & 2.467401 \end{pmatrix}$$

with seven digits of accuracy. The vector of expansion coefficients then is

$$\underline{u} \;\approx\; \begin{pmatrix} -0.3047378 \\ 0.03956838 \end{pmatrix}$$

With these values of the coefficients, and with the approximating functions of (3.6), the numerical solution becomes

$$u_{\text{num}} \;\approx\; -\,0.3047378\left(x-\frac{\pi}{2}\right)\left(x+\frac{\pi}{2}\right) \;+\; 0.03956838\left(x-\frac{\pi}{2}\right)^2\left(x+\frac{\pi}{2}\right)^2 \tag{3.11}$$

The numerical error is shown in Fig. 3.2 and its absolute value is in the range of $(3 \div 8) \times 10^{-3}$. The energy norm of this error is ~ 0.0198. (Energy norm is defined as

$$\|w\|_E \;=\; \left[\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \left(\frac{dw}{dx}\right)^2 dx\right]^{\frac{1}{2}} \tag{3.12}$$

for any differentiable function $w(x)$ satisfying the Dirichlet boundary conditions.)[3] Given that the numerical solution involves only two approximating functions with only two free parameters, the result certainly appears to be remarkably accurate.[4]

This example, with its more than satisfactory end result, is a good first illustration of variational techniques. Nevertheless the approach described above is difficult to turn into a systematic and robust methodology, for the following reasons:

1. The approximating functions and test functionals (more specifically, the collocation points) have been chosen in an *ad hoc* way; no systematic strategy is apparent from the example.

[3] In a more rigorous mathematical context, w would be treated as a function in the Sobolev space $H_0^1[-\frac{\pi}{2}, \frac{\pi}{2}]$, but for the purposes of this introduction this is of little consequence.

[4] Still, an even better numerical solution will be obtained in the following example (Example 2 on p. 73).

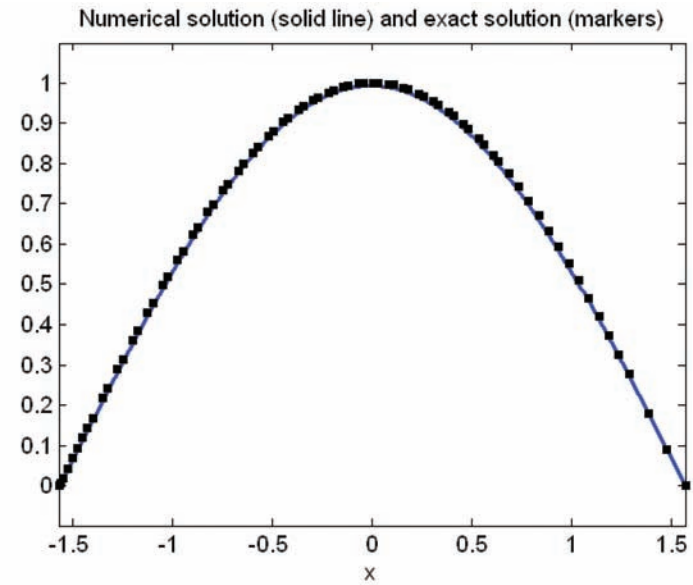


Fig. 3.1. Solution by collocation (3.11) in Example 1 (solid line) is almost indistinguishable from the exact solution $u^* = \cos x$ (markers). See also error plot in Fig. 3.2.

2. It is difficult to establish convergence of the numerical solution as the number of approximating functions increases, even if a reasonable way of choosing the approximating functions and collocation points is found.
3. As evident from (3.10), the approximating functions must be twice differentiable. This may be too strong a constraint. It will become apparent in the subsequent sections of this chapter that the smoothness requirements should be, from both theoretical and practical point of view, as weak as possible.

The following example (Example 2) addresses the convergence issue and produces an even better numerical solution for the 1D Poisson equation considered above. The Finite Element Method covered in the remainder of this chapter provides an elegant framework for resolving all three matters on the list.

Example 2. Let us consider the same Poisson equation as in the previous example and the same approximating functions $\psi_{1,2}$ (3.6). However, the test functionals $f_{1,2}$ are now chosen in a different way:

$$f_\alpha(R) \;=\; \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} R(x)\,\psi_\alpha(x)\,dx \tag{3.13}$$

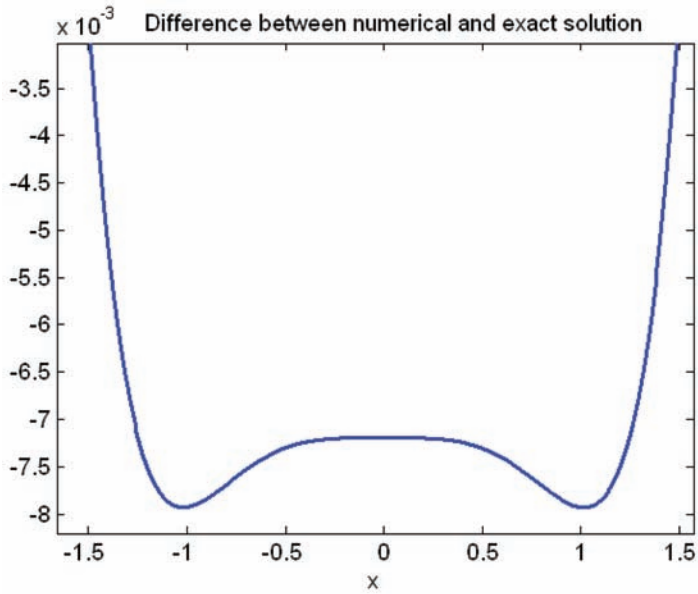


Fig. 3.2. Error of solution by collocation (3.11) in Example 1. (Note the 10^{-3} scaling factor.)

In contrast with collocation, these functionals "measure" *weighted averages* rather than point-wise values of R.[5] Note that the weights are taken to be exactly the same as the approximating functions ψ; this choice signifies the *Galerkin method*.

Substituting $R(x)$ (3.8) into Galerkin equations (3.13), we obtain a linear system

$$L\underline{u} \;=\; \underline{\rho}, \quad L_{\alpha\beta} \;=\; -\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \psi''_\beta(x)\,\psi_\alpha(x)\,dx; \quad \underline{\rho}_\alpha \;=\; \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \rho(x)\,\psi_\alpha(x)\,dx \quad (3.14)$$

Notably, the expression for matrix entries $L_{\beta\alpha}$ can be made more elegant using integration by parts and taking into account zero boundary conditions:

$$L_{\alpha\beta} \;=\; \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \psi'_\alpha(x)\,\psi'_\beta(x)\,dx \qquad (3.15)$$

This reveals the symmetry of the system matrix. The symmetry is due to two factors: (i) the operator $\mathcal{L}$ of the problem – in this case, Laplacian in the space of functions with zero Dirichlet conditions – is self-adjoint; this allowed

[5] Loosely speaking, collocation can be viewed as a limiting case of weighted averaging, with the weight concentrated at one point as the Dirac delta.

the transformation of the integrand to the symmetric form; (ii) the Galerkin method was used.

The Galerkin integrals in the expressions for the system matrix (3.15) and the right hand side (3.14) can be calculated explicitly:[6]

$$L \;=\; \frac{\pi^3}{105}\begin{pmatrix} 35 & -7\pi^2 \\ -7\pi^2 & -2\pi^4 \end{pmatrix}; \qquad \underline{\rho} \;=\; \begin{pmatrix} -4 \\ 48-4\pi^2 \end{pmatrix} \tag{3.16}$$

Naturally, this matrix is different from the matrix in the collocation method of the previous example (albeit denoted with the same symbol). In particular, the Galerkin matrix is symmetric, while the collocation matrix is not.

The expansion coefficients in the Galerkin method are

$$\underline{u} \;=\; L^{-1}\underline{\rho} \;=\; \frac{1}{\pi^7}\begin{pmatrix} -60\pi^2\,(3\pi^2-28) \\ -840\,(\pi^2-10) \end{pmatrix} \;\approx\; \begin{pmatrix} -0.3154333 \\ 0.03626545 \end{pmatrix}$$

The numerical values of these coefficients differ slightly from the ones obtained by collocation in the previous example. The Galerkin solution is

$$u_{\text{num}} \;\approx\; -\,0.3154333\left(x-\frac{\pi}{2}\right)\left(x+\frac{\pi}{2}\right) \;+\; 0.03626545\left(x-\frac{\pi}{2}\right)^2\left(x+\frac{\pi}{2}\right)^2 \tag{3.17}$$

The error of solution (3.17) is plotted in Fig. 3.3; it is seen to be substantially smaller than the error for collocation. Indeed, the energy norm of this error is ~ 0.004916, which is almost exactly four times less than the same error measure for collocation.

The higher accuracy of the Galerkin solution (at least in the energy norm) is not an accident. The following section shows that the Galerkin solution in fact minimizes the energy norm of the error; in that sense, it is the "best" of all possible numerical solutions representable as a linear combination of a given set of approximating functions ψ.

3.2 The Weak Formulation and the Galerkin Method

In this section, the variational approach outlined above is cast in a more general and precise form; however, it does make sense to keep the last example (Example 2) in mind for concreteness. Let us consider a generic problem of the form

$$\mathcal{L}u \;=\; \rho, \quad u \in V = V(\Omega) \tag{3.18}$$

of which the Poisson equation (3.1) on p. 69 is a simple particular case. Here operator $\mathcal{L}$ is assumed to be self-adjoint with respect to a given inner product $(\cdot\,,\,\cdot)$ in the functional space V under consideration:

[6] In more complicated cases, numerical quadratures may be needed.

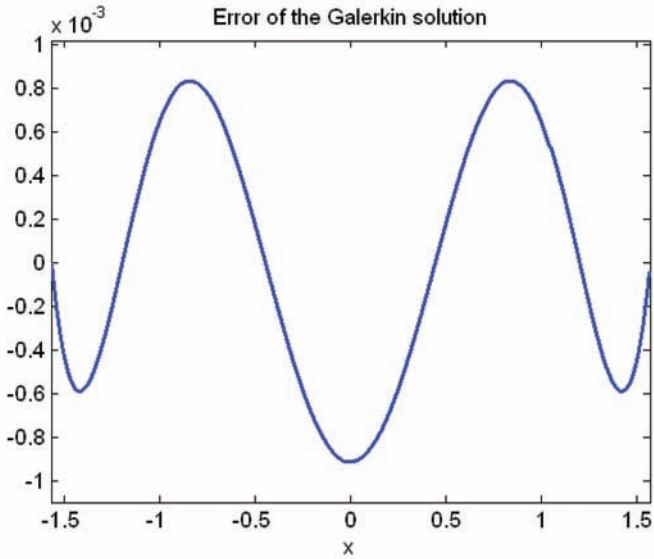


Fig. 3.3. Error of the Galerkin solution (3.7) in Example 2. (Note the 10^{-3} scaling factor.)

$$(\mathcal{L}u, v) \;=\; (u, \mathcal{L}v), \quad \forall u, v \in V \tag{3.19}$$

The reader unfamiliar with the notion of inner product may view it just as a shorthand notation for integration:

$$(w, v) \;\equiv\; \int_\Omega wv \, d\Omega$$

This definition is not general[7] but sufficient in the context of this section.

Note that operators defined in different functional spaces (or, more generally, in different domains) are mathematically different, even if they can be described by the same expression. For example, the Laplace operator in a functional space with zero boundary conditions is not the same as the Laplace operator in a space without such conditions. One manifestation of this difference is that the Laplace operator is self-adjoint in the first case but not so in the second.

Applying to the operator equation (3.18) inner product with an arbitrary function $v \in V$ (in the typical case, multiplying both sides with v and integrating), we obtain

$$(\mathcal{L}u, v) \;=\; (\rho, v), \quad \forall v \in V \tag{3.20}$$

[7] Generally, inner product is a bilinear (sesquilinear in the complex case) (conjugate-)symmetric positive definite form.

Clearly, this inner-product equation follows from the original one (3.18). At the same time, because v is arbitrary, it can be shown under fairly general mathematical assumptions that the converse is true as well: original equation (3.18) follows from (3.20); that is, these two formulations are equivalent (see also p. 84).

The left hand side of (3.20) is a bilinear form in u, v; in addition, if $\mathcal{L}$ is self-adjoint, this form is symmetric. This bilinear form will be denoted as $\mathcal{L}(u, v)$ (making symbol $\mathcal{L}$ slightly overloaded):

$$\mathcal{L}(u, v) \;\equiv\; (\mathcal{L}u, v), \quad \forall v \in V \tag{3.21}$$

To illustrate this definition: in Examples 1, 2 this bilinear form is

$$\mathcal{L}(u, v) \;\equiv\; -\int_{\frac{\pi}{2}}^{\frac{\pi}{2}} u''\, v\, dx \;=\; \int_{\frac{\pi}{2}}^{\frac{\pi}{2}} u'\, v'\, dx \tag{3.22}$$

The last integration-by-parts transformation appears innocuous but has profound consequences. It replaces the second derivative of u with the first derivative, thereby relaxing the required level of smoothness of the solution.

The following illustration is simple but instructive. Let u be a function with a "sharp corner" – something like $|x|$ in Fig. 3.4: it has a discontinuous first derivative and no second derivative (in the sense of regular calculus) at $x = 0$. However, this function can be approximated, with an arbitrary degree of accuracy, by a smooth one – it is enough just to "round off" the corner.

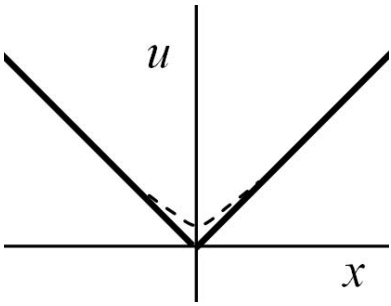


Fig. 3.4. Rounding off the corner provides a smooth approximation.

"Accuracy" here is understood in the energy-norm sense: if the smoothed function is denoted with $\tilde{u}$, then the approximation error is

$$\|\tilde{u} - u\|_E \;\equiv\; \left[\int_\Omega \left(\frac{d\tilde{u}}{dx} - \frac{du}{dx}\right)^2 dx\right]^{\frac{1}{2}} \tag{3.23}$$

where the precise specification of domain (segment) Ω is unimportant.

For the smooth function $\tilde{u}$, both expressions for the bilinear form (3.21) are valid and equal. For u, the first definition, containing u'' in the integrand, is not valid, but the second one, with u', is. It is quite natural to extend the definition of the bilinear form to functions that, while not necessarily smooth enough themselves, can be approximated arbitrarily well – in the energy norm sense – by smooth functions:

$$\mathcal{L}(u,v) \;\equiv\; \int_\Omega \frac{du}{dx}\frac{dv}{dx}\,d\Omega, \quad u\,,v \in H_0^1(\Omega) \tag{3.24}$$

Such functions form the Sobolev space $H^1(\Omega)$. The subspace $H_0^1(\Omega) \subset H^1(\Omega)$ contains functions with zero Dirichlet conditions at the boundary of domain Ω.[8]

Similarly, for the electrostatic equation (with the dielectric permittivity normalized to unity)

$$\mathcal{L}u \;\equiv\; -\nabla\cdot\epsilon\nabla u \;=\; \rho \tag{3.25}$$

in a two- or three-dimensional domain Ω with zero Dirichlet boundary conditions,[9] the weak formulation is

$$\mathcal{L}(u,v) \;\equiv\; (\epsilon\nabla u, \nabla v) \;=\; (\rho\,,v) \quad u\,,v \in H_0^1(\Omega) \tag{3.26}$$

where the parentheses denote the inner product of vector functions

$$(\mathbf{v},\mathbf{w}) \;\equiv\; \int_\Omega \mathbf{v}\cdot\mathbf{w}\,d\Omega, \quad \mathbf{v}\,,\mathbf{w} \in L_2^3(\Omega) \tag{3.27}$$

The analysis leading to the weak formulation (3.26) is analogous to the 1D case: the differential equation is inner-multiplied (i.e. multiplied and integrated) with a "test" function v; then integration by parts moves one of the ∇ operators over from u to v, so that the formulation can be extended to a broader class of admissible functions, with the smoothness requirements relaxed.

The weak formulation (3.20) (of which (3.26) is a typical example) provides a very natural way of approximating the problem. All that needs to be done is to restrict both the unknown function u and the test function v in (3.20) to a finite-dimensional subspace $V_h \subset V$:

$$\mathcal{L}(u_h, v_h) \;=\; (\rho, v_h), \quad \forall v_h \in V_h(\Omega) \tag{3.28}$$

In Examples 1 and 2 space V_h had just two dimensions; in engineering practice, the dimension of this space can be on the order of hundreds of thousands and

[8] The rigorous mathematical characterization of "boundary values" (more precisely, *traces*) of functions in Sobolev spaces is quite involved. See R.A. Adams [AF03] or K. Rektorys [Rek80].

[9] Neumann conditions on the domain boundary and interface boundary conditions between different media will be considered later.

even millions. Also in practice, construction of V_h typically involves a mesh (this was not the case in Examples 1 and 2, but will be the case in the subsequent sections in this chapter); then subscript "h" indicates the mesh size. If a mesh is not used, h can be understood as some small parameter; in fact, one usually has in mind a *family* of spaces V_h that can approximate the solution of the problem with arbitrarily high accuracy as $h \to 0$.

Let us assume that an approximating space V_h of dimension n has been chosen and that ψ_α ($\alpha = 1, \ldots, n$) is a known basis set in this space. Then the approximate solution is a linear combination of the basis functions:

$$u_h \;=\; \sum_{\alpha=1}^{n} \underline{u}_\alpha \psi_\alpha \tag{3.29}$$

Here $\underline{u}$ is a Euclidean vector of coefficients in $\mathbb{R}^n$ (or, in the case of problems with complex solutions, in $\mathbb{C}^n$).

This expansion establishes an intimate relationship between the functional space V_h to which u_h belongs and the Euclidean space of coefficient vectors $\underline{u}$. If functions ψ_α are linearly independent, there is a one-to-one correspondence between u_h and $\underline{u}$. Moreover, the bilinear form $\mathcal{L}(u_h, u_h)$ induces an equivalent bilinear form over Euclidean vectors:

$$(L\underline{u}, \underline{v}) \;=\; \mathcal{L}(u_h, v_h) \tag{3.30}$$

for any two functions $u_h, v_h \in V_h$ and their corresponding Euclidean vectors $\underline{u}$, $\underline{v} \in \mathbb{R}^n$. The left hand side of (3.30) is the usual Euclidean inner product of vectors, and L is a square matrix. From basic linear algebra, each entry $L_{\alpha\beta}$ of this matrix is equal to (Le_α, e_β), where e_α is column $\#\alpha$ of the identity matrix (the only nonzero entry $\#\alpha$ is equal to one); similarly for e_β. At the same time, (Le_α, e_β) is, by definition of L, equal to the bilinear form involving ψ_α, ψ_β; hence

$$L_{\alpha\beta} \;=\; (Le_\alpha, e_\beta) \;=\; \mathcal{L}(\psi_\alpha, \psi_\beta) \tag{3.31}$$

The equivalence of bilinear forms (3.30) is central in Galerkin methods in general and FEM in particular; it can also be viewed as an operational definition of matrix L. Explicitly the entries of L are defined by the right hand side of (3.31). Example 3 below should clarify this matter further.

The Galerkin formulation (3.28) is just a restriction of the weak continuous formulation to a finite-dimensional subspace, and therefore the numerical bilinear form inherits the algebraic properties of the continuous one. In particular, if the bilinear form $\mathcal{L}$ is elliptic, i.e. if

$$\mathcal{L}(u, u) \;\geq\; c\,(u, u), \qquad \forall u \in V \quad (c > 0) \tag{3.32}$$

where c is a constant, then matrix L is strictly positive definite and, moreover,

$$(L\underline{u}, \underline{u}) \;\geq\; c\,(M\underline{u}, \underline{u}), \qquad \forall \underline{u} \in \mathbb{R}^n \tag{3.33}$$

Matrix M is such that the Euclidean form $(M\underline{u}, \underline{v})$ corresponds to the L_2 inner product of the respective functions:

$$(M\underline{u}, \underline{v}) \;=\; (u_h, v_h) \tag{3.34}$$

so that the entries are

$$M_{\alpha\beta} \;=\; (\psi_\alpha, \psi_\beta) \tag{3.35}$$

These expressions for matrix M are analogous to expressions (3.30) and (3.31) for matrix L. In FEM, M is often called the *mass matrix* and L – the *stiffness matrix*, due to the roles they play in problems of structural mechanics where FEM originated.

Example 3. To illustrate the connection between Euclidean inner products and the respective bilinear forms of functions, let us return to Example 2 on p. 73 and choose the two coefficients arbitrarily as $\underline{u}_1 = 2$, $\underline{u}_2 = -1$. The corresponding function is

$$u_h \;=\; \underline{u}_1\psi_1 + \underline{u}_2\psi_2 \;=\; 2\left(x - \frac{\pi}{2}\right)\left(x + \frac{\pi}{2}\right) \;-\; \left(x - \frac{\pi}{2}\right)^2\left(x + \frac{\pi}{2}\right)^2 \tag{3.36}$$

This function of course lies in the two-dimensional space V_h spanned by $\psi_{1,2}$.

Similarly, let $\underline{v}_1 = 4$, $\underline{v}_2 = -3$ (also as an arbitrary example); then

$$v_h \;=\; \underline{v}_1\psi_1 + \underline{v}_2\psi_2 \;=\; 4\left(x - \frac{\pi}{2}\right)\left(x + \frac{\pi}{2}\right) \;-\; 3\left(x - \frac{\pi}{2}\right)^2\left(x + \frac{\pi}{2}\right)^2 \tag{3.37}$$

In the left hand side of (3.30), matrix L was calculated to be (3.16), and the Euclidean inner product is

$$(L\underline{u}, \underline{v}) \;=\; \left(\frac{\pi^3}{105}\begin{pmatrix} 35 & -7\pi^2 \\ -7\pi^2 & -2\pi^4 \end{pmatrix}\begin{pmatrix} 2 \\ -1 \end{pmatrix}, \begin{pmatrix} 4 \\ -3 \end{pmatrix}\right) \;=\; \frac{8\pi^3}{3} + \frac{2\pi^5}{3} + \frac{2\pi^7}{35} \tag{3.38}$$

The right hand side of (3.30) is

$$\int_{\frac{\pi}{2}}^{\frac{\pi}{2}} u_h'\, v_h'\, dx$$

where functions u_h, v_h are given by their expansions (3.36), (3.37). Substitution of these expansions into the integrand above yields exactly the same result as the right hand side of (3.38), namely

$$\frac{8\pi^3}{3} + \frac{2\pi^5}{3} + \frac{2\pi^7}{35}$$

This illustrates that the Euclidean inner product of vectors $\underline{u}$, $\underline{v}$ in (3.30) (of which the left hand side of (3.38) is a particular case) is equivalent to the bilinear form $\mathcal{L}(u, v)$ of functions u, v (of which the *right* hand side of (3.38) is a particular case).

By setting v_h consecutively to $\psi_1, \psi_2, \ldots, \psi_n$ in (3.28), one arrives at the following matrix-vector form of the variational formulation (3.28):

$$L\underline{u} \;=\; \underline{\rho} \tag{3.39}$$

with

$$L_{\alpha\beta} \;=\; \mathcal{L}(\psi_\alpha, \psi_\beta); \qquad \underline{\rho}_\alpha \;=\; (\rho, \psi_\alpha) \tag{3.40}$$

This is a direct generalization of system (3.14) on p. 74.

3.3 Variational Methods and Minimization

3.3.1 The Galerkin Solution Minimizes the Error

The analysis in this section is restricted to operator $\mathcal{L}$ that is self-adjoint in a given functional space V, and the corresponding symmetric (conjugate-symmetric in the complex case) form $\mathcal{L}(u, v)$. In addition, if

$$\mathcal{L}(u, u) \;\geq\; c(u, u), \quad \forall u \in V \tag{3.41}$$

for some positive constant c, the form is called *elliptic* (or, synonymously, *coercive*).

The weak continuous problem is

$$\mathcal{L}(u, v) \;=\; (\rho, v), \quad u \in V; \;\; \forall v \in V \tag{3.42}$$

We shall assume that this problem has a unique solution $u^* \in V$ and shall refer to u^* as *the exact solution* (as opposed to a numerical one). Mathematical conditions for the existence and uniqueness are cited in Section 3.5.

The numerical Galerkin problem is obtained by restricting this formulation to a finite-dimensional subspace $V_h \subset V$:

$$\mathcal{L}(u_h, v_h) \;=\; (\rho, v_h), \quad u_h \in V_h; \;\; \forall v_h \in V_h \tag{3.43}$$

where u_h is the numerical solution. Keep in mind that u_h solves the Galerkin problem in the finite-dimensional subspace V_h only; in the full space V there is, in general, a nonzero *residual*

$$\mathcal{R}(u_h, v) \;\equiv\; (\rho, v) \;-\; \mathcal{L}(u_h, v) \;=\; , \quad v \in V \tag{3.44}$$

In matrix-vector form, this problem is

$$L\underline{u} \;=\; \underline{\rho} \tag{3.45}$$

with matrix L and the right hand side $\underline{\rho}$ defined in (3.40). If matrix L is nonsingular, a unique numerical solution exists. For an elliptic form $\mathcal{L}$ – a

particularly important case in theory and practice – matrix L is positive definite and hence nonsingular.

The numerical error is

$$\epsilon_h \;=\; u_h \;-\; u \tag{3.46}$$

A remarkable property of the Galerkin solution for a symmetric form $\mathcal{L}$ is that it minimizes the error functional

$$\mathcal{E}(u_h) \;\equiv\; \mathcal{L}(\epsilon_h,\, \epsilon_h) \;\equiv\; \mathcal{L}(u_h - u,\, u_h - u) \tag{3.47}$$

In other words, of all functions in the finite-dimensional space V_h, the Galerkin solution u_{hG} is the best approximation of the exact solution in the sense of measure (3.47). For coercive forms $\mathcal{L}$, this measure usually has the physical meaning of energy.

To prove this minimization property, let us analyze the behavior of functional (3.47) in the vicinity of some u_h – that is, examine $\mathcal{E}(u_h + \lambda v_h)$, where $v_h \in V_h$ is an increment and λ is an adjustable numerical factor introduced for mathematical convenience. (This factor could be absorbed into v_h but, as will soon become clear, it makes sense *not* to do so. Also, λ can be intuitively understood as "small" but this has no bearing on the formal analysis.) Then, assuming a real form for simplicity,

$$\mathcal{E}(u_h+\lambda v_h) \;=\; \mathcal{L}(\epsilon_h+\lambda v_h, \epsilon_h+\lambda v_h) \;=\; \mathcal{L}(\epsilon_h, \epsilon_h) + 2\lambda\mathcal{L}(\epsilon_h, v_h) + \lambda^2\mathcal{L}(v_h, v_h) \tag{3.48}$$

At a stationary point of $\mathcal{E}$ – and in particular at a maximum or minimum – the term linear in λ must vanish:

$$\mathcal{L}(\epsilon_h, v_h) \;=\; 0, \quad \forall v_h \in V_h$$

This condition is nothing other than

$$\mathcal{L}(u_h, v_h) \;=\; \mathcal{L}(u, v_h) \;=\; (f, v_h)$$

(The last equality follows from the fact that u is the solution of the weak problem.) This is precisely the Galerkin equation.

Thus the Galerkin solution is a stationary point of functional (3.47). If the bilinear form $\mathcal{L}$ is elliptic, expression (3.48) for the variation of the energy functional then indicates that this stationary point is in fact a minimum: the term linear in λ vanishes and the quadratic term is positive for a nonzero v_h.

3.3.2 The Galerkin Solution and the Energy Functional

Error minimization (in the energy norm sense) is a significant strength of the Galerkin method. A practical limitation of the error functional (3.47), however, is that it cannot be computed explicitly: this functional depends on the exact solution that is unknown. At the same time, for self-adjoint problems there is another – and computable – functional for which both the exact

solution (in the original functional space V) and the numerical solution (in the chosen finite-dimensional space V_h) are stationary points. This functional is

$$\mathcal{F}u \;=\; (\rho, u) \;-\; \frac{1}{2}\,\mathcal{L}(u,u), \quad u \in V \tag{3.49}$$

Indeed, for an increment λv, where λ is an arbitrary number and $v \in V$, we have

$$\Delta\mathcal{F} \;\equiv\; \mathcal{F}(u+\lambda v) - \mathcal{F}u \;=\; (\rho, \lambda v) - \frac{1}{2}\,\mathcal{L}(\lambda v, u) - \frac{1}{2}\,\mathcal{L}(u, \lambda v) - \frac{1}{2}\,\mathcal{L}(\lambda v, \lambda v)$$

which for a symmetric real form $\mathcal{L}$ is

$$\Delta\mathcal{F} \;=\; \lambda[(\rho, v) \;-\; \mathcal{L}(u,v)] \;-\; \frac{1}{2}\,\lambda^2\mathcal{L}(v,v)$$

The zero linear term in λ thus corresponds precisely to the weak formulation of the problem. By a very similar argument, the Galerkin solution is a stationary point of $\mathcal{F}$ in V_h. Furthermore, if the bilinear form $\mathcal{L}$ is elliptic, the quadratic term $\lambda^2\mathcal{L}(v,v)$ is nonnegative, and the stationary point is *a maximum.*

In electrostatics, magnetostatics and other physical applications functional $\mathcal{F}$ is often interpreted as energy. It is indeed equal to field energy if u is the exact solution of the underlying differential equation (or, almost equivalently, of the weak problem). Other values of u are not physically realizable, and hence $\mathcal{F}$ in general lacks physical significance as energy and should rather be interpreted as "action" (an integrated Lagrangian). It is not therefore paradoxical that the solution *maximizes* – not minimizes – the functional.[10] This matter is taken up again in Section 6.11 on p. 328 and in Appendix 6.14 on p. 338 in the context of electrostatic simulation.

Functional $\mathcal{F}$ (3.49) is part of a broader picture of complementary variational principles; see the book by A.M. Arthurs [Art80] (in particular, examples in Section 1.4 of his book[11]).

3.4 Essential and Natural Boundary Conditions

So far, for brevity of exposition, only Dirichlet conditions on the exterior boundary of the domain were considered. Now let us turn our attention to

[10] One could reverse the sign of $\mathcal{F}$, in which case the stationary point would be a *minimum.* However, this functional would no longer have the meaning of field energy, as its value at the exact solution u would be negative, which is thermodynamically impossible for electromagnetic energy (see L.D. Landau & E.M. Lifshitz [LL84].

[11] A note for the reader interested in the Arthurs book and examples therein. In the electrostatic case, the quantities in these examples are interpreted as follows: $U \equiv D$ (the electrostatic displacement field), $v \equiv \epsilon$ (the permittivity), $\Phi = u$ (potential), $q \equiv \rho$ (charge density).

quite interesting, and in practice very helpful, circumstances that arise if conditions on part of the boundary are left *unspecified* in the weak formulation.

We shall use the standard electrostatic equation in 1D, 2D or 3D as a model:

$$-\nabla \cdot \epsilon \nabla u \;=\; \rho \quad \text{in } \Omega; \quad u = 0 \text{ on } \partial\Omega_D \subset \partial\Omega \tag{3.50}$$

At first, the dielectric permittivity ϵ will be assumed a smooth function of coordinates; later, we shall consider the case of piecewise-smooth ϵ (e.g. dielectric bodies in a host medium). Note that u satisfies the zero Dirichlet condition only *on part of* the domain boundary; the condition on the remaining part is left unspecified for now, so the boundary value problem is not yet fully defined.

The weak formulation is

$$(\epsilon \nabla u, \nabla v) \;=\; (\rho, v), \qquad u, \forall v \in H_0^1(\Omega, \partial\Omega_D) \tag{3.51}$$

$H_0^1(\Omega, \partial\Omega_D)$ is the Sobolev space of functions that have a generalized derivative and satisfy the zero Dirichlet condition on $\partial\Omega_D$.[12]

Let us now examine, a little more carefully than we did before, the relationship between the weak problem (3.51) and the differential formulation (3.50). To convert the weak problem into a strong one, one integrates the left hand side of (3.51) by parts:

$$\int_{\partial\Omega - \partial\Omega_D} v\, \epsilon \frac{\partial u}{\partial n}\, dS \;-\; (\nabla \cdot \epsilon \nabla u, v) \;=\; (\rho, v) \tag{3.52}$$

It is tacitly assumed that u is such that the differential operator $\nabla \cdot \epsilon \nabla u$ makes sense. Note that the surface integral is taken over the non-Dirichlet part of the boundary only, as the "test" function v vanishes on the Dirichlet part by definition.

The key observation is that v is arbitrary. First, as a particular choice, let us consider test functions v vanishing on the domain boundary. In this case, the surface integral in (3.52) disappears, and we have

$$(\nabla \cdot \epsilon \nabla u \,+\, \rho,\; v) \;\equiv\; \int_\Omega v(\nabla \cdot \epsilon \nabla u \,+\, \rho)\, d\Omega \;=\; 0 \tag{3.53}$$

This may hold true for *arbitrary* v only if the integrand

$$I \;\equiv\; \nabla \cdot \epsilon \nabla u \,+\, \rho \tag{3.54}$$

in (3.53) is identically zero. The proof, at least for continuous I, is simple. Indeed, if I were strictly positive at some point r_0 inside the domain, it would,

[12] These are functions that are either smooth themselves or can be approximated by smooth functions, in the H^1-norm sense, with any degree of accuracy. Boundary values, strictly speaking, should be considered in the sense of traces (R.A. Adams & J.J.F. Fournier [AF03], K. Rektorys [Rek80]).

by continuity, have to be positive in some neighborhood of that point. By choosing the test function that is positive in the same neighborhood and zero elsewhere (imagine a sharp but smooth peak centered at r_0 as such a test function), one arrives at a contradiction, as the integral in (3.53) is positive rather than zero.

This argument shows that the Poisson equation must be satisfied for the solution u of the weak problem. Further observation can be made if we now consider a test function that is nonzero on the non-Dirichlet part of the boundary. In the integrated-by-parts weak formulation (3.52), the volume integrals, as we now know, must vanish if u is the solution, because the Poisson equation is satisfied. Then we have

$$\int_{\partial\Omega-\partial\Omega_D} v\,\epsilon\,\frac{\partial u}{\partial n}\,dS \;=\; 0 \tag{3.55}$$

Since v is arbitrary, the integrand must be identically zero – the proof is essentially the same as for the volume integrand I in (3.54). We come to the conclusion that solution u must satisfy the Neumann boundary condition

$$\frac{\partial u}{\partial n} \;=\; 0 \tag{3.56}$$

on the non-Dirichlet part of the domain boundary (for $\epsilon \neq 0$).

This is really a notable result. In the weak formulation, if no boundary condition is explicitly imposed on part of the boundary, then the solution will satisfy the Neumann condition. Such "automatic" boundary conditions that follow from the weak formulation are called *natural*. In contrast, conditions that have to be imposed explicitly are called *essential*. Dirichlet conditions are essential.

For cases other than the model electrostatic problem, a similar analysis is needed to identify natural boundary conditions. As a rule of thumb, conditions containing the normal derivative at the boundary are natural. For example, Robin boundary conditions (a combination of values of u and its normal derivative) are natural.

Importantly, the continuity of flux $\epsilon\,\partial u/\partial n$ across material interfaces is also a natural condition. The analysis is similar to that of the Neumann condition. Indeed, let Γ be the boundary between materials #1,2 with their respective parameters $\epsilon_{1,2}$. Separately within each material, ϵ varies smoothly, but a jump may occur across Γ.

With the weak problem (3.51) taken as a starting point, integration by parts yields

$$\int_{\partial\Omega-\partial\Omega_D}[\ldots] \;+\; \int_{\Gamma} v\left[\left(\epsilon\,\frac{\partial u}{\partial n}\right)_1 - \left(\epsilon\,\frac{\partial u}{\partial n}\right)_2\right] dS \;-\; (\nabla\cdot\epsilon\nabla u, v) \;=\; (\rho, v) \tag{3.57}$$

Subscripts 1 and 2 indicate that the respective electric flux density $\epsilon\,\partial u/\partial n$ is taken in materials 1, 2; n is the unit normal to Γ, directed into material #2

(this choice of direction is arbitrary). The integrand on the exterior boundary is omitted for brevity, as it is the same as considered previously and leads, as we already know, to the Neumann boundary condition on $\Omega - \partial\Omega_D$.

Consider first the volume integrals (inner products) in (3.57). Using the fact that v is arbitrary, one can show in exactly the same way as before that the electrostatic differential equation must be satisfied throughout the domain, except possibly for the interface boundary where the differential operator may not be valid in the sense of ordinary calculus. Turning then to the surface integral over Γ and again noting that v is arbitrary on that surface, one observes that the integrand – i.e. the flux jump – across the surface must be zero if u is the solution of the weak problem.

This is a great practical advantage because no special treatment of material interfaces is needed. For the model electrostatic problem, the finite element algorithm for heterogeneous media is essentially the same as for the homogeneous case. However, for more complicated problems interface conditions may need special treatment and may result in additional surface integrals.[13]

It is in principle possible to impose natural conditions explicitly – that is, incorporate them into the definition of the functional space and choose the approximating and test functions accordingly. However, this is usually inconvenient and redundant, and therefore is hardly ever done in practice.

3.5 Mathematical Notes: Convergence, Lax–Milgram and Céa's Theorems

This section summarizes some essential facts about weak formulations and convergence of Galerkin solutions. The mathematical details and proofs are omitted, one exception being a short and elegant proof of Céa's theorem. There are many excellent books on the mathematical theory: an elaborate exposition of variational methods by K. Rektorys [Rek80] and by S.G. Mikhlin [Mik64, Mik65], as well as the well-known text by R. Weinstock [Wei74]; classical monographs on FEM by P.G. Ciarlet [Cia80], by B. Szabó & I. Babuška [SB91], and a more recent book by S.C. Brenner & L.R. Scott [BS02], among others.

Those readers who are not interested in the mathematical details may skip this section – a digest of the underlying mathematical theory – without substantial harm to their understanding of the rest of the chapter.

Theorem 2. *(***Lax–Milgram***.) [BS02, Rek80]*

[13] One interesting example is a hybrid formulation of eddy current problems, with the magnetic vector potential inside a conducting body and the magnetic scalar potential outside. The weak formulation contains a surface integral on the boundary of the conductor. The interested reader may see C.R.I. Emson & J. Simkin [ES83], D. Rodger [Rod83] for the formulation and [Tsu90] for a mathematical analysis.

Given a Hilbert space V, a continuous and elliptic bilinear form $\mathcal{L}(\cdot\,,\,\cdot)$ and a continuous linear functional $f \in V'$, there exists a unique $u \in V$ such that

$$\mathcal{L}(u,v) \;=\; f(v), \qquad \forall v \in V \tag{3.58}$$

As a reminder, a bilinear form is *elliptic* if

$$\mathcal{L}(u,u) \;\geq\; c_1(u,u), \qquad \forall u \in V$$

and continuous if

$$\mathcal{L}(u,v) \;\leq\; c_2\|u\|\,\|v\|, \qquad \forall u,v \in V$$

for some positive constants $c_{1,2}$. Here the norm is induced by the inner product:

$$\|v\| \;\equiv\; (v,v)^{\frac{1}{2}} \tag{3.59}$$

Finally in the formulation of the Lax–Milgram theorem, V' is the space of continuous linear functionals over V. A linear functional is continuous if $f(v) \leq c\|v\|$, where c is some constant.

The reason why the Lax–Milgram theorem is important is that its conditions correspond to the weak formulations of many problems of mathematical physics, including the model electrostatic problem of the previous section. The Lax–Milgram theorem establishes uniqueness and existence of the (exact) solution of such problems. Under the Lax–Milgram conditions, it is clear that uniqueness and existence also hold in any subspace of V – in particular, for the approximate Galerkin solution.

The Lax–Milgram theorem can be proved easily for *symmetric* forms. Indeed, if $\mathcal{L}$ is symmetric (in addition to its continuity and ellipticity required by the conditions of the theorem), this form represents an inner product in V: $[u,v] \equiv \mathcal{L}(u,v)$. Then $f(v)$, being a linear continuous functional, can be by the Riesz Representation Theorem (one of the basic properties of Hilbert spaces) expressed via this new inner product as $f(v) = [u,v] \equiv \mathcal{L}(u,v)$, which is precisely what the Lax–Milgram theorem states. The more complicated proof for nonsymmetric forms is omitted.

Theorem 3. *(**Céa**) [BS02, Rek80]*

Let V be a subspace of a Hilbert space H and $\mathcal{L}(\cdot\,,\,\cdot)$ be a continuous elliptic (but not necessarily symmetric) bilinear form on V. Let $u \in V$ be the solution of equation (3.58) from the Lax–Milgram theorem. Further, let u_h be the solution of the Galerkin problem

$$\mathcal{L}(u_h,v_h) \;=\; f(v_h), \qquad \forall v_h \in V_h \tag{3.60}$$

in some finite-dimensional subspace $V_h \subset V$. Then

$$\|u-u_h\| \;\leq\; \frac{c_2}{c_1}\min_{v\in V_h}\|u-v\| \tag{3.61}$$

where c_1 and c_2 are the ellipticity and continuity constants of the bilinear form $\mathcal{L}$.

Céa's theorem is a principal result, as it relates the error of the Galerkin solution to the approximation error. The latter is much more easily amenable to analysis: good *approximation* can be produced by various forms of interpolation, while the *Galerkin solution* emerges from solving a large system of algebraic equations. For a symmetric form $\mathcal{L}$ and for the norm induced by $\mathcal{L}$, constants $c_{1,2} = 1$ and the Galerkin solution is best in the energy-norm sense, as we already know.

Proof. The error of the Galerkin solution is

$$\epsilon_h \;\equiv\; u_h \,-\, u, \qquad u_h \in V_h \tag{3.62}$$

where u is the (exact) solution of the weak problem (3.58) and u_h is the solution of the Galerkin problem (3.60). This error itself satisfies a weak problem obtained simply by subtracting the Galerkin equation from the exact one:

$$\mathcal{L}(\epsilon_h, v_h) \;=\; 0, \qquad \forall v_h \in V_h \tag{3.63}$$

This can be interpreted as a generalized orthogonality relationship: the error is "$\mathcal{L}$-orthogonal" to V_h. (If $\mathcal{L}$ is not symmetric, it does not induce an inner product, so the *standard* definition of orthogonality does not apply.) Such an interpretation has a clear geometric meaning: the Galerkin solution is a projection (in a proper sense) of the exact solution onto the chosen approximation space.

Then we have

$$\mathcal{L}(\epsilon_h, \epsilon_h) \;\equiv\; \mathcal{L}(\epsilon_h, u_h - u) \;=\; \mathcal{L}(\epsilon_h, u_h - v_h - u), \;\equiv\; \mathcal{L}(\epsilon_h, w_h - u); \quad v_h \in V_h$$

The first identity is trivial, as it reiterates the definition of the error. The second equality is crucial and is due to the generalized orthogonality (3.63). The last identity is just a variable change, $w_h = u_h - v_h$.

Using now the ellipticity and continuity of the bilinear form, we get

$$c_1 \|\epsilon_h\|_2^2 \;=\; c_1(\epsilon_h, \epsilon_h) \;\leq\; \mathcal{L}(\epsilon_h, \epsilon_h) \;=\; \mathcal{L}(\epsilon_h, w_h \,-\, u) \;\leq\; c_2 \|\epsilon_h\| \, \|w_h \,-\, u\|$$

which, after dividing through by $\|\epsilon_h\|$, yields precisely the result of Céa's theorem:

$$c_2 \|\epsilon_h\| \;\leq\; c_1 \|w_h \,-\, u\|$$

□

Céa's theorem simplifies error analysis greatly: it is in general extremely difficult to evaluate the Galerkin error directly because the Galerkin solution emerges as a result of solving a (usually large) system of equations; it is much easier to deal with *some* good approximation w_h of the exact solution (e.g. via an interpolation procedure). Céa's theorem relates the Galerkin solution error to the approximation error via the stability and continuity constants of the bilinear form.

From a practical point of view, Céa's theorem is the source of robustness of the Galerkin method. In fact, the Galerkin method proves to be surprisingly reliable even for non-elliptic forms: although Céa's theorem is silent about that case, a more general result known as the Ladyzhenskaya–Babuška–Brezzi (or just LBB) condition[14] is available (O.A. Ladyzhenskaya [Lad69], I. Babuška, [Bab58], F. Brezzi [Bre74]; see also B. Szabó & I. Babuška [SB91], I. Babuška & T. Strouboulis [BS01] and Appendix 3.10).

3.6 Local Approximation in the Finite Element Method

Remember the shortcomings of collocation – the first variational technique to be introduced in this chapter? The Galerkin method happily resolves (at least for elliptic problems) two of the three issues listed on p. 72. Indeed, the way to choose the test functions is straightforward (they are the same as the approximating functions), and Céa's theorem provides an error bound for the Galerkin solution.

The only missing ingredient is a procedure for choosing "good" approximating functions. The Finite Element Method does provide such a procedure, and the following sections explain how it works in one, two and three dimensions.

The guiding principle is *local* approximation of the solution. This usually makes perfect physical sense. It is true that in a limited number of cases a global approximation over the whole computational domain is effective – these cases usually involve homogeneous media with a smooth distribution of sources or no sources at all, with the field approximated by a Fourier series or a polynomial expansion. However, in practical problems, *local* geometric and physical features of systems and devices, with the corresponding local behavior of fields and potentials, is typical. Discontinuities at material interfaces, peaks, boundary layers, complex behavior at edges and corners, and many other features make it all but impossible to approximate the solution globally.[15]

Local approximation in FEM is closely associated with a mesh: the computational domain is subdivided into small subdomains – *elements*. A large assortment of geometric shapes of elements can be used: triangular or quadrilateral are most common in 2D, tetrahedral and hexahedral – in 3D. Note that the term "element" is overloaded: depending on the context, it may mean just the geometric figure or, in addition to that, the relevant approximating space and degrees of freedom (more about that later). For example, linear and

[14] Occasionally used with some permutations of the names.

[15] Analytical approximations over homogeneous *subdomains*, with proper matching conditions at the interfaces of these subdomains, can be a viable alternative but is less general than FEM. One example is the Multiple Multipole Method popular in some areas of high frequency electromagnetic analysis and optics; see e.g. T. Wriedt (ed.), [Wri99].

quadratic approximations over a triangle give rise to different finite elements in the sense of FEM, even though the geometric figure is the same.

For illustration, Fig. 3.5 – Fig. 3.7 present FE meshes for a few particles of arbitrary shapes – the first two of these figures in 2D, and the third one in 3D. The mesh in the second figure (Fig. 3.6) was obtained by *global refinement* of the mesh in the first figure: each triangular element was subdivided into four. Mesh refinement can be expected to produce a more accurate numerical solution, albeit at a higher computational cost. *Global* refinement is not the most effective procedure: a smarter way is to make an effort to identify the areas where the numerical solution is least accurate and refine the mesh there. This idea leads to *local adaptive mesh refinement* (Section 3.13).

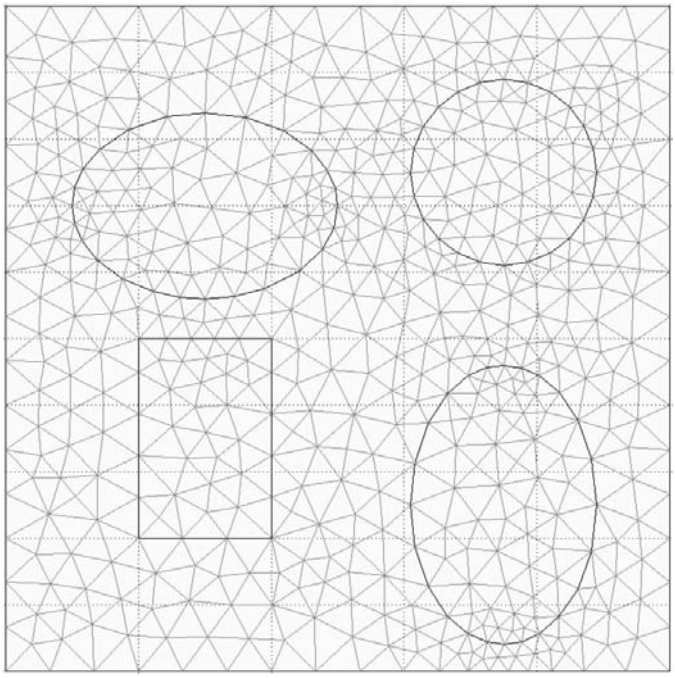


Fig. 3.5. An illustrative example of a finite element mesh in 2D.

Each approximating function in FEM is nonzero only over a small number of adjacent elements and is thus responsible for local approximation without affecting the approximation elsewhere. The following sections explain how this is done.

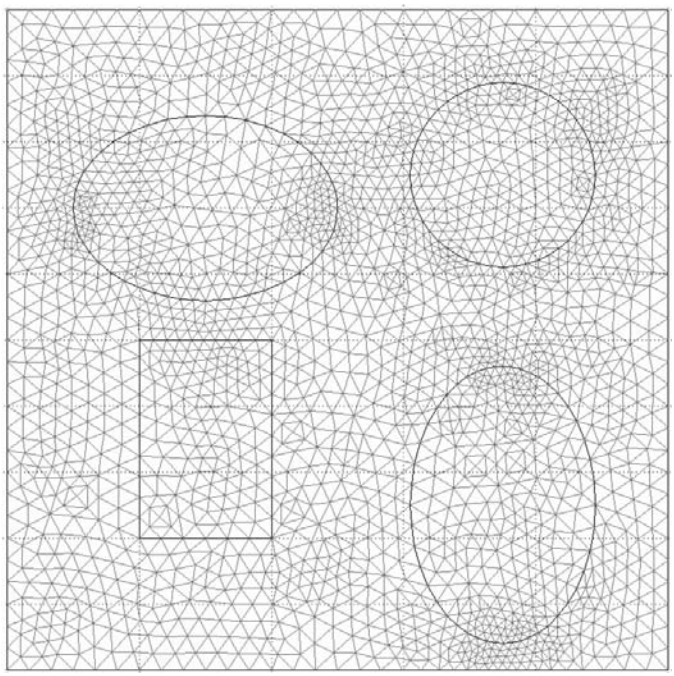


Fig. 3.6. Global refinement of the mesh of Fig. 3.5, with each triangular element subdivided into four by connecting the midpoints of the edges.

3.7 The Finite Element Method in One Dimension

3.7.1 First-Order Elements

In one dimension, the computational domain is a segment $[a, b]$, the mesh is a set of nodes $x_0 = a$, $x_1, \ldots, x_n = b$, and the elements (in the narrow geometric sense) are the segments $[x_{i-1}, x_i]$, $i = 1,2, \ldots, n$. The simplest approximating function is shown in Fig. 3.8 and is commonly called a "hat function" or, much less frequently, a "tent function".[16] The hat functions form a convenient basis of the simplest finite element vector space, as discussed in more detail below.

For notational convenience only, we shall often assume that the grid is uniform, i.e. the grid size $h = x_i - x_{i-1}$ is the same for all nodes i. For nonuniform grids, there are no conceptual changes and only trivial differences in the algebraic expressions. A formal expression for ψ_i on a uniform grid is

$$\psi_i(x) \;=\; \begin{cases} h^{-1}\,(x - x_{i-1}), & x_{i-1} \le x \le x_i \\ h^{-1}\,(x_{i+1} - x), & x_i \le x \le x_{i+1} \\ 0 \quad \text{otherwise} \end{cases} \tag{3.64}$$

[16] About 50 times less, according to Google. "Hut function" also makes some intuitive sense but is used very infrequently.

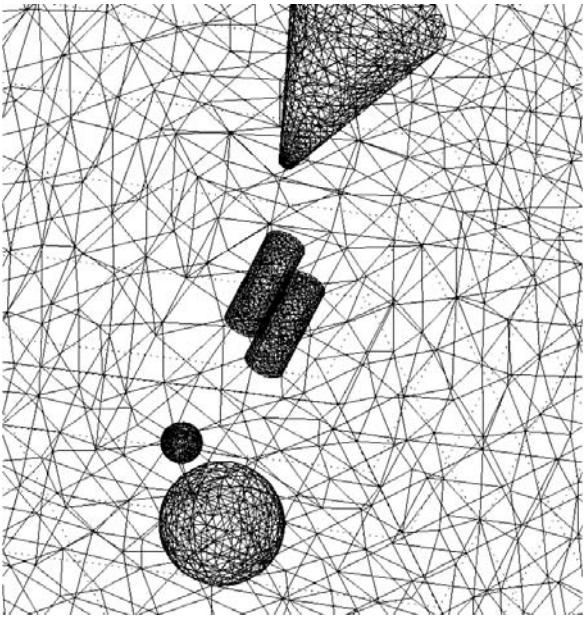


Fig. 3.7. An example of a finite element mesh in 3D.

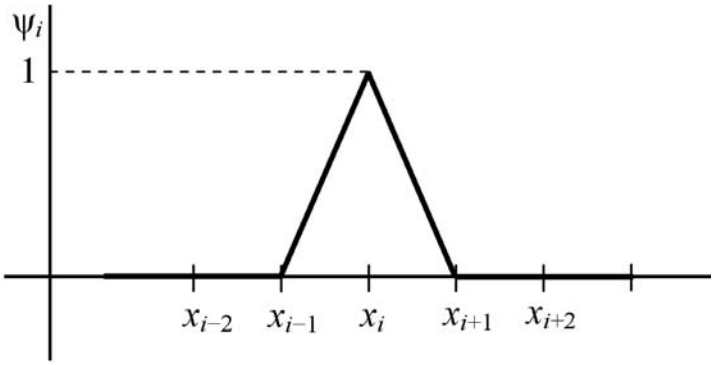


Fig. 3.8. The "hat" function for first order 1D elements.

The hat function ψ_i straddles two adjacent elements (segments) and satisfies the obvious Kronecker-delta property on the grid: it is equal to one at x_i and zero at all other nodes. This property is not critical in theoretical analysis but is very helpful in practice. In particular, for any smooth function $u(x)$, piecewise-linear interpolation on the grid can be written simply as the linear combination

$$u_{\text{interp}}(x) \;=\; \sum_{i=1}^{n} u(x_i)\,\psi_i$$

Indeed, the fact that the nodal values of u and u_{interp} are the same follows directly from the Kronecker-delta property of the ψs.

We now have all the prerequisites for solving an example problem.

Example 4.

$$-\frac{d^2u}{dx^2} \;=\; \sin x, \qquad \Omega = [0,\pi], \qquad u(0) = u(\pi) = 0 \tag{3.65}$$

The obvious theoretical solution $u(x) = \sin x$ is available for evaluating the accuracy of the finite element result.

Let us use a uniform grid $x_0 = 0$, $x_1 = h, \ldots, x_n = \pi$ with the grid size $h = \pi/n$. In numerical experiments, the number of nodes will vary, and we can expect higher accuracy (at higher computational cost) for larger values of n.

The weak formulation of the problem is

$$\int_0^\pi \frac{du}{dx}\,\frac{dv}{dx}\,dx \;=\; \int_0^\pi \sin x\, v(x)\,dx, \qquad u, \forall v \in H_0^1([0,\pi]) \tag{3.66}$$

The FE-Galerkin formulation is simply a restriction of the weak problem to the subspace $\mathcal{P}_{0h}([0,\pi])$ of piecewise-linear functions satisfying zero Dirichlet conditions; this is precisely the subspace spanned by the hat functions $\psi_1, \ldots, \psi_{n-1}$:[17]

$$\int_0^\pi \frac{du_h}{dx}\,\frac{dv_h}{dx}\,dx \;=\; \int_0^\pi v_h(x)\,\sin x\,dx, \qquad u_h, \forall v_h \in \mathcal{P}_{0h}([0,\pi]) \tag{3.67}$$

As we know, this formulation can be cast in matrix-vector form by substituting the expansion $\sum_{i=1}^{n-1} u_{hi}\psi_i$ for u_h and by setting v_h, sequentially, to $\psi_1, \ldots, \psi_{n-1}$ to obtain $(n-1)$ equations for $(n-1)$ unknown nodal values u_{hi}:

$$L\underline{u} \;=\; \underline{f}, \qquad \underline{u}, \underline{f} \in \mathbb{R}^{n-1} \tag{3.68}$$

where, as we also know, the entries of matrix L and the right hand side $\underline{f}$ are

[17] Functions ψ_0 and ψ_n are not included, as they do not satisfy the Dirichlet conditions. Implementation of boundary conditions will be discussed in more detail later.

$$L_{ij} = \int_0^\pi \frac{d\psi_i}{dx}\frac{d\psi_j}{dx}\,dx; \qquad \underline{f}_i = \int_0^\pi \psi_i(x)\,\sin x\,dx \tag{3.69}$$

As already noted, the discrete problem, being just a restriction of the continuous one to the finite-dimensional FE space, inherits the algebraic properties of the continuous formulation. This implies that the global stiffness matrix L is positive definite in this example (and in all cases where the biliniear form of the problem is elliptic).

Equally important is the *sparsity* of the stiffness matrix: most of its entries are zero. Indeed, the Galerkin integrals for L_{ij} in (3.69) are nonzero only if ψ_i and ψ_j are simultaneously nonzero over a certain finite element. This implies that either $i = j$ or nodes i and j are immediate neighbors. In 1D, the global matrix is therefore tridiagonal. In 2D and 3D, the sparsity pattern of the FE matrix depends on the topology of the mesh and on the node numbering (see Sections 3.8 and 3.8).

Algorithmically, it is convenient to compute these integrals on an element-by-element basis, gradually accumulating the contributions to the integrals as the loop over all elements progresses. Clearly, for each element the nonzero contributions will come only from functions ψ_i and ψ_j *that are both nonzero* over this element. For element $\#i$ – that is, for segment $[x_{i-1}, x_i]$ – there are four such nonzero contributions altogether:

$$L^{\text{elem } i}_{i-1,i-1} = \int_{x_{i-1}}^{x_i} \frac{d\psi_{i-1}}{dx}\frac{d\psi_{i-1}}{dx}\,dx = \int_{x_{i-1}}^{x_i} \frac{1}{h}\frac{1}{h}\,dx = \frac{1}{h}$$

$$L^{\text{elem } i}_{i-1,i} = \int_{x_{i-1}}^{x_i} \frac{d\psi_{i-1}}{dx}\frac{d\psi_i}{dx}\,dx = \int_{x_{i-1}}^{x_i} \frac{1}{h}\frac{-1}{h}\,dx = -\frac{1}{h}$$

$$L^{\text{elem } i}_{i,i-1} = L^{\text{elem } i}_{i-1,i} \quad \text{by symmetry}$$

$$L^{\text{elem } i}_{i,i} = \frac{1}{h} \quad (\text{same as } L^{\text{elem } i}_{i-1,i-1})$$

$$\underline{f}_{i-1} = \int_{x_{i-1}}^{x_i} \psi_{i-1}(x)\,\sin x\,dx = \frac{\sin x_i - x_i\cos x_i + x_{i-1}\cos x_i - \sin x_{i-1}}{h}$$

$$\underline{f}_i = \int_{x_{i-1}}^{x_i} \psi_i(x)\,\sin x\,dx = -\frac{\sin x_i - x_i\cos x_{i-1} + x_i\cos x_i - \sin x_{i-1}}{h}$$

These results can be conveniently arranged into a 2×2 matrix

$$L^{\text{elem } i} = \frac{1}{h}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \tag{3.70}$$

called, for historical reasons, the *element stiffness matrix*, and the element contribution to the right hand side is a vector

$$\underline{f}^{\text{elem } i} = \frac{1}{h}\begin{pmatrix} \sin x_i - x_i\cos x_i + x_{i-1}\cos x_i - \sin x_{i-1} \\ -\sin x_i + x_i\cos x_{i-1} - x_{i-1}\cos x_{i-1} + \sin x_{i-1} \end{pmatrix} \tag{3.71}$$

Remark 2. A word of caution: in the engineering literature, it is not uncommon to introduce "element equations" of the form

$$L^{\text{elem } i} u^{\text{elem } i} \;=\; f^{\text{elem } i} \quad (!!??)$$

Such equations are devoid of mathematical meaning. The actual Galerkin equation involves a test function that spans *a group* of adjacent elements (two in 1D), and so there is no valid equation for a single element. Incidentally, triangular meshes have approximately two times more elements than nodes; so, if "element equations" were taken seriously, there would be about twice as many equations as unknowns!

A sample Matlab code at the end of this subsection (p. 100) gives a "no-frills" implementation of the FE algorithm for the 1D model problem. To keep the code as simple as possible, much of the formulation is hard-coded, including the specific interval Ω, expressions for the right hand side and (for verification and error analysis) the exact solution. The only free parameter is the number of elements n. In actual computational practice, such hard-coding should of course be avoided. Commercial FE codes strive to provide maximum flexibility in setting up geometrical and physical parameters of the problem, with convenient user interface.

Some numerical results are shown in the following figures. Fig. 3.9 provides a visual comparison of the FE solutions for 6 and 12 finite elements with the exact solution. Not surprisingly, the solution with 12 elements is more accurate.

Fig. 3.10 displays several precise measures of the error:

- The relative nodal error defined as

$$\epsilon_{\text{nodal}} \;=\; \frac{\|\underline{u} - \mathcal{N}u^*\|}{\|\mathcal{N}u^*\|}$$

  where $\underline{u} \in \mathbb{R}^{n-1}$ is the Euclidean vector of nodal values of the FE solution, $u^*(x)$ is the exact solution, and $\mathcal{N}u^*$ denotes the vector of nodal values of u^* on the grid.
- The L_2 norm of the error

$$\epsilon_{L2} \;=\; \|u_h - u^*\|$$

  This error measures the discrepancy between the numerical and exact solutions as *functions* over $[0, \pi]$ rather than Euclidean vectors of nodal values.
- The L_2 norm of the *derivative*

$$\epsilon_{H1} \;=\; \left\| \frac{d(u_h - u^*)}{dx} \right\|$$

  Due to the zero Dirichlet boundary conditions, this norm differs by no more than a constant factor from the H_1-norm; hence the notation.

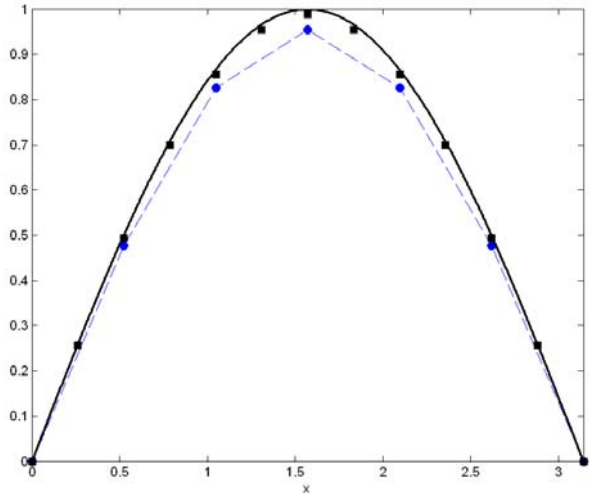


Fig. 3.9. FE solutions with 6 elements (circles) and 12 elements (squares) vs. the exact solution $\sin x$ (solid line).

Due to the simplicity of this example and of the exact solution, these measures can be computed up to the roundoff error. For more realistic problems, particularly in 2D and 3D, the errors can only be estimated.

In Fig. 3.10 the three error measures are plotted vs. the number of elements. The linearity of the plots on the log-log scale implies that the errors are proportional to h^γ, and the slopes of the lines correspond to $\gamma = 2$ for the nodal and L_2 errors and $\gamma = 1$ for the H_1 error. The *derivative* of the solution is computed less accurately than the solution itself. This certainly makes intuitive sense and also agrees with theoretical results quoted in Section 3.10.

Example 5. How will the numerical procedure change if the boundary conditions are different?

First consider inhomogeneous Dirichlet conditions. Let us assume that in the previous example the boundary values are $u(0) = 1$, $u(\pi) = -1$, so that the exact solution is now $u^*(x) = \cos x$. In the hat-function expansion of the (piecewise-linear) FE solution

$$u_h(x) \;=\; \sum_{i=0}^{n} u_{hi}\psi_i(x)$$

the summation now includes boundary nodes in addition to the interior ones. However, the coefficients u_{h0} and u_{hn} at these nodes are the known Dirichlet values, and hence no Galerkin equations with test functions ψ_0 and ψ_n are necessary. In the Galerkin equation corresponding to the test function ψ_1,

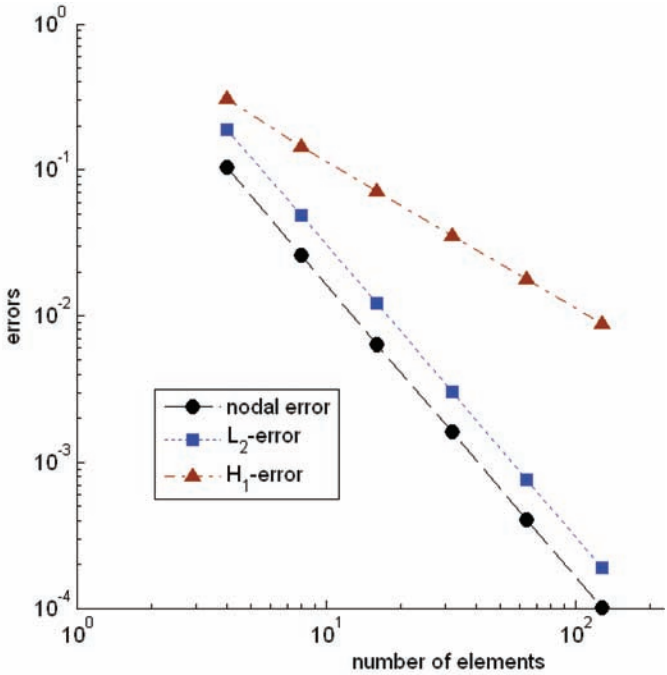


Fig. 3.10. Several measures of error vs. the number of elements for the 1D model problem: relative nodal error (circles), L_2-error (squares), H_1-error (triangles). Note the log–log scale.

$$(\psi_0', \psi_1')\, u_{h0} \;+\; (\psi_1', \psi_1')\, u_{h1} \;=\; (f, \psi_1)$$

the first term is known and gets moved to the right hand side:

$$(\psi_1', \psi_1')\, u_{h1} \;=\; (f, \psi_1) \;-\; (\psi_0', \psi_1')\, u_{h0} \tag{3.72}$$

As usual, parentheses in these expressions are L_2 inner products and imply integration over the computational domain.

The necessary algorithmic adjustments should now be clear. There is no change in the computation of element matrices. However, whenever an entry of the element matrix corresponding to a Dirichlet node is encountered,[18] this entry is *not* added to the global system matrix. Instead, the right hand side is adjusted as prescribed by (3.72). A similar adjustment is made for the other boundary node ($x_n = \pi$) as well. In 2D and 3D problems, there may be many Dirichlet nodes, and all of them are handled in a similar manner. The appropriate changes in the Matlab code are left as an exercise for the interested reader. The FE solution for a small number of elements is compared with the

[18] Clearly, this may happen only for elements adjacent to the boundary.

exact solution ($\cos x$) in Fig. 3.11, and the error measures are shown as a function of the number of elements in Fig. 3.12.

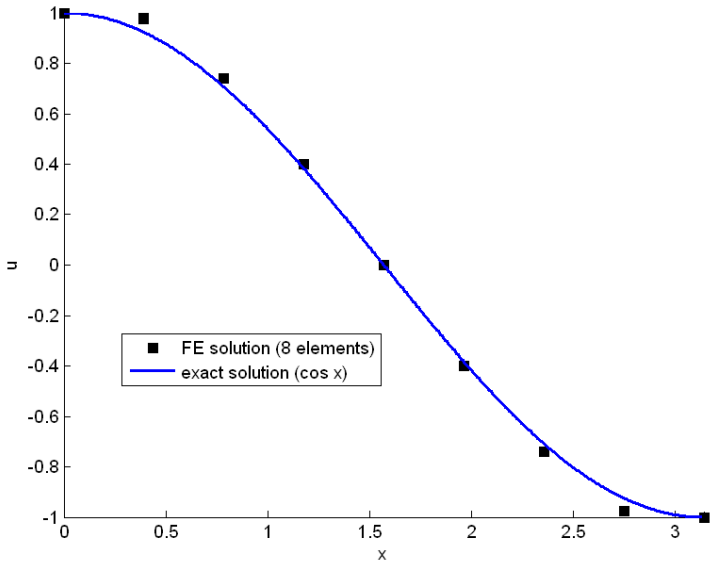


Fig. 3.11. FE solution with 8 elements (markers) vs. the exact solution $\cos x$ (solid line).

Neumann conditions in the Galerkin formulation are natural[19] and therefore do not require any algorithmic treatment: elements adjacent to the Neumann boundary are treated exactly the same as interior elements.

Despite its simplicity, the one-dimensional example above contains the key ingredients of general FE algorithms:

1. **Mesh generation and the choice of FE approximating functions.** In the 1D example, "mesh generation" is trivial, but it becomes complicated in 2D and even more so in 3D. Only piecewise-linear approximating functions have been used here so far; higher-order functions are considered in the subsequent sections.
2. **Local and global node numbering.** For the computation of element matrices (see below), it is convenient to use local numbering (e.g. nodes 1, 2 for a segment in 1D, nodes 1, 2, 3 for a triangular element in 2D, etc.) At

[19] We showed that Neumann conditions are natural – i.e. automatically satisfied – by the solution of the *continuous* weak problem. The FE solution does not, as a rule, satisfy the Neumann conditions exactly but should do so in the limit of $h \to 0$, although this requires a separate proof.

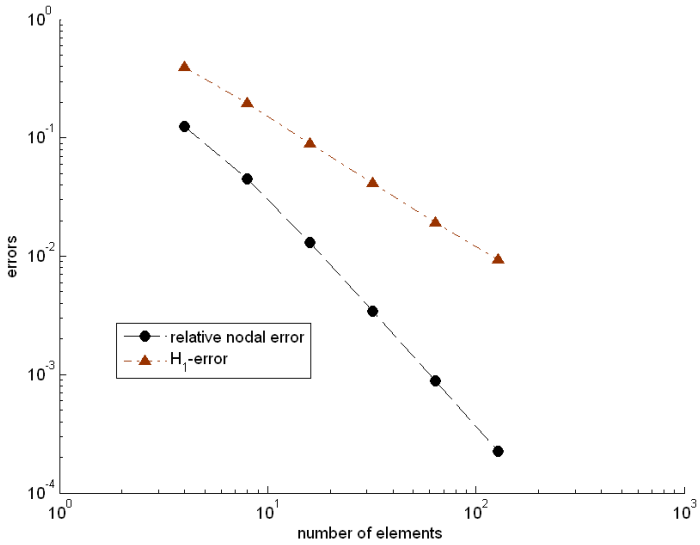


Fig. 3.12. The relative nodal error (circles) and the H_1-error (triangles) for the model Dirichlet problem. Note the log–log scale.

the same time, some *global* numbering of all mesh nodes from 1 to n is also needed. This global numbering is produced by a mesh generator that also puts local node numbers for each element in correspondence with their global numbers. In the 1D example, mesh generation is trivial, and so is the local-to-global association of node numbers: for element (segment) $\#i$, ($i = 1,2, \ldots, n$), local node 1 (the left node) corresponds to global node $i-1$, and local node 2 corresponds to global node i. The 2D and 3D cases are considered in Section 3.8 and Section 3.9.

3. **Computation of element matrices and of element-wise contributions to the right hand side.** In the 1D example, these quantities were computed analytically; in more complicated cases, when analytical expressions are unavailable (this is frequently the case for curved or high order elements in 2D and 3D), Gaussian quadratures are used.
4. **Assembly of the global matrix and of the right hand side.** In a loop over all elements, the element contributions are added to the global matrix and to the right hand side; in the FE langauge, the matrix and the right hand side are "assembled" from element-wise contributions. The entries of each element matrix are added to the respective entries of the global matrix and right hand side. See Section 3.8 for more details in the 2D case.
5. **The treatment of boundary conditions.** The Neumann conditions in 1D, 2D or 3D do not require any special treatment – in other words, the

FE algorithm may simply "ignore" these conditions and the solution will, in the limit, satisfy them automatically. The Robin condition containing a combination of the potential and its normal derivative is also natural but results in an additional boundary integral that will not be considered here. Finally, the Dirichlet conditions have to be taken into account explicitly. The following algorithmic adjustment is made in the loop over all elements. If L_{ij} is an entry of the element matrix and j is a Dirichlet node but i isn't, then L_{ij} is *not* added to the global stiffness matrix. Instead, the quantity $L_{ij}u_j$, where u_j is the known Dirichlet value of the solution at node j, is subtracted from the right hand side entry $\underline{f}_i$, as prescribed by equation (3.72). If *both* i and j are Dirichlet nodes, L_{ij} is set to zero.

6. **Solution of the FE system of equations.** System solvers are reviewed in Section 3.11.
7. **Postprocessing of the results.** This may involve differentiation of the solution (to compute fields from potentials), integration over surfaces (to find field fluxes, etc.), and various contour, line or surface plots. Modern commercial FE packages have elaborate postprocessing capabilities and sophisticated graphical user interface; this subject is largely beyond the scope of this book, but some illustrations can be found in Chapter 7.

At the same time, there are several more advanced features of FE analysis that are not evident from the 1D example and will be considered (at a varying level of detail) in the subsequent sections of this chapter:

- Curved elements – used in 2D and 3D for more accurate approximation of curved boundaries.
- Adaptive mesh refinement (Section 3.13). The mesh is refined locally, in the subregions where the numerical error is estimated to be highest. (In addition, the mesh may be un-refined in subregions with lower errors.) The problem is then solved again on the new grid. The key to the success of this strategy is a sensible error indicator that is computed *a posteriori*, i.e. after the FE solution is found.
- Vector finite elements (Section 3.12). The most straightforward way of dealing with vector fields in FE analysis is to approximate each Cartesian component separately by scalar functions. While this approach is adequate in some cases, it turns out not to be the most solid one in general. One deficiency is fairly obvious from the outset: some field components are discontinuous at material interfaces, which is not a natural condition for scalar finite elements and requires special constraints. This is, however, only one manifestation of a deeper mathematical structure: fundamentally, electromagnetic fields are better understood as differential forms (Section 3.12).

A Sample Matlab Code for the 1D Model Problem

```
function FEM_1D_example1 = FEM_1D_example1 (n)
% Finite element solution of the Poisson equation
```

```
% -u'' = sin x  on [0, pi];   u(0) = u(pi) = 0
% Input:
% n  -- number of elements

domain_length = pi;  % hard-coded for simplicity of this sample code
h = domain_length / n;  % mesh size (uniform mesh assumed)

% Initialization:
system_matrix = sparse(zeros(n-1, n-1));

rhs = sparse(zeros(n-1, 1));

% Loop over all elements (segments)
for elem_number = 1 : n
  node1 = elem_number - 1;
  node2 = elem_number;

  % Coordinates of nodes:
  x1 = h*node1;
  x2 = x1 + h;

  % Element stiffness matrix:
  elem_matrix = 1/h * [1 -1; -1 1];
  elem_rhs = 1/h * [sin(x2) - x2 * cos(x2) + x1 * cos(x2) - sin(x1);
    ...  -(sin(x2) - x2 * cos(x1) + x1 * cos(x1) - sin(x1))];

  % Add element contribution to the global matrix
  if node1 ~= 0 % contribution for nonzero Dirichlet condition only
      system_matrix(node1, node1) = system_matrix(node1, node1) ...
                                        + elem_matrix(1, 1);
      rhs(node1) = rhs(node1) + elem_rhs(1);
  end
  if (node1 ~= 0) & (node2 ~= n) % contribution for nonzero
                                 % Dirichlet condition only
      system_matrix(node1, node2) = system_matrix(node1, node2) ...
                                        + elem_matrix(1, 2);
      system_matrix(node2, node1) = system_matrix(node2, node1) ...
                                        + elem_matrix(2, 1);
  end
  if node2 ~= n % contribution for nonzero Dirichlet condition only
      system_matrix(node2, node2) = system_matrix(node2, node2) ...
                                        + elem_matrix(2, 2);
      rhs(node2) = rhs(node2) + elem_rhs(2);
  end
end   % end element cycle

u_FEM = system_matrix \ rhs;  % refrain from using
                              % matrix inversion inv()!
```

```
FEM_1D_example1.a = 0;
FEM_1D_example1.b = pi;
FEM_1D_example1.n = n;
FEM_1D_example1.u_FEM = u_FEM;

return;
```

3.7.2 Higher-Order Elements

There are two distinct ways to improve the numerical accuracy in FEM. One is to reduce the size h of (some or all) the elements; this approach is known as (local or global) h-refinement.

Remark 3. It is very common to refer to a single parameter h as the "mesh size," even if finite elements in the mesh have different sizes (and possibly even different shapes). With this terminology, it is tacitly assumed that the ratio of maximum/minimum element sizes is bounded and not too large; then the difference between the minimum, maximum or some average size is relatively unimportant. However, several recursive steps of local mesh refinement may result in a large disparity of the element sizes; in such cases, reference to a single mesh size would be misleading.

The other way to improve the accuracy is to increase the polynomial order p of approximation within (some or all) elements; this is (local or global) p-refinement.

Let us start with second-order elements in one dimension. Consider a geometric element – in 1D, a segment of length h. We are about to introduce *quadratic* polynomials over this element; since these polynomials have three free parameters, it makes sense to deal with their values at *three* nodes and to place these nodes at $x = 0, h/2, h$ relative to a local coordinate system.

The canonical approximating functions satisfy the Kronecker-delta conditions at the nodes. The first function is thus equal to one at node #1 and zero at the other two nodes; this function is easily found to be

$$\psi_1 = \frac{2}{h^2}\left(x - \frac{h}{2}\right)(x - h) \tag{3.73}$$

(The factors in the parentheses are due to the roots at $h/2$ and h; the scaling coefficient $2/h^2$ normalizes the function to $\psi_1(0) = 1$.)

Similarly, the remaining two functions are

$$\psi_2 = \frac{4}{h^2}\, x(h - x) \tag{3.74}$$

$$\psi_3 = \frac{2}{h^2}\, x\left(x - \frac{h}{2}\right) \tag{3.75}$$

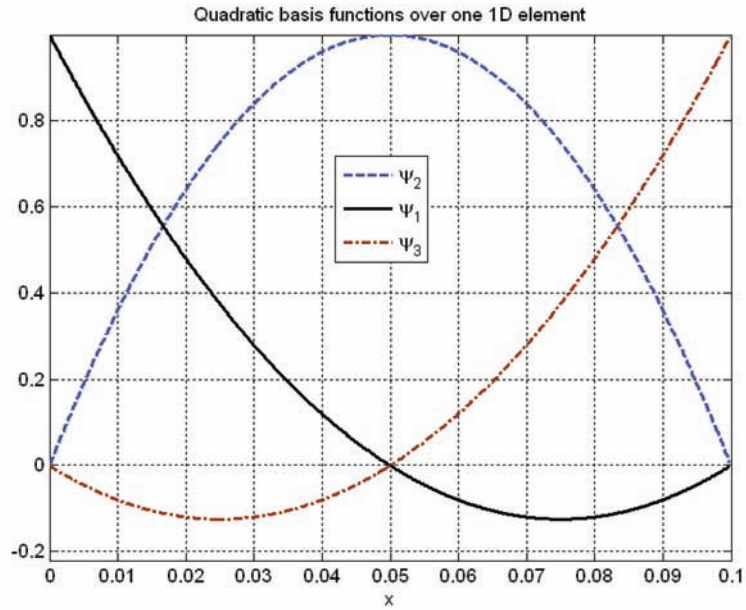


Fig. 3.13. Three quadratic basis functions over one 1D element. $h = 0.1$ as an example.

Fig. 3.13 displays all three quadratic approximating functions over a single 1D element. While the "bubble" ψ_2 is nonzero within one element only, functions $\psi_{1,3}$ actually span two adjacent elements, as shown in Fig. 3.14.

The entries of the element stiffness matrix L and mass matrix M (that is, the Gram matrix of the ψs) are

$$L_{ij} = \int_0^h \psi_i' \psi_j' \, dx$$

where the prime sign denotes the derivative, and

$$M_{ij} = \int_0^h \psi_i \psi_j \, dx$$

These matrices can be computed by straightforward integration:

$$L = \frac{1}{3h} \begin{pmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{pmatrix} \tag{3.76}$$

$$M = \frac{h}{30} \begin{pmatrix} 4 & 2 & -1 \\ 2 & 16 & 2 \\ -1 & 2 & 4 \end{pmatrix} \tag{3.77}$$

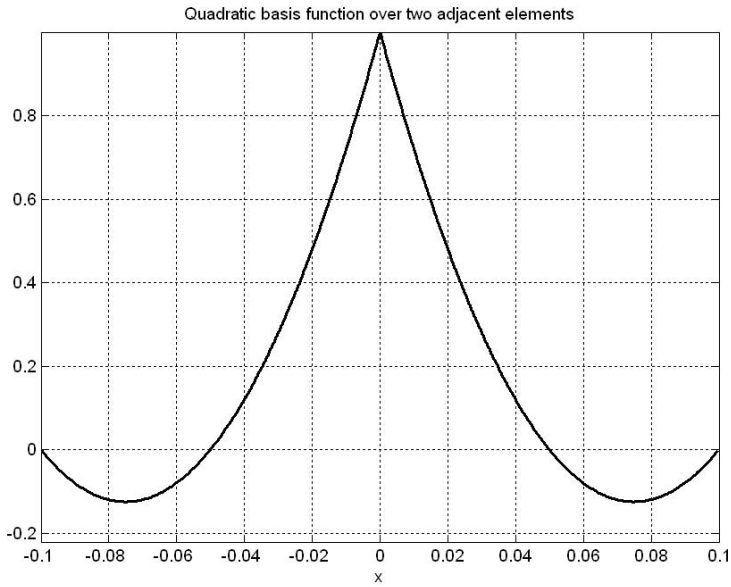


Fig. 3.14. Quadratic basis function over two adjacent 1D elements. $h = 0.1$ as an example.

Naturally, both matrices are symmetric.

The matrix assembly procedure for second-order elements in 1D is conceptually the same as for first-order elements. There are some minor differences:

- For second-order elements, the number of nodes is about double the number of elements.
- Consequently, the correspondence between the local node numbers (1, 2, 3) in an element and their respective global numbers in the grid is a little less simple than for first-order elements.
- The element matrix is 3×3 for second order elements vs. 2×2 for first order ones; the global matrices are five- and three-diagonal, respectively.

Elements of order higher than two can be introduced in a similar manner. The element of order n is, in 1D, a segment of length h with $n+1$ nodes x_0, x_1, $\ldots$, $x_n = x_0 + h$. The approximating functions are polynomials of order n. As with first- and second-order elements, it is most convenient if polynomial #i has the Kronecker-delta property: equal to one at the node x_i and zero at the remaining n nodes. This is the Lagrange interpolating polynomial

$$\Lambda_i(x) = \frac{(x-x_0)(x-x_1)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)} \quad (3.78)$$

Indeed, the roots of this polynomial are $x_0, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$, which immediately leads to the expression in the numerator. The denominator is the normalization factor needed to make $\Lambda_i(x)$ equal to one at $x = x_i$.

The focus of this chapter is on the main ideas of finite element analysis rather than on technical details. With regard to the computation of element matrices, assembly procedures and other implementation issues for high order elements, I defer to more comprehensive FE texts cited at the end of this chapter.

3.8 The Finite Element Method in Two Dimensions

3.8.1 First-Order Elements

In two dimensions, most common element shapes are triangular (by far) and quadrilateral. Fig. 3.15 gives an example of a triangular mesh, with the global node numbers displayed. Element numbering is not shown to avoid congestion in the figure.

This section deals with first-order triangular elements. The approximating functions are linear over each triangle and continuous in the whole domain. Each approximating function spans a cluster of elements (Fig. 3.16) and is zero outside that cluster.

Expressions for element-wise basis functions can be derived in a straightforward way. Let the element nodes be numbered $1, 2, 3$[20] in the counter-clockwise direction[21] and let the coordinates of node i (i = 1,2,3) be x_i, y_i. As in the 1D case, it is natural to look for the basis functions satisfying the Kronecker-delta condition.

More specifically, the basis function $\psi_1 = a_1 x + b_1 y + c_1$, where a_1, b_1 and c_1 are coefficients to be determined, is equal to one at node #1 and zero at the other two nodes:

$$\begin{aligned} a_1 x_1 + b_1 y_1 + c_1 &= 1 \\ a_1 x_2 + b_1 y_2 + c_1 &= 0 \\ a_1 x_3 + b_1 y_3 + c_1 &= 0 \end{aligned} \tag{3.79}$$

or equivalently in matrix-vector form

$$X d_1 = e_1, \quad X = \begin{pmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{pmatrix}; \quad d_1 = \begin{pmatrix} a_1 \\ b_1 \\ c_1 \end{pmatrix}; \quad e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \tag{3.80}$$

Similar relationships hold for the other two basis functions, ψ_2 and ψ_3, the only difference being the right hand side of system (3.80). It immediately

[20] These are *local* numbers that have their corresponding global numbers in the mesh; for example, in the shaded element of Fig. 3.15 (bottom) global nodes 179, 284 and 285 could be numbered as 1, 2, 3, respectively.

[21] The significance of this choice of direction will become clear later.

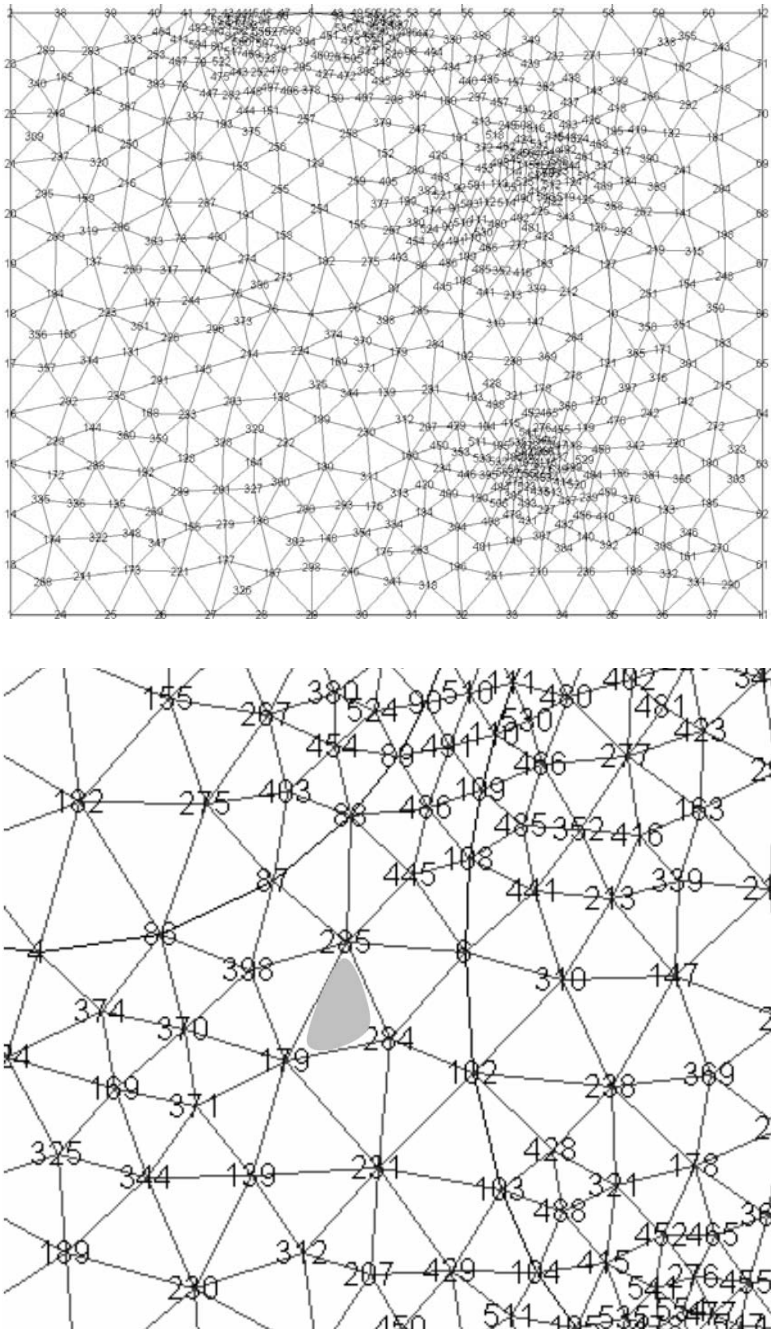


Fig. 3.15. An example of a triangular mesh with node numbering (top) and a fragment of the same mesh (bottom).

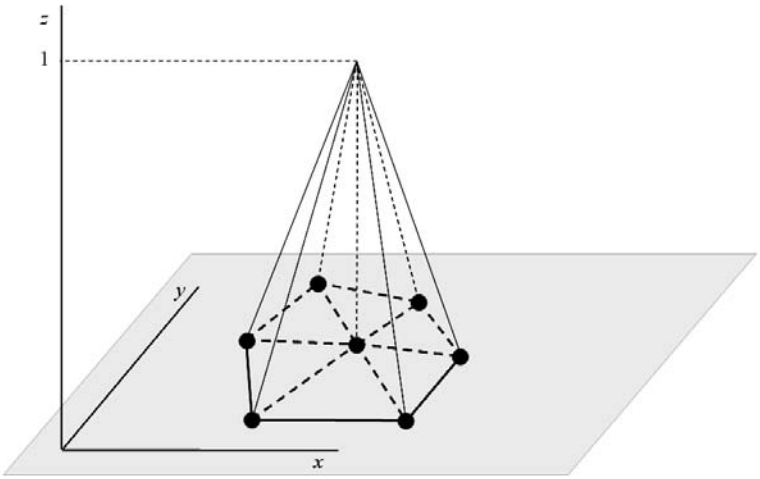


Fig. 3.16. A piecewise-linear basis function in 2D over a cluster of triangular elements. Circles indicate mesh nodes. The basis function is represented by the surface of the pyramid.

follows from (3.80) that the coefficients a, b, c for all three basis functions can be collected together in a compact way:

$$XD = I, \qquad D = \begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{pmatrix} \tag{3.81}$$

where I is the 3×3 identity matrix. Hence the coefficients of the basis functions can be expressed succinctly as

$$D = X^{-1} \tag{3.82}$$

From analytical geometry, the determinant of X is equal to $2S_\Delta$, where S_Δ is the area of the triangle. (That is where the counter-clockwise numbering of nodes becomes important; for clockwise numbering, the determinant would be equal to *minus* $2S_\Delta$.) This leads to simple explicit expressions for the basis functions:

$$\psi_1 = \frac{(y_2 - y_3)x + (x_3 - x_2)y + (x_2y_3 - x_3y_2)}{2S_\Delta} \tag{3.83}$$

with the other two functions obtained by cyclic permutation of the indexes.

Since the basis functions are linear, their gradients are just constants:

$$\nabla\psi_1 = \frac{y_2 - y_3}{2S_\Delta}\,\hat{x} + \frac{x_3 - x_2}{2S_\Delta}\,\hat{y} \tag{3.84}$$

with the formulas for $\psi_{2,3}$ again obtained by cyclic permutation. These expressions are central in the FE-Galerkin formulation.

It would be straightforward to verify from (3.83), (3.84) that

$$\psi_1 + \psi_2 + \psi_3 = 1 \tag{3.85}$$

$$\nabla\psi_1 + \nabla\psi_2 + \nabla\psi_3 = 0 \tag{3.86}$$

However, these results can be obtained without any algebraic manipulation. Indeed, due to the Kronecker delta property of the basis, any function $u(x, y)$ linear over the triangle can be expressed via its nodal values $u_{1,2,3}$ as

$$u(x, y) = u_1\psi_1 + u_2\psi_2 + u_3\psi_3$$

Equation (3.85) follows from this simply for $u(x, y) \equiv 1$.

Functions $\psi_{1,2,3}$ are also known as *barycentric coordinates* and have an interesting geometric interpretation (Fig. 3.17). For any point x, y in the plane, $\psi_1(x, y)$ is the ratio of the shaded area to the area of the whole triangle: $\psi_1(x, y) = S_1(x, y)/S_\Delta$. Similar expressions are of course valid for the other two basis functions.

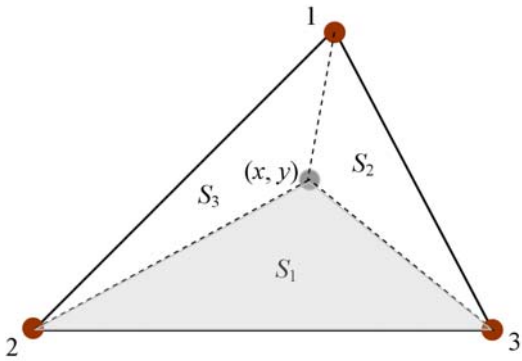


Fig. 3.17. Geometric interpretation of the linear basis functions: $\psi_1(x, y) = S_1(x, y)/S_\Delta$, where S_1 is the shaded area and S_Δ is the area of the whole triangle. (Similar for $\psi_{2,3}$.)

Indeed, the fact that S_1/S_Δ is equal to one at node #1 and zero at the other two nodes is geometrically obvious. Moreover, it is a linear function of coordinates because S_1 is proportional to height l of the shaded triangle (the "elevation" of point x, y over the "base" segment 2–3), and l can be obtained by a linear transformation of coordinates (x, y).

The three barycentric coordinates are commonly denoted with $\lambda_{1,2,3}$, so the linear FE basis functions are just $\psi_i \equiv \lambda_i$ (i = 1,2,3). Higher-order FE bases can also be conveniently expressed in terms of λ (Section 3.8.2).

The element stiffness matrix for first order elements is easy to compute because the gradients (3.84) of the basis functions are constant:

$$(\nabla\lambda_i, \nabla\lambda_j) \;\equiv\; \int_\Delta \nabla\lambda_i \cdot \nabla\lambda_j \, dS \;=\; \nabla\lambda_i \cdot \nabla\lambda_j \, S_\Delta, \quad i, j = 1, 2, 3 \tag{3.87}$$

where the integration is over a triangular element and S_Δ is the area of this element. Expressions for the gradients are available (3.84) and can be easily substituted into (3.87) if an explicit formula for the stiffness matrix in terms of the nodal coordinates is desired.

Computation of the element mass matrix (the Gram matrix of the basis functions) is less simple but the result is quite elegant. The integral of, say, the product $\lambda_1\lambda_2$ over the triangular element can be found using an affine transformation of this element to the "master" triangle with nodes 1, 2, 3 at (1, 0), (0, 1) and (0, 0), respectively. Since the area of the master triangle is 1/2, the Jacobian of this transformation is equal to $2S_\Delta$ and we have[22]

$$(\lambda_1,\, \lambda_2) \;\equiv\; \int_\Delta \lambda_1\lambda_2 \, dS \;=\; 2S_\Delta \int_0^1 x dx \int_0^{1-x} y dy \;=\; \frac{S_\Delta}{12}$$

Similarly,

$$(\lambda_1,\, \lambda_1) \;=\; 2\,\frac{S_\Delta}{12}$$

and the complete element mass matrix is

$$M \;=\; \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} \frac{S_\Delta}{12} \tag{3.88}$$

The expressions for the inner products of the barycentric coordinates are a particular case of a more general formula that appears in many texts on FE analysis and is quoted here without proof:

$$\int_\Delta \lambda_1^i \lambda_2^j \lambda_3^k \, dS \;=\; \frac{i!\, j!\, k!}{(i+j+k+2)!}\, 2S_\Delta \tag{3.89}$$

for any nonnegative integers i, j, k. M_{11} of (3.88) corresponds to $i = 2$, $j = k = 0$; M_{12} corresponds to $i = j = 1$, $k = 0$; etc.

Remark 4. The notion of "master element" (or "reference element") is useful and long-established in finite element analysis. Properties of FE matrices and FE approximations are usually examined via affine transformations of elements to the "master" ones. In that sense, analysis of finite element interpolation errors in Section 3.14.2 below (p. 160) is less typical.

[22] The Jacobian is positive for the counter-clockwise node numbering convention.

Example 6. Let us find the basis functions and the FE matrices for a right triangle with node #1 at the origin, node #2 on the x-axis at $(h_x, 0)$, and node #3 on the y-axis at $(0, h_y)$ (mesh sizes h_x, h_y are positive numbers). The coordinate matrix is

$$X = \begin{pmatrix} 0 & 0 & 1 \\ h_x & 0 & 1 \\ 0 & h_y & 1 \end{pmatrix}$$

which yields the coefficient matrix

$$D = X^{-1} = \begin{pmatrix} -h_x^{-1} & h_x^{-1} & 0 \\ -h_y^{-1} & 0 & h_y^{-1} \\ 1 & 0 & 0 \end{pmatrix}$$

Each column of this matrix is a set of three coefficients for the respective basis function; thus the three columns translate into

$$\begin{aligned} \psi_1 &= 1 - h_x^{-1}x - h_y^{-1}y \\ \psi_2 &= h_x^{-1}x \\ \psi_3 &= h_y^{-1}y \end{aligned}$$

The sum of these functions is identically equal to one as it should be according to (3.85). Functions ψ_2 and ψ_3 in this case are particularly easy to visualize: ψ_2 is a linear function of x equal to one at node #2 and zero at the other two nodes; ψ_3 is similar. The gradients are

$$\begin{aligned} \nabla\psi_1 &= -h_x^{-1}\hat{x} - h_y^{-1}\hat{y} \\ \nabla\psi_2 &= h_x^{-1}\hat{x} \\ \nabla\psi_3 &= h_y^{-1}\hat{y} \end{aligned}$$

Computing the entries of the element stiffness matrix is easy because the gradients of λs are (vector) constants. For example,

$$(\nabla\lambda_1, \nabla\lambda_1) = \int_\Delta \nabla\lambda_1 \cdot \nabla\lambda_1 \, dS = (h_x^{-2} + h_y^{-2}) \, S_\Delta$$

Since $S_\Delta = h_x h_y / 2$, the complete stiffness matrix is

$$L = \begin{pmatrix} h_x^{-2} + h_y^{-2} & -h_x^{-2} & -h_y^{-2} \\ -h_x^{-2} & h_x^{-2} & 0 \\ -h_y^{-2} & 0 & h_y^{-2} \end{pmatrix} \frac{h_x h_y}{2} \tag{3.90}$$

This expression becomes particularly simple if $h_x = h_y = h$:

$$L = \frac{1}{2} \begin{pmatrix} 2 & -1 & -1 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} \tag{3.91}$$

The mass matrix is, according to the general expression (3.88),

$$M = \frac{S_\Delta}{12}\begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} = \frac{h_x h_y}{24}\begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} \tag{3.92}$$

An example of Matlab implementation of FEM for a triangular mesh is given at the end of this section; see p. 114 for the description and listing of the code. As an illustrative example, consider a dielectric particle with some nontrivial shape – say, T-shaped – in a uniform external field. The geometric setup is clear from Figs. 3.18 and 3.19.

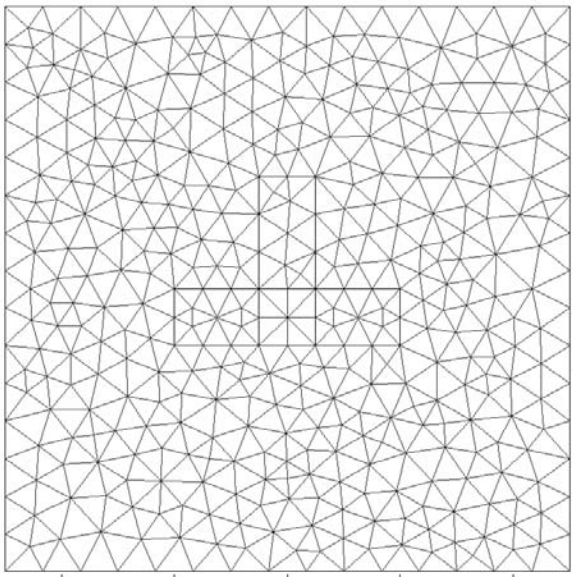


Fig. 3.18. A finite element mesh for the electrostatic problem: a T-shaped particle in an external field. The mesh has 422 nodes and 782 triangular elements.

The potential of the applied external field is assumed to be $u = x$ and is imposed as the Dirichlet condition on the boundary of the computational domain. Since the particle disturbs the field, this condition is not exact but becomes more accurate if the domain boundary is moved farther away from the particle; this, however, increases the number of nodes and consequently the computational cost of the simulation. Domain truncation is an intrinsic difficulty of electromagnetic FE analysis (unlike, say, analysis of stresses and strains confined to a finite mechanical part). Various ways of reducing the domain truncation error are known: radiation boundary conditions and Perfectly Matched Layers (PML) for wave problems (e.g. Z.S. Sacks [SKLL93], Jo-Yu Wu *et al.* [WKLL97]), hybrid finite element/boundary element methods,

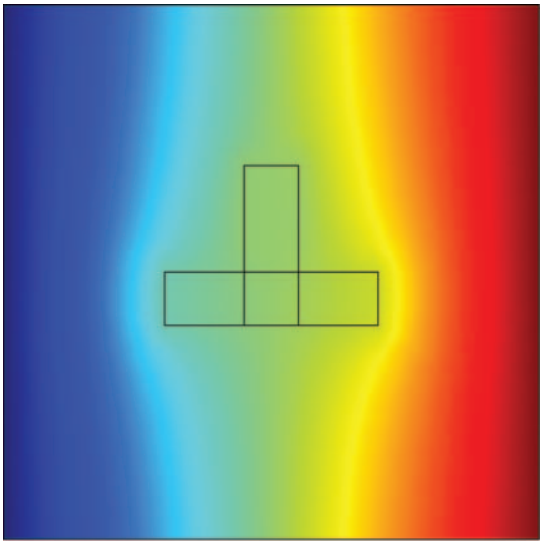


Fig. 3.19. The potential distribution for the electrostatic example: a T-shaped particle in an external field.

infinite elements, "ballooning," spatial mappings (A. Plaks *et al.* [PTPT00]) and various other techniques (see Q. Chen & A. Konrad [CK97] for a review). Since domain truncation is only tangentially related to the material of this section, it is not considered here further but will reappear in Chapter 7.

For inhomogeneous Dirichlet conditions, the weak formulation of the problem has to be modified, with the corresponding minor adjustments to the FE algorithm. The underlying mathematical reason for this modification is that functions satisfying a given inhomogeneous Dirichlet condition form an affine space rather than a linear space (e.g. the sum of two such functions has a different value at the boundary). The remedy is to split the original unknown function u up as

$$u \;=\; u_0 \;+\; u_{\neq 0} \tag{3.93}$$

where $u_{\neq 0}$ is some sufficiently smooth function satisfying the given inhomogeneous boundary condition, while the remaining part u_0 satisfies the homogeneous one. The weak formulation is

$$\mathcal{L}(u_0, v_0) \;=\; (f, v_0) \;-\; \mathcal{L}(u_{\neq 0}, v_0), \quad u_0 \in H_0^1(\Omega), \;\; \forall v_0 \in H_0^1(\Omega) \tag{3.94}$$

In practice, the implementation of this procedure is more straightforward than it may appear from this expression. The inhomogeneous part $u_{\neq 0}$ is spanned by the FE basis functions corresponding to the Dirichlet nodes; the homogeneous part of the solution is spanned by the basis functions for all other nodes. If j is a Dirichlet boundary node, the solution value u_j at this

node is given, and hence the term $L_{ij}u_j$ in the global system of FE equations is known as well. It is therefore moved (with the opposite sign of course) to the right hand side.

In the T-shaped particle example, the mesh has 422 nodes and 782 triangular elements, and the stiffness matrix has 2446 nonzero entries. The sparsity structure of this matrix (also called the *adjacency structure*) – the set of index pairs (i, j) for which $L_{ij} \neq 0$ – is exhibited in Fig. 3.20. The distribution of nonzero entries in the matrix is quasi-random, which has implications for the solution procedures if *direct* solvers are employed. Such solvers are almost invariably based on some form of Gaussian elimination; for symmetric positive definite matrices, it is Cholesky decomposition $U^T U$, where U is an upper triangular matrix.[23] While Gaussian elimination is a very reliable[24] and relatively simple procedure, for sparse matrices it unfortunately produces "fill-in": zero entries become nonzero in the process of elimination (or Cholesky decomposition), which substantially degrades the computational efficiency and memory usage.

In the present example, Cholesky decomposition applied to the original stiffness matrix with 2446 nonzero entries[25] produces the Cholesky factor with 24,969 nonzeros and hence requires about 20 times more memory (if symmetry is taken advantage of); compare Figs. 3.20 and 3.21. For more realistic practical cases, where matrix sizes are much greater, the effect of fill-in is even more dramatic.

It is worth noting – in passing, since this is not the main theme of this section – that several techniques are available for reducing the amount of fill-in in Cholesky factorization. The main ideas behind these techniques are clever permutations of rows and columns (equivalent to renumbering of nodes in the FE mesh), block algorithms (including divide-and-conquer type recursion), and combinations thereof. A. George & J.W.H. Liu give a detailed and lucid exposition of this subject [GL81]. In the current example, the so-called reverse Cuthill–McKee ordering reduces the number of nonzero entries in the Cholesky factor to 7230, which is more than three times better than for the original numbering of nodes (Figs. 3.22 and 3.23).

The "minimum degree" ordering [GL81] is better by another factor of ~ 2: the number of nonzeros in the Cholesky triangular matrix is equal to 3717 (Figs. 3.24 and 3.25). These permutation algorithms will be revisited in the solver section (p. 129).

---

[23] Cholesky decomposition is usually written in the equivalent form of LL^T, where L is a *lower* triangular matrix, but symbol L in this chapter is already used for the FE stiffness matrix.

[24] It is known to be stable for symmetric positive definite matrices but may require pivoting in general.

[25] Of which only a little more than one half need to be stored due to matrix symmetry.

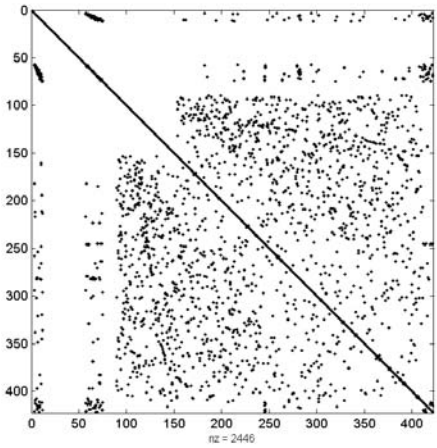


Fig. 3.20. The sparsity (adjacency) structure of the global FE matrix in the T-shaped particle example.

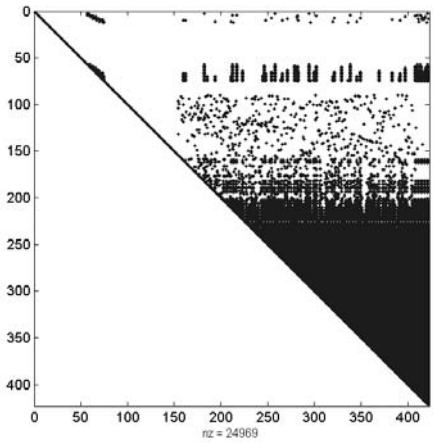


Fig. 3.21. The sparsity structure of the Cholesky factor of the global FE matrix in the T-shaped particle example.

Appendix: Sample Matlab Code for FEM with First-Order Triangular Elements

The Matlab code below is intended to be the simplest possible illustration of the finite element procedure. As such, it uses first order elements and is optimized for algorithmic simplicity rather than performance. For example, there is some duplication of variables for the sake of clarity, and symmetry of

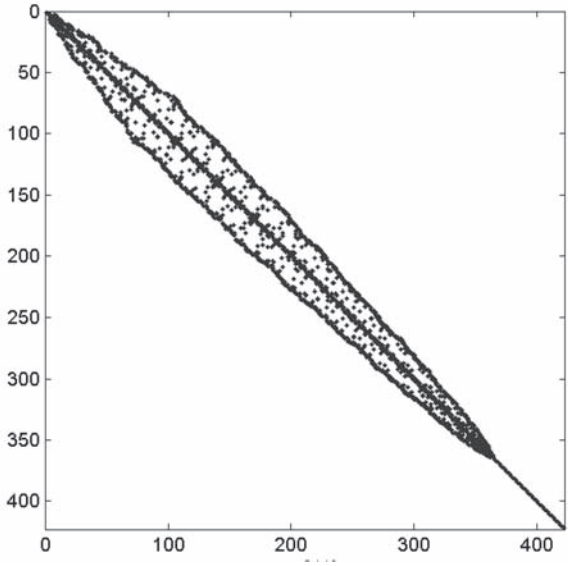


Fig. 3.22. The sparsity structure of the global FE matrix after the reverse Cuthill–McKee reordering of nodes.

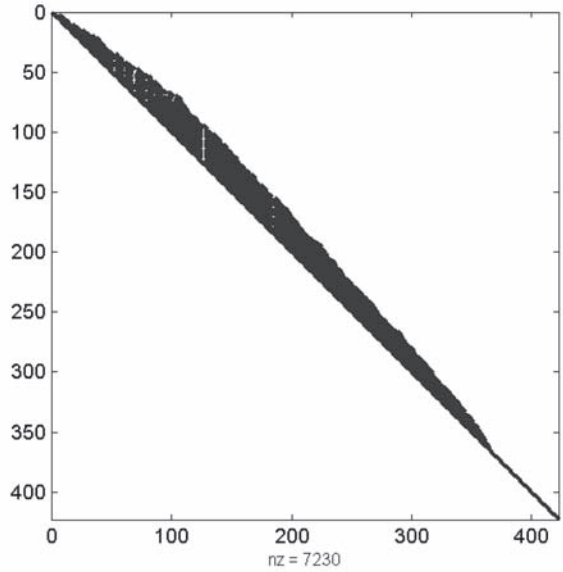


Fig. 3.23. The sparsity structure of the upper-triangular Cholesky factor of the global FE matrix after the reverse Cuthill–McKee reordering of nodes.

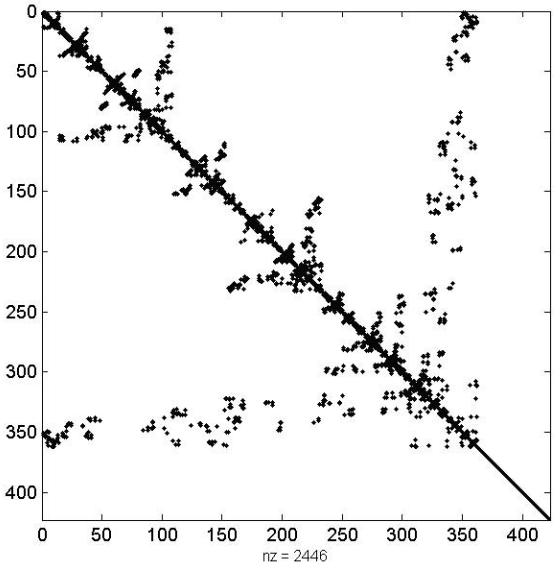


Fig. 3.24. The sparsity structure of the global FE matrix after the minimum degree reordering of nodes.

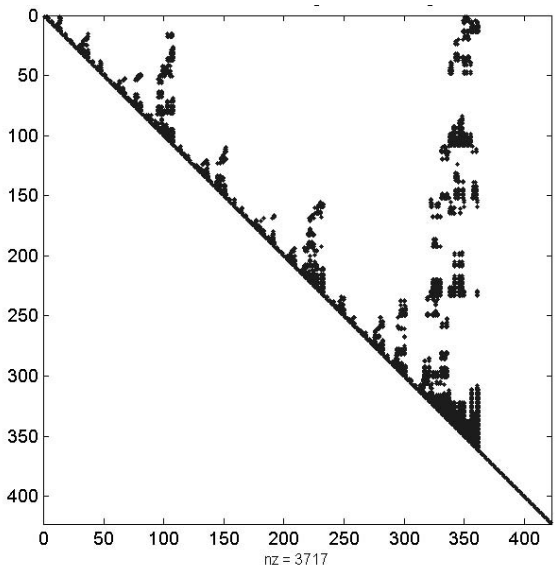


Fig. 3.25. The sparsity structure of the upper-triangular Cholesky factor of the global FE matrix after the minimum degree reordering of nodes.

the FE stiffness matrix is not taken advantage of. Improvements become fairly straightforward to make once the essence of the algorithm is understood.

The starting point for the code is a triangular mesh generated by FEMLAB^{TM}, a commercial finite element package[26] integrated with Matlab. The input data structure `fem` generated by FEMLAB in general contains the geometric, physical and FE mesh data relevant to the simulation. For the purposes of this section, only mesh data (the field `fem.mesh`) is needed. Second-order elements are the default in FEMLAB, and it is assumed that this default has been changed to produce first-order elements for the sample Matlab code.

The `fem.mesh` structure (or simply `mesh` for brevity) contains several fields:

- `mesh.p` is a $2 \times n$ matrix, where n is the number of nodes in the mesh. The i-th column of this matrix contains the (x, y) coordinates of node $\#i$.
- `mesh.e` is a $7 \times n_{\text{be}}$ matrix, where n_{be} is the number of element edges on all boundaries: the exterior boundary of the domain and material interfaces. The first and second rows contain the node numbers of the starting and end points of the respective edge. The sixth and seventh row contain the region (subdomain) numbers on the two sides of the edge. Each region is a geometric entity that usually corresponds to a particular medium, e.g. a dielectric particle or air. Each region is assigned a unique number. By convention, the region outside the computational domain is labeled as zero, which is used in the Matlab code below to identify the exterior boundary edges and nodes in `mesh.e`. The remaining rows of this matrix will not be relevant to us here.
- `mesh.t` is a $4 \times n_{\text{elems}}$ matrix, where n_{elems} is the number of elements in the mesh. The first three rows contain node numbers of each element in counter-clockwise order. The fourth row is the region number identifying the medium where the element resides.

The second input parameter of the Matlab code, in addition to the `fem` structure, is an array of dielectric permittivities by region number. In the T-shaped particle example, region #1 is air, and the particle includes regions #2–#4, all with the same dielectric permittivity. The following sequence of commands could be used to call the FE solver:

```
% Set parameters:
epsilon_air = 1; epsilon_particle = 10;

epsilon_array = [epsilon_air epsilon_particle*ones(1, 5)];

% Solve the FE problem
FEM_solve = FEM_triangles (fem, epsilon_array)
```

The operation of the Matlab function `FEM_triangles` below should be clear from the comments in the code and from Section 3.8.1.

[26] www.comsol.com

```
function FEM_triangles = FEM_triangles (fem, epsilon_array)
% Input parameters:
% fem  --  structure generated by FEMLAB.
% (See comments in the code and text.)
% epsilon_array -- material parameters by region number.

mesh = fem.mesh;   % duplication for simplicity
n_nodes = length(mesh.p); % array p has dimension 2 x n_nodes;
                          % contains x- and y-coordinates of the nodes.
n_elems = length(mesh.t); % array t has dimension 4 x n_elements.
                          % First three rows contain node numbers
                          % for each element.
                          % The fourth row contains region number
                          % for each element.

% Initialization
rhs = zeros(n_nodes, 1);

global_stiffness_matrix = sparse(n_nodes, n_nodes);
dirichlet = zeros(1, n_nodes);  % flags Dirichlet conditions
                                % for the nodes (=1 for Dirichlet
                                % nodes, 0 otherwise)

% Use FEMLAB data on boundary edges to determine Dirichlet nodes:

boundary_edge_data = mesh.e;  % mesh.e contains FEMLAB data
                              % on element edges at the domain boundary
number_of_boundary_edges = size(boundary_edge_data, 2); for
boundary_edge = 1 : number_of_boundary_edges
    % Rows 6 and 7 in the array are region numbers
    % on the two sides of the edge
    region1 = boundary_edge_data(6, boundary_edge);
    region2 = boundary_edge_data(7, boundary_edge);
    % If one of these region numbers is zero, the edge is at the
    % boundary, and the respective nodes are Dirichlet nodes:
    if (region1 == 0) | (region2 == 0)  % boundary edge
        node1 = boundary_edge_data(1, boundary_edge);
        node2 = boundary_edge_data(2, boundary_edge);
        dirichlet(node1) = 1;
        dirichlet(node2) = 1;
    end
end

% Set arrays of nodal coordinates:
for elem = 1 : n_elems  % loop over all elements
    elem_nodes = mesh.t(1:3, elem);  % node numbers for the element
    for node_loc = 1 : 3
        node = elem_nodes(node_loc);
        x_nodes(node) = mesh.p(1, node);
```

```
        y_nodes(node) = mesh.p(2, node);
    end
end

% Matrix assembly -- loop over all elements:
for elem = 1 : n_elems
    elem_nodes = mesh.t(1:3, elem);
    region_number = mesh.t(4, elem);
    for node_loc = 1 : 3
        node = elem_nodes(node_loc);
        x_nodes_loc(node_loc) = x_nodes(node);
        y_nodes_loc(node_loc) = y_nodes(node);
    end
    % Get element matrices:
    [stiff_mat, mass_mat] = elem_matrices_2D(x_nodes_loc, y_nodes_loc);
    for node_loc1 = 1 : 3
        node1 = elem_nodes(node_loc1);
        if dirichlet(node1) ~= 0
            continue;
        end
        for node_loc2 = 1 : 3
            % symmetry not taken advantage of, to simplify code
            node2 = elem_nodes(node_loc2);
            if dirichlet(node2) == 0  % non-Dirichlet node
                global_stiffness_matrix(node1, node2) = ...
                    global_stiffness_matrix(node1, node2) ...
                    + epsilon_array(region_number) ...
                    * stiff_mat(node_loc1, node_loc2);
            else   % Dirichlet node; update rhs
                rhs(node1) = rhs(node1) - ...
                stiff_mat(node_loc1, node_loc2) * ...
                dirichlet_value(x_nodes(node2), y_nodes(node2));
            end
        end
    end
end

% Equations for Dirichlet nodes are trivial:
for node = 1 : n_nodes
    if dirichlet(node) ~= 0  % a Dirichlet node
        global_stiffness_matrix(node, node) = 1;
        rhs(node) = dirichlet_value(x_nodes(node), y_nodes(node));
    end
end

solution = global_stiffness_matrix \ rhs;

% Output fields:
FEM_triangles.fem = fem;  % record the fem structure
```

```
FEM_triangles.epsilon_array = epsilon_array;  %  material parameters
                                              %  by region number
FEM_triangles.n_nodes = n_nodes;  % number of nodes in the mesh
FEM_triangles.x_nodes = x_nodes;  % array of x-coordinates of the nodes
FEM_triangles.y_nodes = y_nodes;  % array of y-coordinates of the nodes
FEM_triangles.dirichlet = dirichlet;  % flags for the Dirichlet nodes
FEM_triangles.global_stiffness_matrix = global_stiffness_matrix;
                                 % save matrix for testing
FEM_triangles.rhs = rhs;  % right hand side for testing
FEM_triangles.solution = solution;  % nodal values of the potential

return;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

 function [stiff_mat, mass_mat] = elem_matrices_2D(x_nodes, y_nodes)
 % Compute element matrices for a triangle.
 % Input parameters:
 % x_nodes -- x-coordinates of the three nodes,
 %            in counter-clockwise order
 % y_nodes -- the corresponding y-coordinates

 coord_mat = [x_nodes' y_nodes' ones(3, 1)];
       % matrix of nodal coordinates, with an extra column of ones
 coeffs = inv(coord_mat);  % coefficients of the linear basis functions
 grads = coeffs(1:2, :);  % gradients of the linear basis functions

 area = 1/2 * abs(det(coord_mat));  % area of the element
 stiff_mat = area * grads' * grads;  % the FE stiffness matrix
 mass_mat = area / 12 * (eye(3) + ones(3, 3));
                                    % the FE mass matrix

 return;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function dirichlet_value = dirichlet_value (x, y)
% Set the Dirichlet boundary condition

dirichlet_value = x;  % as a simple example

return;
```

3.8.2 Higher-Order Triangular Elements

The discussion in Section 3.8.1 suggests that in a triangular element the barycentric variables λ (p. 108) form a natural set of coordinates (albeit not

independent, as their sum is equal to unity). For first order elements, the barycentric coordinates themselves double as the basis functions. They can also be used to generate FE bases for higher order triangular elements.

A second order element has three corner nodes #1–#3 and three midpoint nodes (Fig. 3.26). All six nodes can be labeled with triplets of indexes (k_1, k_2, k_3); each index k_i increases from 0 to 1 to 2 along the edges toward node i ($i = 1, 2, 3$).

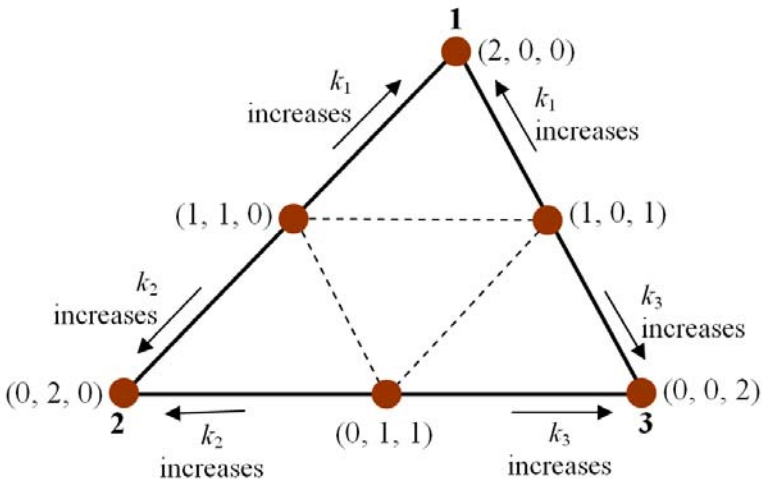


Fig. 3.26. Second order triangular element. The six nodes can be labeled with triplets of indexes (k_1, k_2, k_3), $k_i = 0, 1, 2$. Each node has the corresponding basis function $\Lambda_{k_1}^{k_1}(\lambda_1)\Lambda_{k_2}^{k_2}(\lambda_2)\Lambda_{k_3}^{k_3}(\lambda_3)$.

To each node, there corresponds an FE basis function that is a second order polynomial in λ with the Kronecker-delta property. The explicit expression for this polynomial is $\Lambda_{k_1}^{k_1}(\lambda_1)\Lambda_{k_2}^{k_2}(\lambda_2)\Lambda_{k_3}^{k_3}(\lambda_3)$. For example, the basis function corresponding to node (0, 1, 1) – the midpoint node at the bottom – is $\Lambda_1(\lambda_2)\Lambda_1(\lambda_3)$. Indeed, it is the Lagrange polynomial Λ_1 that is equal to one at the midpoint and to zero at the corner nodes of a given edge, and it is the barycentric coordinates $\lambda_{2,3}$ that vary (linearly) along the bottom edge.

This construction can be generalized to elements of order p. Each side of the triangle is subdivided into p segments; the nodes of the resulting triangular grid are again labeled with triplets of indexes, and the corresponding basis functions are defined in the same way as above. Details can be found in the FE monographs cited at the end of the chapter.

3.9 The Finite Element Method in Three Dimensions

Tetrahedral elements, by analogy with triangular ones in 2D, afford the greatest flexibility in representing geometric shapes and are therefore the most common type in many applications. Hexahedral elements are also frequently used. This section describes the main features of tetrahedral elements; further information about elements of other types can be found in specialized FE books (Section 3.16).

Due to a direct analogy between tetrahedral and triangular elements (Section 3.8), results for tetrahedra are presented below without further ado. Let the coordinates of the four nodes be x_i, y_i, z_i (i = 1,2,3,4). A typical linear basis function – say, ψ_1 – is

$$\psi_1 = a_1 x + b_1 y + c_1 z + d_1$$

with some coefficients a_1, b_1, c_1, d_1. The Kronecker-delta property is desired:

$$\begin{aligned} a_1x_1 + b_1y_1 + c_1z_1 + d_1 &= 1 \\ a_1x_2 + b_1y_2 + c_1z_2 + d_2 &= 0 \\ a_1x_3 + b_1y_3 + c_1z_3 + d_3 &= 0 \\ a_1x_4 + b_1y_4 + c_1z_4 + d_4 &= 0 \end{aligned} \tag{3.95}$$

Equivalently in matrix-vector form

$$Xf_1 = e_1, \quad X = \begin{pmatrix} x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ x_3 & y_3 & z_3 & 1 \\ x_4 & y_4 & z_4 & 1 \end{pmatrix}; \quad f_1 = \begin{pmatrix} a_1 \\ b_1 \\ c_1 \\ d_1 \end{pmatrix}; \quad e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{3.96}$$

with similar relationships for the other three basis functions. In compact notation,

$$XF = I, \quad F = \begin{pmatrix} a_1 & a_2 & a_3 & a_4 \\ b_1 & b_2 & b_3 & b_4 \\ c_1 & c_2 & c_3 & c_4 \\ d_1 & d_2 & d_3 & d_4 \end{pmatrix} \tag{3.97}$$

where I is the 4×4 identity matrix. The coefficients of the basis functions thus are

$$F = X^{-1} \tag{3.98}$$

The determinant of X is equal to $6V$, where V is the volume of the tetrahedron (assuming that the nodes are numbered in a way that produces a *positive* determinant). The basis functions can be found from (3.98), say, by Cramer's rule. Since the basis functions are linear, their gradients are constants.

The sum of the basis functions is unity, for the same reason as for triangular elements:

$$\psi_1 + \psi_2 + \psi_3 + \psi_4 = 1 \tag{3.99}$$

The sum of the gradients is zero:

$$\nabla\psi_1 + \nabla\psi_2 + \nabla\psi_3 + \nabla\psi_4 \;=\; 0 \tag{3.100}$$

Functions $\psi_{1,2,3,4}$ are identical with the barycentric coordinates $\lambda_{1,2,3,4}$ of the tetrahedron. They have a geometric interpretation as ratios of tetrahedral volumes – an obvious analog of the similar property for triangles (Fig. 3.17 on p. 108).

The element stiffness matrix for first order elements is (noting that the gradients are constant)

$$(\nabla\lambda_i, \nabla\lambda_j) \;\equiv\; \int_\Delta \nabla\lambda_i \cdot \nabla\lambda_j \, dV \;=\; \nabla\lambda_i \cdot \nabla\lambda_j V, \quad i,j = 1,2,3,4 \tag{3.101}$$

where the integration is over the tetrahedron and V is its volume. The element mass matrix (the Gram matrix of the basis functions) turns out to be

$$M \;=\; \begin{pmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 \end{pmatrix} \frac{V}{20} \tag{3.102}$$

which follows from the formula

$$\int_\Delta \lambda_1^i \lambda_2^j \lambda_3^k \lambda_4^l \, dV \;=\; \frac{i!\, j!\, k!\, l!}{(i+j+k+l+3)!}\, 6V \tag{3.103}$$

for any nonnegative integers i, j, k, l.

Higher-order tetrahedral elements are constructed in direct analogy with the triangular ones (Section 3.8.2). The second-order tetrahedron has ten nodes (four main vertices and six edge midpoints); the cubic tetrahedral element has 20 nodes (two additional nodes per edge subdividing it into three equal segments, and four nodes at the barycenters of the faces). Detailed descriptions of tetrahedral elements, as well as first- and high-order elements of other shapes (hexahedra, triangular prisms, and others) are easy to find in FE monographs (Section 3.16).

3.10 Approximation Accuracy in FEM

Theoretical considerations summarized in Section 3.5 show that the accuracy of the finite element solution is directly linked, and primarily depends on, the approximation accuracy. In particular, for symmetric elliptic forms $\mathcal{L}$, the Galerkin solution is actually the best approximation of the exact solution in the sense of the $\mathcal{L}$-norm (usually interpreted as an energy norm). In the case of a continuous elliptic, but not necessarily symmetric, form, the solution error depends also on the ellipticity and continuity constants, according to Céa's

theorem; however, the approximation error is still key. The same is true in the general case of continuous but not necessarily symmetric or elliptic forms; then the so-called Ladyzhenskaya–Babuška–Brezzi (LBB) condition relates the solution error to the approximation error via the inf-sup constant (Section 3.10, p. 126).

In all cases, the central role of FE approximation is clear. The main theoretical results on approximation accuracy in FEM are summarized below. But first, let us consider a simple intuitive 1D picture. The exact solution (solid line in Fig. 3.27) is approximated on a FE grid of size h; several finite elements (e) are shown in the figure. The most natural and easy to analyze form of approximation is interpolation, with the exact and approximating functions sharing the same nodal values on the grid.

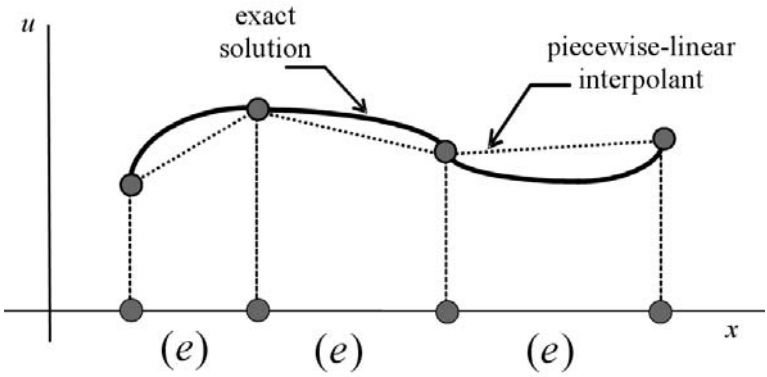


Fig. 3.27. Piecewise-linear FE interpolation of the exact solution.

The FE solution of a boundary value problem in general will *not* interpolate the exact one, although there is a peculiar case where it does (see the Appendix on p. 127). However, due to Céa's theorem (or Galerkin error minimization or the LBB condition, whichever may be applicable), the smallness of the interpolation error guarantees the smallness of the solution error.

It is intuitively clear from Fig. 3.27 that the interpolation error decreases as the mesh size becomes smaller. The error will also decrease if higher-order interpolation – say, piecewise-quadratic – is used. (Higher-order nodal elements have additional nodes that are not shown in the figure.) If the derivative of the exact solution is only piecewise-smooth, the approximation will not suffer as long as the points of discontinuity – typically, material interfaces – coincide with some of the grid nodes. The accuracy will degrade significantly if a material interface boundary passes through a finite element. For this reason, FE meshes in any number of dimensions are generated in such a way that each element lies entirely within one medium. For curved material boundaries, this

is strictly speaking possible only if the elements themselves are curved; nevertheless, approximation of curved boundaries by piecewise-planar element FE surfaces is often adequate in practice.

P.G. Ciarlet & P.A. Raviart gave the following general and powerful mathematical characterization of interpolation accuracy [CR72]. Let Σ be a finite set in $\mathbb{R}^n$ and let polynomial $\mathcal{I}u$ interpolate a given function u, in the Lagrange or Hermite sense, over a given set of points in Σ. Notably, the only significant assumption in the Ciarlet–Raviart theory is uniqueness of such a polynomial. Then

$$\sup\{\|D^m u(x) - D^m \mathcal{I}u(x)\|\,;\ \ x \in K\} \ \le\ CM_{p+1}\frac{h^{p+1}}{\rho^m}, \quad 0 \le m \le p \quad (3.104)$$

Here

K is the closed convex hull of Σ;
h – diameter of K;
p – maximum order of the interpolating polynomial;
$M_{p+1} = \sup\{\|D_{p+1}u(x)\|;\ \ x \in K\}$;
ρ – supremum of the diameters of spheres inscribed in K.
C – a constant.

While the result is applicable to abstract sets, in the FE context K is a finite element (as a geometric figure).

Let us examine the factors that the error depends upon. M_{p+1}, being the magnitude of the $(p+1)$st derivative of u, characterizes the level of smoothness of u; naturally, the polynomial approximation is better for smoother functions. The geometric factor can be split up into the shape and size components:

$$\frac{h^{p+1}}{\rho^m} \ = \ \left(\frac{h}{\rho}\right)^m h^{p+1-m}$$

h/ρ is dimensionless and depends only on the shape of K; we shall return to the dependence of FE errors on element shape in Section 3.14. The following observations about the second factor, h^{p+1-m}, can be made:

- Example: the maximum interpolation error by linear polynomials is $\mathcal{O}(h^2)$ ($p = 1$, $m = 0$). The error in the first derivative is asymptotically higher, $\mathcal{O}(h)$ ($p = 1$, $m = 1$).
- The interpolation error behaves as a power function of element size h but depends *exponentially* on the interpolation order p, provided that the exact solution has at least $p + 1$ derivatives.
- The interpolation accuracy is lower for higher-order derivatives (parameter m).

Most of these observations make clear intuitive sense. A related result is cited in Section 4.4.4 on p. 209.

Appendix: The Ladyzhenskaya–Babuška–Brezzi Condition

For elliptic forms, the Lax–Milgram theorem guarantees well-posedness of the weak problem and Céa's theorem relates the error of the Galerkin solution to the approximation error (Section 3.5 on p. 86). For non-elliptic forms, the Ladyzhenskaya–Babuška–Brezzi (LBB) condition plays a role similar to the Lax–Milgram–Céa results, although analysis is substantially more involved. Conditions for the well-posedness of the weak problem were derived independently by O.A. Ladyzhenskaya, I. Babuška & F. Brezzi [Lad69, BA72, Bre74]. In addition, the Babuška and Brezzi theories provide error estimates for the numerical solution.

Unfortunately, the LBB condition is in many practical cases not easy to verify. As a result, less rigorous criteria are common in engineering practice; for example, the "patch test" that is not considered in this book but is easy to find in the FE literature (e.g. O.C. Zienkiewicz *et al.* [ZTZ05]). Non-rigorous conditions should be used with caution; I. Babuška & R. Narasimhan [BN97] give an example of a finite element formulation that satisfies the patch test but not the LBB condition. They also show, however, that convergence can still be established in that case, provided that the input data (and hence the solution) are sufficiently smooth.

A mathematical summary of the LBB condition is given below for reference. It is taken from the paper by J. Xu & L. Zikatanov [XZ03].

Let U and V be two Hilbert spaces, with inner products $(\cdot,\cdot)_U$ and $(\cdot,\cdot)_V$, respectively. Let $\mathcal{B}(\cdot,\cdot)$: $U \times V \mapsto \mathbb{R}$ be a continuous bilinear form

$$\mathcal{B}(u,v) \;\leq\; \|\mathcal{B}\|\,\|u\|_U\,\|v\|_V \tag{3.105}$$

Consider the following variational problem: Find $u \in U$ such that

$$\mathcal{B}(u,v) \;=\; \langle f, v\rangle, \quad \forall v \in V \tag{3.106}$$

where $f \in V^*$ (the space of continuous linear functionals on V and $\langle\cdot,\cdot\rangle$ is the usual pairing between V^* and V.
... problem (3.106) is well posed if and only if the following conditions hold ...:

$$\inf_{u\in U}\sup_{v\in V}\frac{\mathcal{B}(u,v)}{\|u\|_U\,\|v\|_V} \;>\; 0 \tag{3.107}$$

Furthermore, if (3.107) hold, then

$$\inf_{u\in U}\sup_{v\in V}\frac{\mathcal{B}(u,v)}{\|u\|_U\,\|v\|_V} \;=\; \inf_{v\in V}\sup_{u\in U}\frac{\mathcal{B}(u,v)}{\|u\|_U\,\|v\|_V} \;\equiv\; \alpha \;>\; 0 \tag{3.108}$$

and the unique solution of (3.106) satisfies

$$\|u\|_U \;\leq\; \frac{\|f\|_{V^*}}{\alpha} \tag{3.109}$$

... Let $U_h \subset U$ and $V_h \subset V$ be two nontrivial subspaces of U and V, respectively. We consider the following variational problem: Find $u_h \in U_h$ such that

$$\mathcal{B}(u_h, v_h) \;=\; \langle f, v_h \rangle, \quad \forall v_h \in V_h \tag{3.110}$$

... problem (3.110) is uniquely solvable if and only if the following conditions hold:

$$\inf_{u_h \in U_h} \sup_{v_h \in V_h} \frac{\mathcal{B}(u_h, v_h)}{\|u_h\|_{U_h} \, \|v\|_{V_h}} \;=\; \inf_{v_h \in V_h} \sup_{u_h \in U_h} \frac{\mathcal{B}(u_h, v_h)}{\|u_h\|_{U_h} \, \|v\|_{V_h}} \;\equiv\; \alpha_h \;>\; 0 \tag{3.111}$$

(End of quote from J. Xu & L. Zikatanov [XZ03].)

The LBB result, slightly strengthened by Xu & Zikatanov, for the Galerkin approximation is

Theorem 4. *Let (3.105), (3.107) and (3.111) hold. Then*

$$\|u - u_h\|_U \;\leq\; \frac{\|\mathcal{B}\|}{\alpha_h} \inf_{w_h \in V_h} \|u - w_h\|_U \tag{3.112}$$

Appendix: A Peculiar Case of Finite Element Approximation

The curious special case considered in this Appendix is well known to the expert mathematicians but much less so to applied scientists and engineers. I am grateful to B.A. Shoykhet for drawing my attention to this case many years ago and to D.N. Arnold for insightful comments and for providing a precise reference, the 1974 paper by J. Douglas & T. Dupont [DD74], p. 101.

Consider the 1D Poisson equation

$$-\frac{d^2 u}{dx^2} \;=\; f(x), \quad \Omega = [a, b]; \quad u(a) = u(b) = 0 \tag{3.113}$$

where the zero Dirichlet conditions are imposed for simplicity only. Let us examine the finite element solution u_h of this equation using first-order elements. The Galerkin problem for u_h on a chosen mesh is

$$(u_h', v_h') \;=\; (f, v_h), \quad u_h, \forall v_h \in \mathcal{P}_{0h} \tag{3.114}$$

where the primes denote derivatives and $\mathcal{P}_{0h}$ is the space of continuous functions that are linear within each element (segment) of the chosen grid and satisfy the zero Dirichlet conditions. The inner products are those of L_2.

We know from Section 3.3.1 that the Galerkin solution is the best approximation (in $\mathcal{P}_{0h}$) of the exact solution u^*, in the sense of minimum "energy" $(u_h - u^*, u_h - u^*)$. Geometrically, it is the best (in the same energy sense) representation of the curve $u^*(x)$ by a broken line compatible with a given mesh.

Surprisingly, in the case under consideration the best approximation actually *interpolates* the exact solution; in other words, the nodal values of the exact and numerical solutions are the same. In reference to Fig. 3.27 on p. 124, approximation of the exact solution (solid line) by the the piecewise-linear interpolant (dotted line) on a fixed grid cannot be improved by shifting the dotted line up or down a bit.

Proof. Let us treat v_h in the Galerkin problem (3.114) for u_h as a generalized function (distribution; see Appendix 6.15 on p. 343).[27] Then

$$-\langle u_h,\, v_h''\rangle \;=\; (f,\, v_h), \quad u_h, \forall v_h \in \mathcal{P}_{0h}$$

where the angle brackets denote a linear functional acting on u_h and v_h'' is the second distributional derivative of v_h. This transformation of the left hand side is simply due to the definition of distributional derivative.

The right hand side is transformed in a similar way, after noting that $f = -u''$, where u is the exact solution of the Poisson equation. We obtain

$$\langle u_h,\, v_h''\rangle \;=\; (u,\, v_h'')$$

or

$$\langle u_h - u,\, v_h''\rangle \;=\; 0, \quad \forall v_h \in \mathcal{P}_{0h} \tag{3.115}$$

It remains to be noted that v_h' is a piecewise-constant function[28] and hence v_h'' is a set of Dirac delta-functions residing at the grid nodes. This makes it obvious that (3.115) is satisfied if and only if u_h indeed interpolates the exact solution at the nodes of the grid. □

Exactness of the FE solution at the grid nodes is an extreme particular case of the more general phenomenon of *superconvergence*: the accuracy of the FE solution at certain points (e.g. element nodes or barycenters) is asymptotically higher than the average accuracy. The large body of research on superconvergence includes books, conference proceedings and many journal publications.[29]

[27] The reviewer of this book noted that in a purely mathematical text the use of distributional derivatives would not be appropriate without presenting a rigorous theory first. However, distributions (Dirac delta-functions in particular) make our analysis here much more elegant and simple. I rely on the familiarity of applied scientists and engineers – the intended audience of this book – with delta-functions, even if the usage is not backed up by full mathematical rigor.

[28] With zero mean due to the Dirichlet boundary conditions for v_h, but otherwise arbitrary.

[29] M. Křižek, P. Neittaanmaki & R. Stenberg, eds. *Finite Element Methods: Superconvergence, Post-Processing, and a Posteriori Estimates*, Lecture Notes in Pure and Applied Mathematics, vol. 196, Marcel Dekker: New York, 1998. L.B. Wahlbin, *Superconvergence in Galerkin Finite Element Methods*, Berlin; New York: Springer-Verlag, 1995. M. Křížek, Superconvergence phenomena on three-dimensional meshes, *Int. J. of Num. Analysis and Modeling*, vol. 2, pp. 43–56, 2005. L. Chen has assembled a reference database at http://math.ucsd.edu/c̃long/Paper/html/Superconvergence.html .

3.11 An Overview of System Solvers

The finite element method leads to systems of equations with large matrices – in practice, the dimension of the system can range from thousands to millions. When the method is applied to differential equations, the matrices are *sparse* because each basis function is local and spans only a few neighboring elements; nonzero entries in the FE matrices correspond to the overlapping supports of the neighboring basis functions. (The situation is different when FEM is applied to integral equations. The integral operator is nonlocal and typically all unknowns in the system of equations are coupled; the matrix is full. Integral equations are considered in this book only in passing.)

The sparsity (adjacency) structure of a matrix is conveniently described as a graph. For an $n \times n$ matrix, the graph has n nodes.[30] To each nonzero entry a_{ij} of the matrix there corresponds the graph edge $i - j$. If the structure of the matrix is not symmetric, it is natural to deal with a *directed* graph and distinguish between edges $i \to j$ and $j \to i$ (each of them may or may not be present in the graph, independently of the other one). Symmetric structures can be described by *un*directed graphs.

As an example, the directed graph corresponding to the matrix

$$\begin{pmatrix} 2 & 0 & 3 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ -1 & 0 & 0 & 3 \end{pmatrix} \tag{3.116}$$

is shown in Fig. 3.28. For simplicity, the diagonal entries of the matrix are always tacitly assumed to be nonzero and are not explicitly represented in the graph.

An important question in finite difference and finite element analysis is how to solve such large sparse systems effectively. One familiar approach is Gaussian elimination of the unknowns one by one. As the simplest possible illustration, consider a system of two equations of the form

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \;=\; \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \tag{3.117}$$

For the natural order of elimination of the unknowns (x_1 eliminated from the first equation and substituted into the others, etc.) and for a nonzero a_{11}, we obtain $x_1 = (f_1 - a_{12}x_2)/a_{11}$ and

$$(a_{22} - a_{21}a_{11}^{-1}a_{12})\, x_2 \;=\; f_2 \;-\; a_{11}^{-1} f_1 \tag{3.118}$$

This simple result looks innocuous at first glance but in fact foreshadows a problem with the elimination process. Suppose that in the original system

[30] For matrices arising in finite difference or finite element methods, the nodes of the graph typically correspond to mesh nodes; otherwise graph nodes are abstract mathematical entities.

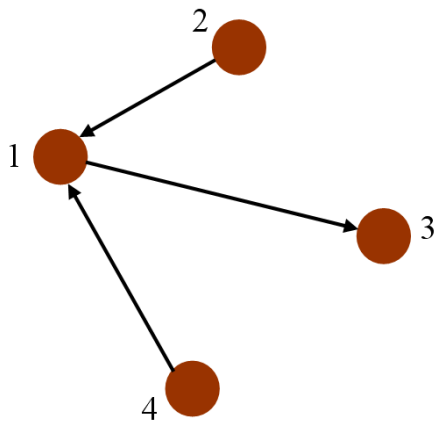


Fig. 3.28. Matrix sparsity structure as a graph: an example.

(3.117) the diagonal entry a_{22} is zero. In the transformed system (3.118) this is no longer so: the entry corresponding to x_2 (the only entry in the remaining 1×1 matrix) is $a_{22} - a_{21}a_{11}^{-1}a_{12}$. Such transformation of zero matrix entries into nonzeros is called "fill-in". For the simplistic example under consideration, this fill-in is of no practical consequence. However, for large sparse matrices, fill-in tends to accumulate in the process of Gaussian elimination and becomes a serious complication.

In our 2×2 example with $a_{22} = 0$, the fill-in disappears if the order of equations (or equivalently the sequence of elimination steps) is changed:

$$\begin{pmatrix} a_{21} & 0 \\ a_{11} & a_{12} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}$$

Obviously, x_1 is now found immediately from the first equation, and x_2 is computed from the second one, with no additional nonzero entries created in the process. In general, permutations of rows and columns of a sparse matrix may have a dramatic effect on the amount of fill-in, and hence on the computational cost and memory requirements, in Gaussian elimination.

Gaussian elimination is directly linked to matrix factorization into lower- and upper-triangular terms. More specifically, the first factorization step can be represented in the following form:

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & & & \\ \dots & & A_1 & \\ a_{n1} & & & \end{pmatrix} = \begin{pmatrix} l_{11} & 0 & \dots & 0 \\ l_{21} & & & \\ \dots & & L_1 & \\ l_{n1} & & & \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ 0 & & & \\ \dots & & U_1 & \\ 0 & & & \end{pmatrix} \tag{3.119}$$

The fact that this factorization is possible (and even not unique) can be verified by direct multiplication of the factors in the right hand side. This

yields, for the first diagonal element, first column and first row, respectively, the following conditions:

$$l_{11}u_{11} = a_{11}$$

$$l_{21}u_{11} = a_{21}, \quad l_{31}u_{11} = a_{31}, \quad \ldots, \quad l_{n1}u_{11} = a_{n1}$$

$$l_{11}u_{12} = a_{12}, \quad l_{11}u_{13} = a_{13}, \quad \ldots, \quad l_{11}u_{1n} = a_{1n}$$

where n is the dimension of matrix A. Fixing l_{11} by, say, setting it equal to one defines the column vector $l_1 = (l_{11}, l_{21}, \ldots, l_{n1})^T$ and the row vector $u_1^T = (u_{11}, u_{12}, \ldots, u_{1n})$ unambiguously:

$$l_{11} = 1; \quad u_{11} = a_{11} \tag{3.120}$$

$$l_{21} = u_{11}^{-1}a_{21}, \quad l_{31} = u_{11}^{-1}a_{31}, \ldots, \quad l_{n1} = u_{11}^{-1}a_{n1} \tag{3.121}$$

$$u_{12} = a_{12}, \quad u_{13} = a_{13}, \ldots, \quad u_{1n} = a_{1n} \tag{3.122}$$

Further, the condition for matrix blocks L_1 and U_1 follows directly from factorization (3.119):

$$L_1U_1 + l_1u_1^T = A_1$$

or equivalently

$$L_1U_1 = \tilde{A}_1$$

where

$$\tilde{A}_1 \equiv A_1 - l_1u_1^T$$

The updated matrix $\tilde{A}_1$ is a particular case of the *Schur complement* (R.A. Horn & C.R. Johnson [HJ90], Y. Saad [Saa03]). Explicitly the entries of $\tilde{A}_1$ can be written as

$$\tilde{a}_{1,ij} = a_{ij} - l_{i1}u_{1j} = a_{ij} - a_{i1}a_{11}^{-1}a_{1j} \tag{3.123}$$

Thus the first step of Gaussian factorization $A = LU$ is accomplished by computing the first column of L (3.120), (3.121), the first row of U (3.120), (3.122) and the updated block $\tilde{A}_1$ (3.123). The factorization step is then repeated for $\tilde{A}_1$, etc., until (at the n-th stage) the trivial case of a 1×1 matrix results. Theoretically, it can be shown that this algorithm succeeds as long as all leading minors of the original matrix are nonzero. In practical computation, however, care should be taken to ensure computational stability of the process (see below).

Once the matrix is factorized, solution of the original system of equations reduces to forward elimination and backward substitution, i.e. to solving systems with the triangular matrices L and U, which is straightforward. An important advantage of Gaussian elimination is that, once matrix factorization has been performed, equations with the same matrix but multiple right hand sides can be solved at the very little cost of forward elimination and backward substitution only.

Let us review a few computational aspects of Gaussian elimination.

1. **Fill-in.** The matrix update formula (3.123) clearly shows that a zero matrix entry a_{ij} can become nonzero in the process of LU-factorization. The 2×2 example considered above is the simplest possible case of such fill-in. A quick look at the matrix update equation (3.123) shows how the fill-in is reflected in the directed sparsity graph. If at some step of the process node k is being eliminated, any two edges $i \to k$ and $k \to j$ produce a new edge $i \to j$ (corresponding to a new nonzero matrix entry ij). This is reminiscent of the usual "head-to-tail" rule of vector addition. Fig. 3.29 may serve as an illustration. Similar considerations apply for symmetric sparsity structures represented by undirected graphs. Methods to reduce fill-in are discussed below.
2. **The computational cost.** For *full* matrices, the number of arithmetic operations (multiplications and additions) in LU-factorization is approximately $2n^3/3$. For sparse matrices, the cost depends very strongly on the adjacency structure and can be reduced dramatically by clever permutations of rows and columns of the matrix and other techniques reviewed later in this section. [31]
3. **Stability.** Detailed analysis of LU factorization (J.H. Wilkinson [Wil94], G.H. Golub & C.F. Van Loan [GL96], G.E. Forsythe & C.B. Moler [FM67], N.J. Higham [Hig02]) shows that numerical errors (due to roundoff) can accumulate if the entries of L and U grow. Such growth can, in turn, be traced back to small diagonal elements arising in the factorization process. To rectify the problem, the leading diagonal element at each step of factorization is maximized either via *complete pivoting* – reshuffling of rows and columns of the remaining matrix block – or via *partial pivoting* – reshuffling of rows only. The existing theoretical error estimates for both types of pivoting are much more pessimistic than practical experience indicates.[32]

---

[31] Incidentally, the $\mathcal{O}(n^3)$ operation count is *not* asymptotically optimal for solving large systems with *full* matrices of size $n \times n$. In 1969, V. Strassen discovered a trick for computing the product of two 2×2 block matrices with seven block multiplications instead of eight that would normally be needed [Str69]. When applied recursively, this idea leads to $\mathcal{O}(n^\gamma)$ operations, with $\gamma = \log_2 7 \approx 2.807$. Theoretically, algorithms with γ as low as 2.375 now exist, but they are computationally unstable and have very large numerical prefactors that make such algorithms impractical. I. Kaporin has developed practical (i.e. stable and faster than straightforward multiplication for matrices of moderate size) algorithms with the asymptotic operation count $\mathcal{O}(N^{2.7760})$ [Kap04]. Note that solution of algebraic systems with full matrices can be reduced to matrix multiplication (V. Pan [Pan84]). See also S. Robinson [Rob05] and H. Cohn *et al.* [CKSU05].

[32] J.H. Wilkinson [Wil61] showed that for complete pivoting the growth factor for the numerical error does not exceed

$$n^{1/2}(2^1 \times 3^{1/2} \times 4^{1/3} \times \ldots \times n^{1/(n-1)})^{1/2} \sim Cn^{0.25 \log n}$$

(which is ~ 3500 for $n = 100$ and $\sim 8.6 \times 10^6$ for $n = 1000$). In practice, however, there are no known matrices with this growth factor higher than n. For partial

> In fact, partial pivoting works so well in practice that it is used almost exclusively: higher stability of complete pivoting is mostly theoretical but its higher computational cost is real. Likewise, orthogonal factorizations such as QR, while theoretically more stable than LU-factorization, are hardly ever used as system solvers because their computational cost is approximately twice that of LU.[33] L.N. Trefethen [Tre85] gives very interesting comments on this and related matters.

Remarkably, the modern use of Gaussian elimination can be traced back to a single 1948 paper by A.M. Turing[34] [Tur48, Bri92]. N.J. Higham writes ([Hig02], pp. 184–185):

> " [Turing] formulated the ... LDU factorization of a matrix, proving [that the factorization exists and is unique if all leading minors of the matrix are nonzero] and showing that Gaussian elimination computes an LDU factorization. He introduced the term "condition number" ... He used the word "preconditioning" to mean improving the condition of a system of linear equations (a term that did not come into popular use until the 1970s). He described iterative refinement for linear systems. He exploited backward error ideas. ... he analyzed Gaussian elimination with partial pivoting for general matrices and obtained [an error bound]. "

The case of sparse symmetric positive definite (SPD) systems has been studied particularly well, for two main reasons. First, such systems are very common and important in both theory and practice. Second, it can be shown that the factorization process for SPD matrices is always numerically stable (A. George & J.W.H. Liu [GL81], G.H. Golub & C.F. Van Loan [GL96], G.E. Forsythe & C.B. Moler [FM67]). Therefore one need not be concerned with pivoting (permutations of rows and columns in the process of factorization) and can concentrate fully on minimizing the fill-in.

The general case of nonsymmetric and/or non-positive definite matrices will not be reviewed here but is considered in several monographs: books by O. Østerby & Z. Zlatev [sZZ83] and by I.S. Duff *et al.* [DER89], as well as a much more recent book by T.A. Davis [Dav06].

The remainder of this section deals exclusively with the SPD case and is, in a sense, a digest of the excellent treatise by A. George & J.W.H. Liu [GL81]. For SPD matrices, it is easy to show that in the LU factorization U can be

---

pivoting, the bound is 2^{n-1}, and this bound can in fact be reached in some exceptional cases.

[33] QR algorithms are central in eigenvalue solvers; see Appendix 7.15 on p. 478.

[34] Alan Mathison Turing (1912–1954), the legendary inventor of the Turing machine and the *Bombe* device that broke (with an improvement by Gordon Welchman) the German Enigma codes during World War II. Also well known is the *Turing test* that defines a "sentient" machine. Overall, Turing lay the foundation of modern computer science. See http://www.turing.org.uk/turing

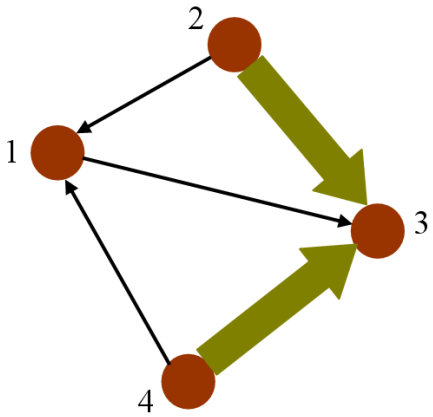


Fig. 3.29. Block arrows indicate fill-in created in a matrix after elimination of unknown #1.

taken as L^T, leading to Cholesky factorization LL^T already mentioned on p. 113. Cholesky decomposition has a small overhead of computing the square roots of the diagonal entries of the matrix; this overhead can be avoided by using the LDL^T factorization instead (where D is a diagonal matrix).

Methods for reducing fill-in are based on reordering of rows and columns of the matrix, possibly in combination with block partitioning. Let us start with the permutation algorithms.

The simplest case where the sparsity structure can be exploited is that of *banded* matrices. The *band* implies part of the matrix between two sub-diagonals parallel to the main diagonal or, more precisely, the set of entries with indexes i, j such that $-k_1 \leq i - j \leq k_2$, where $k_{1,2}$ are nonnegative integers. A matrix is banded if its entries are all zero outside a certain band (in practice, usually $k_1 = k_2 = k$). The importance of this notion for Gaussian (or Cholesky) elimination lies in the easily verifiable fact that the band structure is preserved during factorization, i.e. no additional fill is created outside the band. Cholesky decomposition for a band matrix requires approximately $k(k+3)n/2$ multiplicative operations, which for $k \ll n$ is much smaller than the number of operations needed for the decomposition of a full matrix $n \times n$.

A very useful generalization is to allow the width of the band to vary row-by-row: $k = k(i)$. Such a variable-width band is called an *envelope*. Figs. 3.22 (p. 115) and 3.23 may serve as a helpful illustration. Again, no fill is created outside the envelope. Since the minimal envelope is obviously a subset of the minimal band, the computational cost of the envelope algorithm is generally lower than that of the band method.[35] The operation count for the envelope

[35] I disregard the small overhead related to storage and retrieval of matrix entries in the band and envelope.

method can be found in George & Liu's book [GL81], along with a detailed description and implementation of the *Reverse Cuthill–McKee* ordering algorithm that reduces the envelope size.

There is no known algorithm that would minimize the computational cost and/or memory requirements for a matrix with any given sparsity structure, even if pivoting is not involved, and whether or not the matrix is SPD. D.J. Rose & R.E. Tarjan [RT75] state (but do not include the proof) that this problem for a non-SPD matrix is NP-complete and conjecture that the same is true in the SPD case.

However, powerful heuristic algorithms are available, and the underlying ideas are clear from adjacency graph considerations. Fig. 3.30 shows a small fragment of the adjacency graph; thick lines in Fig. 3.31 represent the corresponding fill-in if node #1 is eliminated first. These figures are very similar to Figs. 3.28 and 3.29, except that the graph for a symmetric structure is unordered.

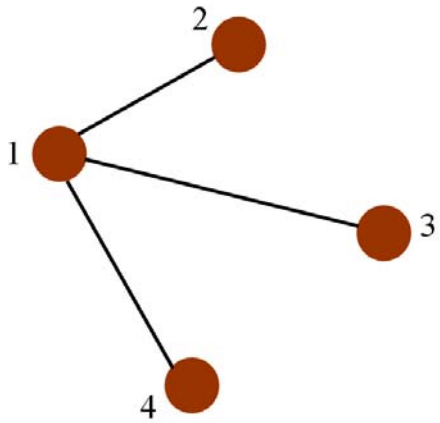


Fig. 3.30. Symmetric sparsity structure as a graph: an example.

Elimination of a node couples all the nodes to which it is connected. If nodes 2, 3 and 4 were to be eliminated *prior* to node 1, there would be no fill-in in this fragment of the graph. This simple example has several ramifications.

First, a useful heuristic is to start the elimination with the graph vertices that have the fewest number of neighbors, i.e. the minimum degree. (Degree of a vertex is the number of edges incident to it.) The minimum degree algorithm, first introduced by W.F. Tinney & J.W. Walker [TW67], is quite useful and effective in practice, although there is of course no guarantee that local minimization of fill-in at each step of factorization will lead to global optimization of the whole process. George & Liu [GL81] describe the Quotient

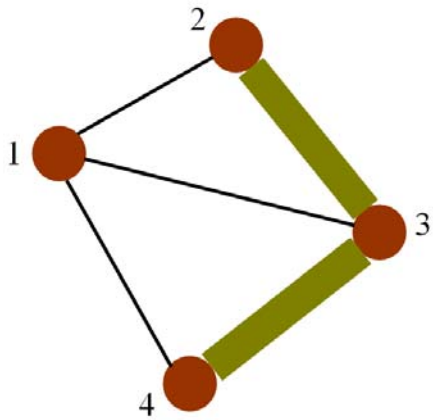


Fig. 3.31. Fill-in (block arrows) created in a matrix with symmetric sparsity structure after elimination of unknown #1.

Minimum Degree (QMD) method, an efficient algorithmic implementation of MD in the SPARSPAK package that they developed.

Second, it is obvious from Fig. 3.31 that elimination of the root of a tree in a graph is disastrous for the fill-in. The opposite is true if one starts with the leaves of the tree. This observation may not seem practical at first glance, as adjacency graphs in FEM are very far from being trees.[36] What makes the idea useful is *block* factorization and partitioning.

Suppose that graph G (or, almost equivalently, the finite element mesh) is split into two parts G_1 and G_2 by a separator S, so that $G = G_1 \bigcup G_2 \bigcup S$ and $G_1 \bigcap G_2 = \emptyset$; this corresponds to block partitioning of the system matrix. The partitioning has a tree structure, with the separator as the root and $G_{1,2}$ as the leaves. The system matrix has the following block form:

$$L = \begin{pmatrix} L_{G1} & 0 & L_{G1,S} \\ 0 & L_{G2} & L_{G2,S} \\ L_{G1,S}^T & L_{G2,S}^T & L_S \end{pmatrix} \tag{3.124}$$

Elimination of block L_{G1} leaves the zero blocks unchanged, i.e. does not – on the block level – generate any fill in the matrix. For comparison, if the "root" block L_S were eliminated first (quite unwisely), zero blocks would be filled.

George & Liu [GL81, GL89] describe two main partitioning strategies: One-Way Dissection (1WD) and Nested Dissection (ND). In 1WD, the graph is partitioned by several dissecting lines that are, if viewed as geometric objects

[36] For first order elements in FEM, the mesh itself can be viewed as the sparsity graph of the system matrix, element nodes corresponding to graph vertices and element edges to graph edges. For a 2D triangular mesh with n nodes, the number of edges is approximately $2n$, whereas for a tree it is $n - 1$.

on the FE mesh, approximately "parallel".[37] Taken together, the separators form the root of a tree structure for the block matrix; the remaining disjoint blocks are the leaves of the tree. Elimination of the leaves generates fill-in in the root block, which is acceptable as long as the size of this block is moderate. To get an idea about the computational savings of 1WD as compared to the envelope method, one may consider an $m \times l$ rectangular grid ($m < l$) in 2D[38] and optimize the number of operations or, alternatively, memory requirements with respect to the chosen number of separators, each separator being a grid line with m nodes. The end result is that the memory in 1WD can be $\sim \sqrt{6/m}$ times smaller than for the envelope method [GL81]. For example, if $m = 100$, the savings are by about a factor of four ($\sqrt{6/100} \approx 0.25$).

A typical ND separator in 2D can geometrically be pictured as two lines, horizontal and vertical, that split the graph into four approximately equal parts. The procedure is then applied recursively to each of the disjoint subgraphs. For a regular $m \times m$ grid in 2D, one can write a recursive relationship for the amount of computer memory $M_{\mathrm{ND}}(m)$ needed for ND; this ultimately yields [GL81]

$$M_{\mathrm{ND}}(m) \;=\; \frac{31}{4}\, m^2 \log_2 m \;+\; \mathcal{O}(m^2)$$

Hence for 2D problems ND is asymptotically almost optimal in terms of its memory requirements: the memory is proportional to the number of nodes times a relatively mild logarithmic factor. However, the computational cost is *not* optimal even for 2D meshes: the number of multiplicative operations is approximately

$$\frac{829}{84}\, m^3 \;+\; \mathcal{O}(m^2 \log_2 m)$$

That is, the computational cost grows as the number of nodes n to the power of 1.5.

Performance of direct solvers further deteriorates in three dimensions. For example, the computational cost and memory for ND scale as $\mathcal{O}(n^2)$ and $\mathcal{O}(n^{4/3})$, respectively, when the number of nodes n is large. Some improvement has been achieved by combining the ideas of 1WD, ND and QMD, with a recursive application of multisection partitioning of the graph. These algorithms are implemented in the SPOOLES software package[39] developed by C. Ashcraft, R. Grimes, J. Liu and others [AL98, AG99]. For illustration, Fig. 3.32 shows the number of nonzero entries in the Cholesky factor for several ordering algorithms as a function of the number of nodes in the finite element mesh. This data is for the scalar electrostatic equation in a cubic

[37] The separators need not be straight lines, as their construction is topological (based on the sparsity graph) rather than geometric. The word "parallel" therefore should not be taken literally.

[38] A similar estimate can also be easily obtained for 3D problems, but in that case 1WD is not very efficient.

[39] SParse Object Oriented Linear Equations Solver, netlib.org/linalg/spooles-/spooles.2.2.html

domain; Nested Dissection and one of the versions of Multistage Minimum Degree from the SPOOLES package perform better than other methods in this case.[40]

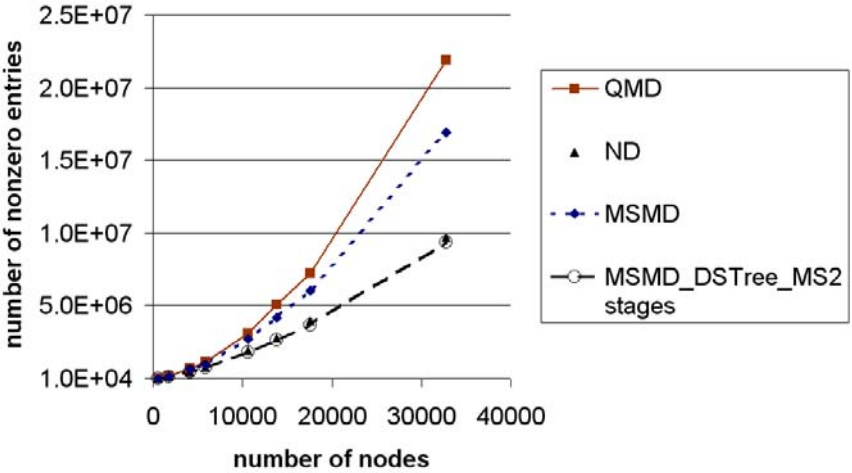


Fig. 3.32. Comparison of memory requirements (number of nonzero entries in the Cholesky factor) as a function of the number of finite element nodes for the scalar electrostatic equation in a cubic domain. Algorithms: Quotient Minimum Degree, Nested Dissection and two versions of Multistage Minimum Degree from the SPOOLES package.

The limitations of direct solvers for 3D finite element problems are apparent, the main bottleneck being memory requirements due to the fill in the Cholesky factor (or the *LU* factors in the nonsymmetric case): tens of millions of nonzero entries for meshes of fairly moderate size, tens of thousands of nodes. The difficulties are exacerbated in vector problems, in particular the ones that arise in electromagnetic analysis in 3D.

Therefore for many 3D problems, and for some large 2D problems, *iterative* solvers are indispensable, their key advantage being a very limited amount of extra memory required.[41] In comparison with direct solvers, iterative ones are arguably more diverse, more dependent on the algebraic properties of matrices, and would require a more wide-ranging review and explanation. To avoid sidetracking the main line of our discussion in this chapter, I refer the reader to the excellent monographs and review papers on iterative solvers by

[40] I thank Cleve Ashcraft for his detailed replies to my questions on the usage of SPOOLES 2.2 when I ran this and other tests in the Spring of 2000.

[41] Typically several auxiliary vectors in Krylov subspaces and sparse preconditioners need to be stored; see references below.

Y. Saad & H.A. van der Vorst [Saa03, vdV03b, SvdV00], L.A. Hageman & D.M. Young [You03, HY04], and O. Axelsson [Axe96].

3.12 Electromagnetic Problems and Edge Elements

3.12.1 Why Edge Elements?

In electromagnetic analysis and a number of other areas of physics and engineering, the unknown functions are often vector rather than scalar fields. A straightforward finite element model would involve approximation of the Cartesian components of the fields. This approach was historically the first to be used and is still in use today. However, it has several flaws – some of them obvious and some hidden.

An obvious drawback is that nodal element discretization of the Cartesian components of a field leads to a *continuous* approximation throughout the computational domain. This is inconsistent with the *dis*continuity of some field components – in particular, the normal components of $\mathbf{E}$ and $\mathbf{H}$ – at material boundaries. The treatment of such conditions by nodal elements is possible but rather awkward: the interface nodes are "doubled," and each of the two coinciding nodes carries the field value on one side of the interface boundary. Constraints then need to be imposed to couple the Cartesian components of the field at the double nodes; the algorithm becomes inelegant.

Although this difficulty is more of a nuisance than a serious obstacle for implementing the component-wise formulation, it is also an indication that something may be "wrong" with this formulation on a more fundamental level (more about that below).

So-called "spurious modes" – the hidden flaw of the component-wise treatment – were noted in the late 1970s and provide further evidence of some fundamental limitations of Cartesian approximation. These modes are frequently branded as "notorious," and indeed hundreds of papers have been published on this subject.[42]

As a representative example, consider the computation of the eigenfrequencies ω and the corresponding electromagnetic field modes in a cavity resonator. The resonator is modeled as a simply connected domain Ω with perfectly conducting walls $\partial\Omega$. The governing equation for the electric field is

$$\nabla \times \mu^{-1} \nabla \times \mathbf{E} \; - \; \omega^2 \epsilon \mathbf{E} \; = \; 0 \;\; \text{in } \Omega; \quad \mathbf{n} \times \mathbf{E} = 0 \;\; \text{on } \partial\Omega \tag{3.125}$$

where the standard notation for the electromagnetic material parameters μ, ϵ and for the exterior normal $\mathbf{n}$ to the domain boundary $\partial\Omega$ is used. The ideally

[42] 320 ISI database references at the end of 2006 for the term "spurious modes". This does not include alternative relevant terminology such as spectral convergence, spurious-free approximation, "vector parasites," etc., so the actual number of papers is much higher.

conducting walls cause the tangential component of the electric field to vanish on the boundary.

Mathematically, the proper functional space for this problem is $H_0(\mathrm{curl}, \Omega)$ – the space of square-integrable vector functions with a square-integrable curl and a vanishing tangential component at the boundary:

$$H_0(\mathrm{curl}, \Omega) \equiv \{\mathbf{E} : \mathbf{E} \in \mathbf{L}_2(\Omega), \ \nabla \times \mathbf{E} \in \mathbf{L}_2(\Omega), \ \mathbf{n} \times \mathbf{E} = 0 \text{ on } \partial\Omega\} \quad (3.126)$$

The weak formulation is obtained by inner-multiplying the eigenvalue equation by an arbitrary test function $\mathbf{E}' \in H_0(\mathrm{curl}, \Omega)$:

$$(\nabla \times \mu^{-1} \nabla \times \mathbf{E}, \mathbf{E}') - \omega^2 (\epsilon \mathbf{E}, \mathbf{E}') = 0, \quad \forall \mathbf{E}' \in H_0(\mathrm{curl}, \Omega) \quad (3.127)$$

where the inner product is that of $\mathbf{L}_2(\Omega)$, i.e.

$$(\mathbf{X}, \mathbf{Y}) \equiv \int_\Omega \mathbf{X} \cdot \mathbf{Y} \, d\Omega$$

for vector fields $\mathbf{X}$ and $\mathbf{Y}$ in $H_0(\mathrm{curl}, \Omega)$.

Using the vector calculus identity

$$\nabla \cdot (\mathbf{X} \times \mathbf{Y}) = \mathbf{Y} \cdot \nabla \times \mathbf{X} - \mathbf{X} \cdot \nabla \times \mathbf{Y} \quad (3.128)$$

with $\mathbf{X} = \mu^{-1} \nabla \times \mathbf{E}$, $\mathbf{Y} = \mathbf{E}'$, equation (3.127) can be integrated by parts to yield

$$(\mu^{-1} \nabla \times \mathbf{E}, \nabla \times \mathbf{E}') - \omega^2 (\epsilon \mathbf{E}, \mathbf{E}') = 0, \quad \forall \mathbf{E}' \in H_0(\mathrm{curl}, \Omega) \quad (3.129)$$

(It is straightforward to verify that the surface integral resulting from to the left hand side of (3.128) vanishes, due to the fact that $\mathbf{n} \times \mathbf{E}' = 0$ on the wall.)

The discrete problem is obtained by restricting $\mathbf{E}$ and $\mathbf{E}'$ to a finite element subspace of $H_0(\mathrm{curl}, \Omega)$; a "good" way of constructing such a subspace is the main theme of this section. The mathematical theory of convergence for the eigenvalue problem (3.129) is quite involved and well beyond the scope of this book;[43] however, some uncomplicated but instructive observations can be made.

The continuous eigenproblem in its strong form (3.125) guarantees, for nonzero frequencies, zero divergence of the $\mathbf{D}$ vector ($\mathbf{D} = \epsilon \mathbf{E}$). This immediately follows by applying the divergence operator to the equation. For the weak formulation (3.129), the zero-divergence condition is satisfied in the generalized form (see Appendix 3.17 on p. 186):

$$(\epsilon \mathbf{E}, \nabla \phi') = 0 \quad (3.130)$$

This follows by using, as a particular case, an arbitrary curl-free test function $\mathbf{E}' = \nabla \phi'$ in (3.129).[44]

[43] References: the book by P. Monk [Mon03], papers by P. Monk & L. Demkowicz [MD01], D. Boffi *et al.* [BFea99, Bof01] and S. Caorsi *et al.* [CFR00].

[44] The equivalence between curl-free fields and gradients holds true for simply connected domains.

It is now intuitively clear that the divergence-free condition will be correctly imposed in the discrete (finite element) formulation if the FE space contains a "sufficiently dense"[45] population of gradients $\mathbf{E}' = \nabla\phi'$. This argument was articulated for the first time (to the best of my knowledge) by A. Bossavit in 1990 [Bos90].

From this viewpoint, a critical deficiency of component-wise nodal approximation is that the corresponding FE space does *not* ordinarily contain "enough" gradients. The reason for that can be inferred from Fig. 3.33 (2D illustration for simplicity). Suppose that there exists a function ϕ vanishing outside a small cluster of elements and such that its gradient is in $\mathcal{P}_1^3$ – i.e. continuous throughout the computational domain and linear within each element. It is clear that ϕ must be a piecewise-quadratic function of coordinates. Furthermore, since $\nabla\phi$ vanishes on the outer side of edge 23, due to the continuity of the gradient along that edge ϕ can only vary in proportion to n_{23}^2 within element 123, where n_{23} is the normal to edge 23. Similarly, ϕ must be proportional to n_{34}^2 in element 134. However, these two quadratic functions are incompatible along the common edge 13 of these two elements, unless the normals n_{23} and n_{34} are parallel.

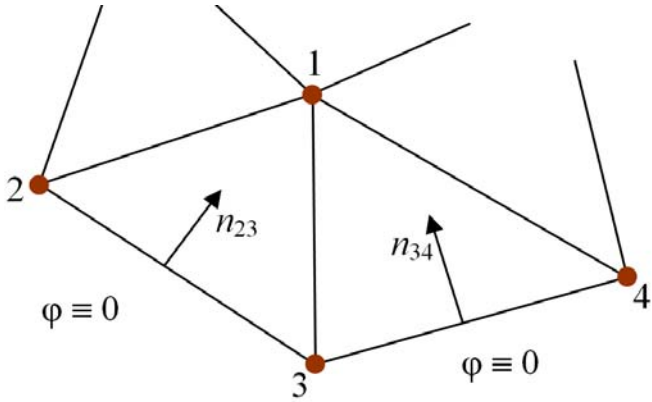


Fig. 3.33. A fragment of a 2D finite element mesh. A piecewise-quadratic function ϕ vanishes outside a cluster of elements. For $\nabla\phi$ to be continuous, ϕ must be proportional to n_{23}^2 within element 123 and to n_{34}^2 within element 134. However, these quadratic functions are incompatible on the common edge 13, unless the normals n_{23} and n_{34} are parallel.

This observation illustrates very severe constraints on the construction of irrotational continuous vector fields that would be piecewise-linear on a given FE mesh. As a result, the FE space does not contain a representative set of

[45] The quotation marks are used as a reminder that this analysis does not have full mathematical rigor.

gradients for the divergence-free condition to be enforced even in weak form. Detailed mathematical analysis and practical experience indicate that this failure to impose the zero divergence condition on the **D** vector usually leads to nonphysical solutions.

The argument presented above is insightful but from a rigorous mathematical perspective incomplete. A detailed analysis can be found in the literature cited in footnote 43 on p. 140. For our purposes, the important conclusion is that the lack of spectral convergence (i.e. the appearance of "spurious modes") is inherent in component-wise finite element approximation of vector fields. Attempts to rectify the situation by imposing additional constraints on the divergence, penalty terms, etc., have had only limited success.

A radical improvement can be achieved by using *edge elements* described in Section 3.12.2 below. As we shall see, the approximation provided by these elements is, in a sense, more "physical" that the component-wise representation of vector fields; the corresponding mathematical structures also prove to be quite elegant.

3.12.2 The Definition and Properties of Whitney-Nédélec Elements

As became apparent in Section 3.8.1 on p. 108 and in Section 3.9 on p. 123, a natural coordinate system for triangular and tetrahedral elements is formed by the barycentric coordinates λ_α ($\alpha = 1, 2, 3$ for triangles and $\alpha = 1, 2, 3, 4$ for tetrahedra). Each function λ is linear and equal to one at one of the nodes and zero at all other nodes. Since the barycentric coordinates play a prominent role in the finite element approximation of scalar fields, it is sensible to explore how they can be used to approximate *vector* fields as well, and not in the component-wise sense.

Remark 5. The most mathematically sound framework for the material of this section is provided by the differential-geometric treatment of physical fields as differential forms rather than vector fields. A large body of material – well written and educational – can be found on A. Bossavit's website.[46] (Bossavit is an authority in this subject area and one of the key developers and proponents of edge element analysis.) Other references are cited in Section 3.12.4 on p. 146 and in Section 3.16 on p. 184. While differential geometry is a standard tool for mathematicians and theoretical physicists, it is not so for many engineers and applied scientists. For this reason, only regular vector calculus is used in this section and in the book in general; this is sufficient for our purposes.

Natural "vector offspring" of the barycentric coordinates are the gradients $\nabla\lambda_\alpha$. These, however, are constant within each element and can therefore represent only piecewise-constant and – even more importantly – only irrotational vector fields. Next, we may consider products $\psi_{\alpha\beta}^{6-12} = \lambda_\alpha \nabla\lambda_\beta$; it

[46] http://www.lgep.supelec.fr/mocosem/perso/ab/bossavit.html

is sufficient to restrict them to $\alpha \neq \beta$ because the gradients are linearly dependent, $\sum \nabla\lambda_\alpha = 0$. Superscript "6-12" indicates that there are six such functions for a triangle and 12 for a tetrahedron. A little later, we shall consider a two times smaller set $\psi^{3-6}_{\alpha\beta}$.

It almost immediately transpires that these new vector functions have one of the desired properties: their tangential components are continuous across element facets (edges for triangles and faces for tetrahedra), while their normal components are in general discontinuous. The most elegant way to demonstrate the tangential continuity is by noting that the generalized curl $\nabla \times \psi^{6-12}_{\alpha\beta} = \nabla \times (\lambda_\alpha \nabla\lambda_\beta) = \nabla\lambda_\alpha \times \nabla\lambda_\beta$ is a regular function, not only a distribution, because the λs are continuous.[47] (A jump in the tangential component would result in a Dirac-delta term in the curl; see Appendix 3.17 on p. 186 and formula (3.215) in particular.)

The tangential components can also be examined more explicitly. The circulation of $\psi^{6-12}_{\alpha\beta}$ over the corresponding edge $\alpha\beta$ is

$$\int_{\text{edge } \alpha\beta} \psi^{6-12}_{\alpha\beta} \cdot \hat{\tau}_{\alpha\beta}\, d\tau \;=\; \int_{\text{edge } \alpha\beta} \lambda_\alpha \nabla\lambda_\beta \cdot \hat{\tau}_{\alpha\beta}\, d\tau$$

$$= \nabla\lambda_\beta \cdot \hat{\tau}_{\alpha\beta} \int_{\text{edge } \alpha\beta} \lambda_\alpha\, d\tau \;=\; \frac{1}{l_{\alpha\beta}} \frac{1}{2} l_{\alpha\beta} \;=\; \frac{1}{2} \tag{3.131}$$

where $\hat{\tau}_{\alpha\beta}$ is the unit edge vector pointing from node α to node β, and $l_{\alpha\beta}$ is the edge length. In the course of the transformations above, it was taken into account that (i) $\nabla\lambda_\beta$ is a (vector) constant, (ii) λ_α is a function varying from zero to one linearly along the edge, so that the component of its gradient along the edge is $1/l_{\alpha\beta}$ and the mean value of λ_α over the edge $\alpha\beta$ is $1/2$.

Thus the circulation of each function $\psi^{6-12}_{\alpha\beta}$ is equal to $1/2$ over its respective edge $\alpha\beta$ and (as is easy to see) zero over all other edges.

One type of edge element is defined by introducing (i) the functional space spanned by the $\psi^{6-12}_{\alpha\beta}$ basis, and (ii) a set of degrees of freedom, two per edge: the tangential components $E_{\alpha\beta}$ of the field (say, electric field $\mathbf{E}$) at each node α along each edge $\alpha\beta$ emanating from that node. The number of degrees of freedom and the dimension of the functional space are six for triangles and 12 for tetrahedra. It is not difficult to verify that the space in fact coincides with the space of linear vector functions within the element. A major difference, however, is that the basis functions for edge elements are only tangentially continuous, in contrast with fully continuous component-wise approximation by nodal elements. The FE representation of the field within the edge element is

$$\mathbf{E}_h \;=\; \sum_{\alpha\neq\beta} E_{\alpha\beta} \psi^{6-12}_{\alpha\beta}$$

[47] Here each barycentric coordinate is viewed as a function defined in the whole domain, continuous everywhere but nonzero only over a cluster of elements sharing the same node.

An interesting alternative is obtained by observing that each pair of functions $\psi_{\alpha\beta}^{6-12}$, $\psi_{\beta\alpha}^{6-12}$ have similar properties: their circulations along the respective edge (but *taken in the opposite directions*) are the same, and their curls are opposite. It makes sense to combine each pair into one new function as

$$\psi_{\alpha\beta}^{3-6} \equiv \psi_{\alpha\beta}^{6-12} - \psi_{\beta\alpha}^{6-12} = \lambda_\alpha \nabla\lambda_\beta - \lambda_\beta \nabla\lambda_\alpha \tag{3.132}$$

It immediately follows from the properties of $\psi_{\alpha\beta}^{6-12}$ that the circulation of $\psi_{\alpha\beta}^{3-6}$ is one along its respective edge (in the direction from node α to node β) and zero along all other edges.

The FE representation of the field is almost the same as before

$$\mathbf{E}_h = \sum_{\alpha\neq\beta} c_{\alpha\beta}\psi_{\alpha\beta}^{3-6}$$

except that summation is now over a twice smaller set of basis functions, one per edge: three for triangles and six for tetrahedra; $c_{\alpha\beta}$ are the circulations of the field along the edges.

Fig. 3.34 helps to visualize two such functions for a triangular element; for tetrahedra, the nature of these functions is similar. Their rotational character is obvious from the figure, the curls being equal to

$$\nabla \times \psi_{\alpha\beta}^{3-6} = 2\nabla\lambda_\alpha \times \nabla\lambda_\beta$$

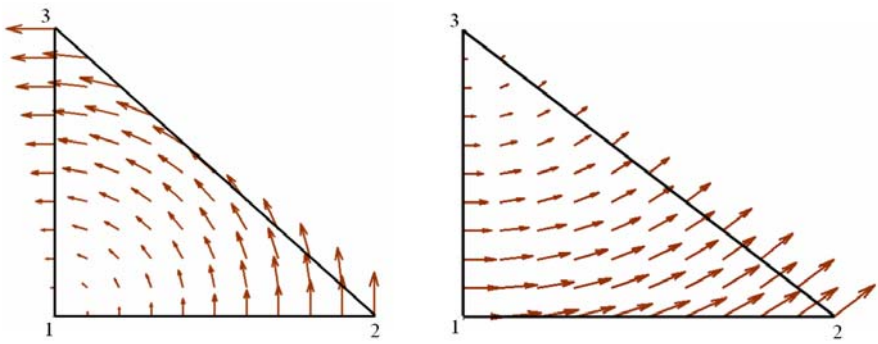


Fig. 3.34. Two basis functions ψ^{3-6} visualized for a triangular element: ψ_{23}^{3-6} (left) and ψ_{12}^{3-6} (right).

The (generalized) divergence of these vector basis functions (see Appendix 3.17, p. 186) is also of interest:

$$\nabla \cdot \psi_{\alpha\beta}^{3-6} = \lambda_\alpha \nabla^2\lambda_\beta - \lambda_\beta \nabla^2\lambda_\alpha$$

When viewed as regular functions within each element, the Laplacians in the right hand side are zero because the barycentric coordinates are linear functions. However, these Laplacians are nonzero in the sense of distributions and contain Dirac-delta terms on the interelement boundaries due to the jumps of the normal component of the gradients of λ. Disregard of the distributional term has in the past been the source of two misconceptions about edge elements:

1. The basis set ψ^{3-6} presumably cannot be used to approximate fields with nonzero divergence. However, if this were true, linear elements, by similar considerations, could not be used to solve the Poisson equation with a nonzero right hand side because the Laplacian of the linear basis functions is zero *within each element*.
2. Since the basis functions have zero divergence, spurious modes are eliminated. While the conclusion is correct, the justification would only be valid if divergence were zero in the distributional sense. Furthermore, there are families of edge elements that are not divergence-free and yet do not produce spurious modes. Rigorous mathematical analysis of spectral convergence is quite involved (see footnote 43 on p. 140).

3.12.3 Implementation Issues

As already noted on p. 140, the finite element formulation of the cavity resonance problem (3.129) is obtained by restricting $\mathbf{E}$ and $\mathbf{E}'$ to a finite element subspace $W_h \subset H_0(\mathrm{curl}, \Omega)$

$$(\mu^{-1}\nabla \times \mathbf{E_h}, \nabla \times \mathbf{E_h}') \; - \; \omega^2(\epsilon \mathbf{E_h}, \mathbf{E_h}') \; = \; 0, \quad \forall \mathbf{E}' \in W_h \tag{3.133}$$

Subspace W_h can be spanned by either of the two basis sets introduced in the previous section for tetrahedral elements (one or two degrees of freedom per edge) or, alternatively, by higher order tetrahedral bases or bases on hexahedral elements (Section 3.12.4).

In the algorithmic implementation of the procedure, the role of the edges is analogous to the role of the nodes for nodal elements. In particular, the matrix sparsity structure is determined by the edge-to-edge adjacency: for any two edges that do not belong to the same element, the corresponding matrix entry is zero. An excellent source of information on adjacency structures and related algorithms (albeit not directly in connection with edge elements) is S. Pissanetzky's monograph [Pis84]. A new algorithmic issue, with no analogs in node elements, is the *orientation* of the edges, as the sign of field circulations depends on it. To make orientations consistent between several elements sharing the same edge, it is convenient to use global node numbers in the mesh. One suitable convention is to define the direction from the smaller global node number to the greater one as positive.

3.12.4 Historical Notes on Edge Elements

In 1980 and 1986, J.-C. Nédélec proposed two families of tetrahedral and hexahedral edge elements [Né80, Né86]. For tetrahedral elements, Nédélec's six- and twelve-dimensional approximation spaces are spanned by the vector basis functions $\lambda_\alpha \nabla \lambda_\beta - \lambda_\beta \nabla \lambda_\alpha$ and $\lambda_\alpha \nabla \lambda_\beta$, respectively, as discussed in the previous section.

Nédélec's exposition is formally mathematical and rooted heavily in the calculus of differential forms. As a result, there was for some time a disconnect between the outstanding mathematical development and its use in the engineering community.

To applied scientists and engineers, finite element analysis starts with the basis functions. This makes practical sense because one cannot actually *solve* an FE problem without specifying a basis. Many practitioners would be surprised to hear that a basis is *not* part of the standard mathematical definition of a finite element. In the mathematical literature, a finite element is defined, in addition to its geometric shape, by a (finite-dimensional) approximation space and a set of *degrees of freedom* – linear functionals over that approximation space (see e.g. the classical book by P.G. Ciarlet [Cia80]). Nodal values are the most typical such functionals, but there certainly are other possibilities as well. As we already know, in Nédélec's elements the linear functionals are circulations of the field along the edges. Nédélec built upon related ideas of P.-A. Raviart & J.M. Thomas who developed special finite elements on triangles in the late 1970s [RT77].

It took almost a decade to transform edge elements from a mathematical theory into a practical tool. A. Bossavit's contribution in that regard is exceptional. He presented, in a very lucid way, the fundamental rationale for edge elements [Bos88b, Bos88a] and developed their applications to eddy current problems [BV82, BV83], scattering [BM89], cavity resonances [Bos90], force computation [Bos92] and other areas. Stimulated by prior work of P.R. Kotiuga[48] and the mathematical papers of J. Dodziuk [Dod76], W. Müller [Mü78] and J. Komorowski [Kom75], Bossavit discovered a link between the tetrahedral edge elements with six degrees of freedom and differential forms in the 1957 theory of H. Whitney [Whi57].

Nédélec's original papers did not explicitly specify any bases for the FE spaces. Since practical computation does rely on the bases, the engineering and computational electromagnetics communities in the late 1980s and in the 1990s devoted much effort to more explicit characterization of edge element spaces. A detailed description of various types of elements would lead us too far astray, as this book is not a treatise on electromagnetic finite element analysis. However, to give the reader a flavor of some developments in this area, and to provide a reference point for the experts, succinct definitions of

[48] Kotiuga was apparently the first to note, in his 1985 Ph.D. thesis, the connection of finite element analysis in electromagnetics with the fundamental branches of mathematics: differential geometry and algebraic topology.

several common edge element spaces are compiled in Appendix 3.12.5 (see also [Tsu03]). Further information can be found in the monographs by P. Monk [Mon03], J. Jin [Jin02] and J.L. Volakis *et al.* [VCK98]. Comparative analysis of edge element spaces by symbolic algebra can be found in [Tsu03]. Families of *hierarchical* and adaptive elements developed independently by J.P. Webb [WF93, Web99, Web02] and by L. Vardapetyan & L. Demkowicz [VD99] deserve to be mentioned separately. In hierarchical refinement, increasingly accurate FE approximations are obtained by adding new functions *to the existing basis set*. This can be done both in the context of h-refinement (reducing the element size and adding functions supported by smaller elements to the existing functions on larger elements) and p-refinement (adding, say, quadratic functions to the existing linear ones). Hierarchical and adaptive refinement are further discussed in Section 3.13 for the scalar case. The vectorial case is much more complex, and I defer to the papers cited above for additional information. One more paper by Webb [Web93] gives a concise but very clear exposition of edge elements and their advantages.

3.12.5 Appendix: Several Common Families of Tetrahedral Edge Elements

Several representative families of elements, with the corresponding bases, are listed below. The list is definitely not exhaustive; for example, Demkowicz–Vardapetyan elements with hp-refinement and R. Hiptmair's general perspective on high order edge elements are not included.

As before, λ_i is the barycentric coordinate corresponding to node i ($i =$ 1,2,3,4) of a tetrahedral element.

1. The Ahagon–Kashimoto basis (20 functions) [AK95].
   $\{12$ "edge" functions $(4\lambda_i - 1)(\lambda_i \nabla\lambda_j - \lambda_j \nabla\lambda_i)$, $i \neq j\} \bigcup \{4\lambda_1(\lambda_2\nabla\lambda_3 - \lambda_3\nabla\lambda_2), 4\lambda_2(\lambda_3\nabla\lambda_1 - \lambda_1\nabla\lambda_3), 4\lambda_1(\lambda_3\nabla\lambda_4 - \lambda_4\nabla\lambda_3), 4\lambda_4(\lambda_1\nabla\lambda_3 - \lambda_3\nabla\lambda_1), 4\lambda_1(\lambda_2\nabla\lambda_4 - \lambda_4\nabla\lambda_2), 4\lambda_2(\lambda_1\nabla\lambda_4 - \lambda_4\nabla\lambda_1), 4\lambda_2(\lambda_3\nabla\lambda_4 - \lambda_4\nabla\lambda_3), 4\lambda_4(\lambda_2\nabla\lambda_3 - \lambda_3\nabla\lambda_2)\}$.
2. The Lee–Sun–Cendes basis (20 functions) [LSC91]. $\{12$ edge-based functions $\lambda_i\nabla\lambda_j$, $i \neq j\} \bigcup \{ \lambda_1\lambda_2\nabla\lambda_3, \lambda_1\lambda_3\nabla\lambda_2, \lambda_2\lambda_3\nabla\lambda_4, \lambda_2\lambda_4\nabla\lambda_3, \lambda_3\lambda_4\nabla\lambda_1, \lambda_3\lambda_1\nabla\lambda_4, \lambda_4\lambda_1\nabla\lambda_2, \lambda_4\lambda_2\nabla\lambda_1\}$.
3. The Kameari basis (24 functions) [Kam99]. $\{$the Lee basis$\} \bigcup \{ \nabla(\lambda_2\lambda_3\lambda_4), \nabla(\lambda_1\lambda_3\lambda_4), \nabla(\lambda_1\lambda_2\lambda_4), \nabla(\lambda_1\lambda_2\lambda_3) \}$.
4. The Ren–Ida basis (20 functions) [RI00]. $\{12$ edge-based functions $\lambda_i\nabla\lambda_j$, $i \neq j\} \bigcup \{ \lambda_1\lambda_2\nabla\lambda_3 - \lambda_2\lambda_3\nabla\lambda_1, \lambda_1\lambda_3\nabla\lambda_2 - \lambda_2\lambda_3\nabla\lambda_1, \lambda_1\lambda_2\nabla\lambda_4 - \lambda_4\lambda_2\nabla\lambda_1, \lambda_1\lambda_4\nabla\lambda_2 - \lambda_4\lambda_2\nabla\lambda_1, \lambda_1\lambda_3\nabla\lambda_4 - \lambda_4\lambda_3\nabla\lambda_1, \lambda_1\lambda_4\nabla\lambda_3 - \lambda_3\lambda_4\nabla\lambda_1, \lambda_2\lambda_3\nabla\lambda_4 - \lambda_4\lambda_3\nabla\lambda_2, \lambda_2\lambda_4\nabla\lambda_3 - \lambda_3\lambda_2\nabla\lambda_4\}$.
5. The Savage–Peterson basis [SP96]. $\{12$ edge-based functions $\lambda_i\nabla\lambda_j, i \neq j\} \bigcup \{ \lambda_i\lambda_j\nabla\lambda_k - \lambda_i\lambda_k\nabla\lambda_j, \lambda_i\lambda_j\nabla\lambda_k - \lambda_j\lambda_k\nabla\lambda_i, 1 \leq i < j < k \leq 4\}$.
6. The Yioultsis–Tsiboukis basis (20 functions) [YT97]. $\{(8\lambda_i 2 - 4\lambda_i)\nabla\lambda_j + (-8\lambda_i\lambda_j + 2\lambda_j)\nabla\lambda_i, i \neq j\} \bigcup \{16\lambda_1\lambda_2\nabla\lambda_3 - 8\lambda_2\lambda_3\nabla\lambda_1 - 8\lambda_3\lambda_1\nabla\lambda_2;$

$16\lambda_1\lambda_3\nabla\lambda_2 - 8\lambda_3\lambda_2\nabla\lambda_1 - 8\lambda_2\lambda_1\nabla\lambda_3$; $16\lambda_4\lambda_1\nabla\lambda_2 - 8\lambda_1\lambda_2\nabla\lambda_4 - 8\lambda_2\lambda_4\nabla\lambda_1$; $16\lambda_4\lambda_2\nabla\lambda_1 - 8\lambda_2\lambda_1\nabla\lambda_4 - 8\lambda_1\lambda_4\nabla\lambda_2$; $16\lambda_2\lambda_3\nabla\lambda_4 - 8\lambda_3\lambda_4\nabla\lambda_2 - 8\lambda_4\lambda_2\nabla\lambda_3$; $16\lambda_2\lambda_4\nabla\lambda_3 - 8\lambda_4\lambda_3\nabla\lambda_2 - 8\lambda_3\lambda_2\nabla\lambda_4$; $16\lambda_3\lambda_1\nabla\lambda_4 - 8\lambda_1\lambda_4\nabla\lambda_3 - 8\lambda_4\lambda_3\nabla\lambda_1$; $16\lambda_3\lambda_4\nabla\lambda_1 - 8\lambda_4\lambda_1\nabla\lambda_3 - 8\lambda_1\lambda_3\nabla\lambda_4\}$.

7. The Webb–Forghani basis (20 functions) [WF93]. $\{$6 edge-based functions $\lambda_i\nabla\lambda_j - \lambda_j\nabla\lambda_i$, $i \neq j\} \bigcup \{$6 edge-based functions $\nabla(\lambda_i\lambda_j)$, $i \neq j\} \bigcup \{ \lambda_1\lambda_2\nabla\lambda_3$, $\lambda_1\lambda_3\nabla\lambda_2$, $\lambda_2\lambda_3\nabla\lambda_4$, $\lambda_2\lambda_4\nabla\lambda_3$, $\lambda_3\lambda_4\nabla\lambda_1$, $\lambda_3\lambda_1\nabla\lambda_4$, $\lambda_4\lambda_1\nabla\lambda_2$, $\lambda_4\lambda_2\nabla\lambda_1\}$.
8. The Graglia–Wilton–Peterson basis (20 functions) [GWP97]. $\{ (3\lambda_i - 1)(\lambda_i\nabla\lambda_j - \lambda_j\nabla\lambda_i)$, $i \neq j\} \bigcup 9/2 \times \{\lambda_2(\lambda_3\nabla\lambda_4 - \lambda_4\nabla\lambda_3)$, $\lambda_3(\lambda_4\nabla\lambda_2 - \lambda_2\nabla\lambda_4)$, $\lambda_3(\lambda_4\nabla\lambda_1 - \lambda_1\nabla\lambda_4)$, $\lambda_4(\lambda_1\nabla\lambda_3 - \lambda_3\nabla\lambda_1)$, $\lambda_4(\lambda_1\nabla\lambda_2 - \lambda_2\nabla\lambda_1)$, $\lambda_1(\lambda_4\nabla\lambda_2 - \lambda_2\nabla\lambda_4)$, $\lambda_1(\lambda_2\nabla\lambda_3 - \lambda_3\nabla\lambda_2)$, $\lambda_2(\lambda_1\nabla\lambda_3 - \lambda_3\nabla\lambda_1)\}$.

3.13 Adaptive Mesh Refinement and Multigrid Methods

3.13.1 Introduction

One of the most powerful ideas that has shaped the development of Finite Element Analysis since the 1980s is adaptive refinement. Once an FE problem has been solved on a given initial mesh, special *a posteriori* error estimates or indicators[49] are used to identify the subregions with relatively high error. The mesh is then refined in these areas, and the problem is re-solved. It is also possible to "unrefine" the mesh in the regions where the error is perceived to be small. The procedure is then repeated recursively and is typically integrated with efficient system solvers such as multigrid cycles or multilevel preconditioners (Section 3.13.4).

There are two main versions of mesh refinement. In h-refinement, the mesh size h is reduced in selected regions to improve the accuracy. In p-refinement, the element-wise order p of local approximating polynomials is increased. The two versions can be combined in an hp-refinement procedure. There are numerous ways of error estimation (Section 3.13.3 on p. 151) and numerous algorithms for effecting the refinement.

To summarize, adaptive techniques are aimed at generating a quasi-optimal mesh adjusted to the local behavior of the solution, while maintaining a high convergence rate of the iterative solver. Three different but related issues arise:

1. Implementation of local refinement without violating the geometric conformity of the mesh.
2. Efficient multilevel iterative solvers.
3. Local *a posteriori* error estimates.

[49] *Estimates* provide an approximate numerical value of the actual error. *Indicators* show whether the error is relatively high or low, without necessarily predicting its numerical value.

Fig. 3.35 shows nonconforming ("slave") nodes appearing on a common boundary between two finite elements e_1 and e_2 if one of these elements (say, e_1) is refined and the other one (e_2) is not. The presence of such nodes is a deviation from the standard set of requirements on a FE mesh. If no restrictions are imposed, the continuity of the solution at slave nodes will generally be violated. One remedy is a transitory (so-called "green") refinement of element e_2 (W.F. Mitchell [Mit89, Mit92], F. Bornemann *et al.* [BEK93]) as shown in Fig. 3.35, right. However, green refinement generally results in non-nested meshes, which may affect the performance of iterative solvers.

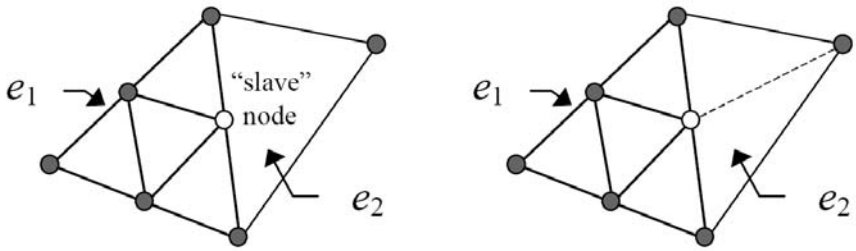


Fig. 3.35. Local mesh refinement (2D illustration for simplicity). Left: continuity of the solution at "slave" nodes must be maintained. Right: "green refinement". (Reprinted by permission from [TP99a] ©1999 IEEE.)

3.13.2 Hierarchical Bases and Local Refinement

Alternatively, nonconforming nodes may be retained if proper continuity conditions are imposed. This can be accomplished in a natural way in the hierarchical basis (H. Yserentant [Yse86], W.F. Mitchell [Mit89, Mit92], U. Rüde [RÖ93]). A simple 1D example (Fig. 3.36) illustrates the hierarchical basis representation of a function.

In the nodal basis a piecewise-linear function has a vector of nodal values $u^{(N)} = (u_1, u_2, u_3, u_4, u_5, u_6)^T$. Nodes 5 and 6 are generated by refining the coarse level elements 1-2 and 2-3. In the hierarchical basis, the degrees of freedom at nodes 5, 6 correspond to the *difference* between the values on the fine level and the interpolated value from the coarse level. Thus the vector in the hierarchical basis is

$$u^{(H)} = (u_1, u_2, u_3, u_4, u_5 - \frac{1}{2}(u_1 + u_2), u_6 - \frac{1}{2}(u_2 + u_3))^T \qquad (3.134)$$

This formula effects the transformation from nodal to hierarchical values of the same piecewise-linear function.

More generally, let a few levels of nested FE meshes (in one, two or three dimensions) be generated by recursively subdividing some or all elements on

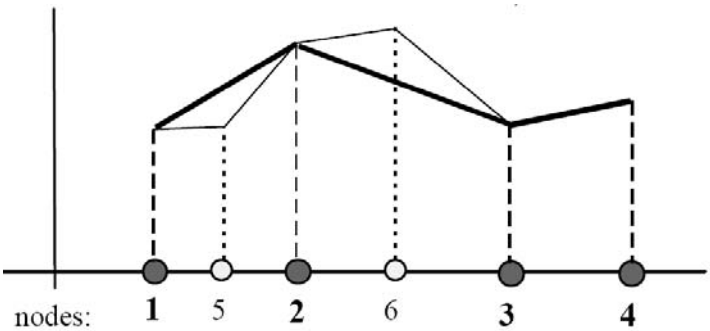


Fig. 3.36. A fragment of a two-level 1D mesh. (Reprinted by permission from [TP99a] ©1999 IEEE.)

a coarser level into several smaller elements. For simplicity, only first order nodal elements will be considered and it will be assumed that new nodes are added at the midpoints of the existing element edges. (The ideas are quite general, however, and can be carried over to high order elements and edge elements; see e.g. P.T.S. Liu & J.P. Webb [LW95], J.P. Webb & B. Forghani [WF93].)

The hierarchical representation of a piecewise-linear function can be obtained from its nodal representation by a recursive application of elementary transforms similar to (3.134). Precise theory and implementation are detailed by H. Yserentant [Yse86].

An advantage of the hierarchical basis is the natural treatment of slave nodes (Fig. 3.35, left). The continuity of the solution is ensured by simply setting the hierarchical basis value at these nodes to zero.

Remark 6. In the nonconforming refinement of Fig. 3.35 (left), element shapes do not deteriorate. However, this advantage is illusory. Indeed, the FE space for the "green refinement" of Fig. 3.35 (right) obviously contains the FE space of Fig. 3.35 (left), and therefore the FE solution with slave nodes cannot be more accurate than for green refinement. Thus the effective "mesh quality," unfortunately, is not preserved with slave nodes.

For tetrahedral meshes, subdividing an element into smaller ones when the mesh is refined is not trivial; careless subdivision may lead to degenerate elements. S.Y. Zhang [Zha95] proposed two schemes: "labeled edge subdivision" and "short-edge subdivision" guaranteeing that tetrahedral elements do not degenerate in the refinement process. The initial stage of both methods is the same: the edge midpoints of the tetrahedron are connected, producing four corner tetrahedra and a central octahedron. The octahedron can be further subdivided into four tetrahedra in three different ways [Zha95] by using one additional edge. The difference between Zhang's two refinement schemes is in the way this additional edge is chosen. The "labeled edge subdivision"

algorithm relies on a numbering convention for nodes being generated (see [Zha95] for details). In the "short edge subdivision" algorithm the shortest of the three possible interior edges is selected. For tetrahedra without obtuse planar angles between edges both refinement schemes are equivalent, provided that the initial refinement is the same – i.e. for a certain numbering of nodes of the initial element [Zha95].

Zhang points out that "in general, it is not simple to find the measure of degeneracy for a given tetrahedron" [Zha95] and uses as such a measure the ratio of the maximum edge length to the radius of the inscribed sphere. A. Plaks and I used a more precise criterion – the *minimum singular value condition* (Section 3.14) to compare the two refinement schemes. Short-edge subdivision in general proves to be better than labeled edge subdivision [TP99b].

3.13.3 *A Posteriori* Error Estimates

Adaptive *hp*-refinement requires some information about the distribution of numerical errors in the computational domain. The FE mesh is refined in the regions where the error is perceived to be higher and left unchanged, or even unrefined, in regions with lower errors. Numerous approaches have been developed for estimating the errors *a posteriori* – i.e. after the FE solution has been found. Some of these approaches are briefly reviewed below; for comprehensive treatment, see monographs by M. Ainsworth & J.T. Oden [AO00], I. Babuška & T. Strouboulis [BS01], R. Verfürth [Ver96], and W. Bangerth & R. Rannacher [BR03].

Much information and many references for this section were provided by S. Prudhomme, the reviewer of this book; his help is greatly appreciated. The overview below follows the book chapter by Prudhomme & Oden [PO02] as well as W.F. Mitchell's paper [Mit89].

Recovery-based error estimators

These methods were proposed by O.C. Zienkiewicz & J.Z. Zhu; as of May 2007, their 1987 and 1992 papers [ZZ87, ZZ92a, ZZ92b] were cited 768, 531 and 268 times, respectively. The essence of the method, in a nutshell, is in field averaging. The computed field within an element is compared with the value obtained by double interpolation: element-to-node first and then node-to-element. The intuitive observation behind this idea is that the field typically has jumps across element boundaries; these jumps are a numerical artifact that can serve as an error indicator. The averaging procedure captures the magnitudes of the jumps. Some versions of the Zienkiewicz–Zhu method rely on superconvergence properties of the FE solution at special points in the elements.

For numerical examples and validation of gradient-recovery estimators, see e.g. I. Babuška *et al.* [BSU$^+$94]. The method is easy to implement and in my experience (albeit limited mostly to magnetostatic problems) works well

[TP99a].[50] One difficulty is in handling nodes at material interfaces, where the field jump can be a valid physical property rather than a numerical artifact. In our implementation [TP99a] of the Zienkiewicz–Zhu scheme, the field values were averaged at the interface nodes separately for each of the materials involved.

Ainsworth & Oden [AO00] note some drawbacks of recovery-based estimators and even present a 1D example where the recovery-based error estimate is zero, while the actual error can be arbitrarily large. Specifically, they consider a 1D Poisson equation with a rapidly oscillating sinusoidal solution. It can be shown (see Appendix 3.10, p. 127) that the FE-Galerkin solution with first-order elements actually *interpolates* the exact solution at the FE mesh nodes. Hence, if these nodes happen to be located at the zeros of the oscillating exact solution, the FE solution, as well as all the gradients derived from it, are identically zero!

Prudhomme & Oden also point out that for problems with shock waves gradient recovery methods tend to indicate mesh refinement around the shock rather than at the shock itself.

Residual-based methods

While the solution error is not directly available, *residual* – the difference between the right and left hand sides of the equation – is. For a problem of the form

$$\mathcal{L}u = \rho \tag{3.135}$$

and the corresponding weak formulation

$$\mathcal{L}(u, v) = (\rho, v) \tag{3.136}$$

the residual is

$$\mathcal{R}u_h \equiv \rho - \mathcal{L}u_h \tag{3.137}$$

or in the weak form

$$\mathcal{R}(u_h, v) \equiv (\rho, v) - \mathcal{L}(u_h, v) \tag{3.138}$$

Symbols $\mathcal{L}$ and $\mathcal{R}$ here are overloaded (with little possibility of confusion) as operators and the corresponding bilinear forms.

The numerical solution u_h satisfies the Galerkin equation in the finite-dimensional subspace V_h. In the full space V residuals (3.137) or (3.138) are, in general, nonzero and can serve as a measure of accuracy. In principle, the error, and hence the exact solution, can be found by solving the problem with the residual in the right hand side. However, doing so is no less difficult than solving the original problem in the first place. Instead, one looks for

[50] Joint work with A. Plaks.

computationally inexpensive ways of extracting useful information about the magnitude of the error from the magnitude of the residual.

One of the simplest element-wise error estimators of this kind combines, with proper weights, two residual-related terms: $(\mathcal{L}u - \rho)^2$ integrated over the volume (area) of the element and the jump of the normal component of flux density, squared and integrated over the facets of the element (R.E. Bank & A.H. Sherman [BS79]). P. Morin *et al.* [MNS02] develop convergence theory for adaptive methods with this estimator and emphasize the importance of the volume-residual term that characterizes possible oscillations of the solution.

A different type of method, proposed by I. Babuška & W.C. Rheinboldt in the late 1970s, makes use of auxiliary problems over small clusters ("patches") of adjacent elements [BR78b, BR78a, BR79]. To gain any additional nontrivial information about the error, the auxiliary local problem must be solved with higher accuracy than the original global problem, i.e. the FE space has to be locally enriched (usually using h- or p-refinement). An alternative interpretation (W.F. Mitchell [Mit89]) is that such an estimator measures how strongly the FE solution would change if the mesh were to be refined locally.

Yet another possibility is to solve the problem with the residual globally but approximately, using only a few iterations of the conjugate gradient method (Prudhomme & Oden [PO02]).

Goal-oriented error estimation

In practice, FE solution is often aimed at finding specific quantities of interest – for example, field, temperature, stress, etc. at a certain point (or points), equivalent parameters (e.g. capacitance or resistance between electrodes), and so on. Naturally, the effort should then be concentrated on obtaining these quantities of interest, rather than the overall solution, with maximum accuracy.

Pointwise estimates have a long history dating back at least to the the 1940s–1950s (H.J. Greenberg [Gre48], C.B. Maple, [Map50]; K. Washizu [Was53]). The key idea can be briefly summarized as follows. One can express the value of solution u at a point r_0 using the Dirac delta functional as

$$u(r_0) \;=\; \langle u,\, \delta(r - r_0)\rangle \tag{3.139}$$

(Appendix 6.15 on p. 343 gives an introduction to generalized functions (distributions), with the Dirac delta among them.) Further progress can be made by using Green's function g of the $\mathcal{L}$ operator:[51] $\mathcal{L}g(r, r_0) = \delta(r - r_0)$. Then

$$u(r_0) \;=\; (u,\, \mathcal{L}g(r, r_0)) \;=\; (\mathcal{L}^* u,\, g(r, r_0)) \;=\; \mathcal{L}^*(u,\, g(r, r_0)) \tag{3.140}$$

where symbol $\mathcal{L}^*$ is the adjoint operator and (again with overloading) the corresponding bilinear form $\mathcal{L}^*(u, v) \equiv \mathcal{L}(v, u)$. The role of Green's function in

[51] The functional space where this operator is defined, and hence the boundary conditions, remain fixed in the analysis.

this analysis is to convert the delta functional (3.139) that is hard to evaluate directly into an $\mathcal{L}$-form that is closely associated with the problem at hand.

The right hand side of (3.140) typically has the physical meaning of the mutual energy of two fields. For example, if $\mathcal{L}$ is the Laplace operator (self-adjoint if the boundary conditions are homogeneous), then the right hand side is $(\nabla u, \nabla g)$ – the inner product (mutual energy) of fields $-\nabla u$ (the solution) and $-\nabla g$ (field of a point source). Importantly, due to the variational nature of the problem, lower and upper bounds can be established for $u(r_0)$ of (3.140) (A.M. Arthurs [Art80]). Moreover, bounds can be established for the *pointwise error* as well. In the finite element context (1D), this was done in 1984 by E.C. Gartland [EG84]. Also in 1984, in a series of papers [BM84a, BM84b, BM84c], I. Babuška & A.D. Miller applied the duality ideas to *a posteriori* error estimates and generalized the method to quantities of physical interest. In Babuška & Miller's example of an elasticity problem of beam deformation, such quantities include the average displacement of the beam, the shear force, the bending moment, etc.

For a contemporary review of the subject, including both the duality techniques and goal-oriented estimates with adaptive procedures, see R. Becker & R. Rannacher [BR01] and J.T. Oden & S. Prudhomme [OP01]. For electromagnetic applications, methods of this kind were developed by R. Albanese, R. Fresa & G. Rubinacci [AF98, AFR00], by J.P. Webb [Web05] and by P. Ingelstrom & A. Bondeson [IB].

Fully Adaptive Multigrid

In this approach, developed by W.F. Mitchell [Mit89, Mit92] and U. Rüde [R̈93]), solution values in the hierarchical basis (Section 3.13.2, p. 149) characterize the difference between numerical solutions at two subsequent levels of refinement and can therefore serve as error estimators.

3.13.4 Multigrid Algorithms

The presentation of multigrid methods in this book faces a dilemma. These methods are first and foremost iterative system solvers – the subject matter not in general covered in the book. On the other hand, multigrid methods, in conjunction with adaptive mesh refinement, have become a truly state-of-the-art technique in modern FE analysis and an integral part of commercial FE packages; therefore the chapter would be incomplete without mentioning this subject.

Fortunately, several excellent books exist, the most readable of them being the one by W.L. Briggs *et al.* [BHM00], with a clear explanation of key ideas and elements of the theory. For a comprehensive exposition of the mathematical theory, the monographs by W. Hackbusch [Hac85], S.F. McCormick

[McC89], P. Wesseling [Wes91] and J.H. Bramble [Bra93], as well as the seminal paper by A. Brandt [Bra77], are highly recommended; see also the review paper by C.C. Douglas [Dou96].

On a historical note, the original development of multilevel algorithms is attributed to the work of the Russian mathematicians R.P. Fedorenko [Fed61, Fed64] and N.S. Bakhvalov [Bak66] in the early 1960s. There was an explosion of activity after A. Brandt further developed the ideas and put them into practice [Bra77].

As a guide for the reader unfamiliar with the essence of multigrid methods, this section gives a narrative description of the key ideas, with "hand-waving" arguments only.

Consider the simplest possible model 1D equation

$$\mathcal{L}u \;\equiv\; -\frac{d^2u}{dx^2} \;=\; f \quad \text{on } \Omega = [0,a]; \qquad u(0) = u(a) = 0 \tag{3.141}$$

where f is a given function of x. FE-Galerkin discretization of this problem leads to a system of equations

$$L\underline{u} \;=\; \underline{f} \tag{3.142}$$

where $\underline{u}$ and $\underline{f}$ are Euclidean vectors and L is a square matrix; $\underline{u}$ represents the nodal values of the FE solution. For first order elements, matrix L is three-diagonal, with 2 on the main diagonal and -1 on the adjacent ones. (The modification of the matrix due to boundary conditions, as described in Section 3.7.1, will not be critical in this general overview.)

Operator $\mathcal{L}$ has a discrete set of spatial eigenfrequencies and eigenmodes, akin to the modes of a guitar string. As Fig. 3.37 illustrates, the *discrete* operator L of (3.142) inherits the oscillating behavior of the eigenmodes but has only a finite number of those. There is the Nyquist limit for the highest spatial frequency that can be adequately represented on a grid of size h. Fig. 3.37 exhibits the eigenmodes with lowest and highest frequency on a uniform grid with 16 elements.

Any iterative solution process for equation (3.142) – including multigrid solvers – involves an approximation $\underline{v}$ to the exact solution vector $\underline{u}$. The error vector

$$\underline{e} \;\equiv\; \underline{u} \;-\; \underline{v} \tag{3.143}$$

is of course generally unknown in practice; however, the *residual* $\underline{r} \;=\; \underline{f} - L\underline{v}$ *is* computable. It is easy to see that the residual is equal to $L\underline{e}$:

$$\underline{r} \;=\; \underline{f} - L\underline{v} \;=\; L\underline{u} - L\underline{v} \;=\; L\underline{e} \tag{3.144}$$

The following sequence of observations leads to the multigrid methodology.

1. *High-frequency* components of the error – or, equivalently, of the residual – (similar to the bottom part of Fig. 3.37) can be easily and rapidly reduced

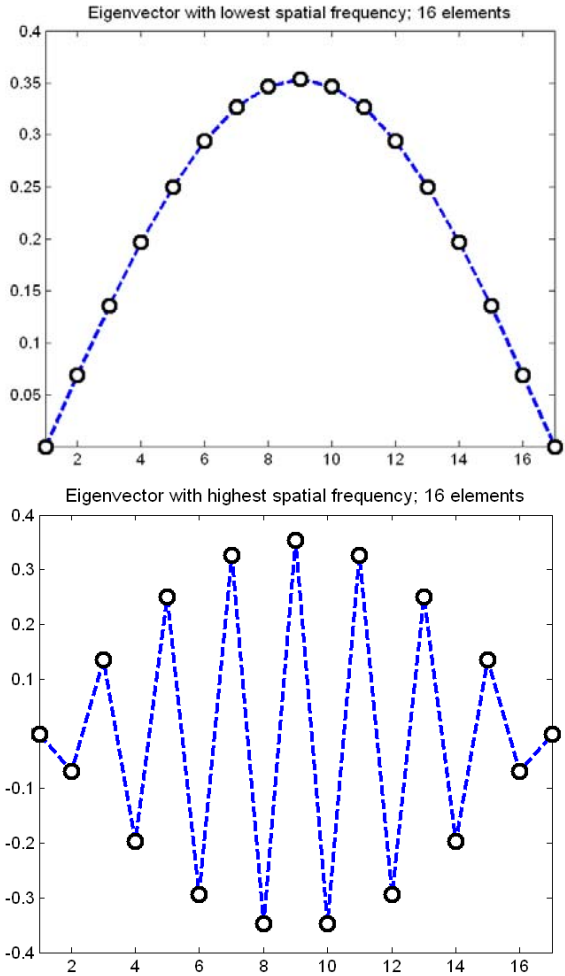


Fig. 3.37. Eigenvectors with lowest (top) and highest (bottom) spatial frequency. Laplace operator discretized on a uniform grid with 16 elements.

by applying basic iterative algorithms such as *Jacobi* or *Gauss–Seidel*. In contrast, low-frequency components of the error decay very slowly. See [BHM00, Tre97, GL96] for details.

2. Once highly oscillatory components of the error have been reduced and the error and the residual have thus become sufficiently smooth, the problem can be effectively transferred to a coarser grid (typically, twice coarser). The procedure for information transfer between the grids is outlined below. The spatial frequency of the eigenmodes *relative to the coarser grid* is higher than on the finer grid, and the components of the error that are

oscillatory relative to the coarse grid can be again eliminated with basic iterative solvers. This is effective not only because the relative frequency is higher, but also because the system size on the coarser grid is smaller.

3. It remains to see how the information transfer between finer and coarser grids is realized. *Residuals* are transferred from finer to coarser grids. Correction vectors obtained after smoothing iterations on coarser grids are transferred to finer grids. There is more than one way of defining the transfer operators. Vectors from a coarse grid can be moved to a fine one by some form of interpolation of the nodal values. The simplest fine-to-coarse transfer is *injection*: the values at the nodes of the coarse grids are taken to be the same as the values at the corresponding nodes of the fine grid.
   However, it is often desirable that the coarse-to-fine and fine-to-coarse transfer operators be adjoint to one another,[52] especially for symmetric problems, to preserve the symmetry. In that case the fine-to-coarse transfer is different from injection.

Multigrid utilizes these ideas recursively, on a sequence of nested grids. There are several ways of navigating these grids. *V-cycle* starts on the finest grid and descends gradually to the coarsest one; then moves back to the finest level. *W-cycle* also starts by traversing all fine-to-coarse levels; then, using the coarsest level as a base, it goes back-and forth in rounds spanning an increasing number of levels. Finally, *full multigrid* cycle starts at the coarsest level and moves back-and-forth, involving progressively more and more finer levels. A precise description and pictorial illustrations of these algorithms can be found in any of the multigrid books.

Convergence of multigrid methods depends on the nature of the underlying problem: primarily, in mathematical terms, on whether or not the problem is elliptic and on the level of regularity of the solution, on the particular type of the multigrid algorithm employed, and to a lesser extent on other details (the norms in which the error is measured, smoothing algorithms, etc.) For elliptic problems, convergence can be close to optimal – i.e. proportional to the size of the problem, possibly with a mild logarithmic factor that in practice is not very critical.

Furthermore, multigrid methods can be used as *preconditioners* in conjugate gradient and similar solvers; particularly powerful are the Bramble–Pasciak–Xu (BPX) preconditioners developed in J. Xu's Ph.D. thesis [Xu89] and in [BPX90]. Since BPX preconditioners are expressed as double sums over all basis functions and over all levels, they are relatively easy to parallelize. A broad mathematical framework for multilevel preconditioners and for the analysis of convergence of multigrid methods in general is established

[52] There is an interesting parallel with Ewald methods of Chapter 5, where charge-to-grid and grid-to-charge interpolation operators must be adjoint for conservation of momentum in a system of charged particles to hold numerically; see p. 262.

in Xu's papers [Xu92, Xu97]. Results of numerical experiments with BPX for several electromagnetic applications are reported by A. Plaks and myself in [Tsu94, TPB98, PTPT00].

Another very interesting development is *algebraic* multigrid (AMG) schemes, where multigrid ideas are applied in an abstract form (K. Stüben *et al.* [Stü83, SL86, Stü00]). The underlying problem may or may not involve any actual geometric grids; for example, there are applications to electric circuits and to coupled field-circuit problems (D. Lahaye *et al.* [LVH04]). In AMG, a hierarchical structure typical of multigrid methods is created artificially, by examining the strength of the coupling between the unknowns. The main advantage of AMG is that it can be used as a "black box" solver. For further information, the interested reader is referred to the books cited above and to the tutorials posted on the MGNet website.[53]

3.14 Special Topic: Element Shape and Approximation Accuracy

The material of this section was inspired by my extensive discussions with Alain Bossavit and Pierre Asselin in 1996–1999. (By extending the analysis of J.L. Synge [Syn57], Asselin independently obtained a result similar to the minimum singular value condition on p. 170.) Numerical experiments were performed jointly with Alexander Plaks. I also thank Ivo Babuška and Randolph Bank for informative conversations in 1998–2000.

3.14.1 Introduction

Common sense, backed up by rigorous error estimates (Section 3.10, p. 125) tells us that the accuracy of the finite element approximation depends on the element size and on the order of polynomial interpolation. More subtle is the dependence of the error on element *shape*. Anyone who has ever used FEM knows that a triangular element similar to the one depicted on the left side of Fig. 3.38 is "good" for approximation, while the element shown on the right is "bad". The flatness of the second element should presumably lead to poor accuracy of the numerical solution.

But how flat are flat elements? How can element shape in FEM be characterized precisely and how can the "source" of the approximation error be identified? Some of the answers to these questions are classical but some are not yet well known, particularly the connection between approximation accuracy and FE matrices (Section 3.14.2), as well as the minimum singular value criterion for the "edge shape matrix" (Sections 3.14.2 and 3.14.3).

The reader need not be an expert in FE analysis to understand the first part of this section; the second part is more advanced. Overall, the section is

[53] http://www.mgnet.org/mgnet-tuts.html

based on my papers [Tsu98b, Tsu98a, Tsu98c, TP98, TP99b, Tsu99] (joint work with A. Plaks).

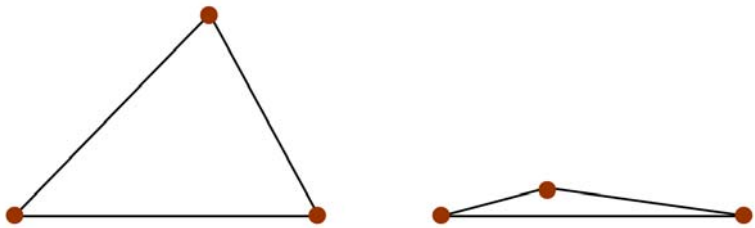


Fig. 3.38. "Good" and "bad" element shape (details in the text).

For triangular elements, one intuitively obvious consideration is that small angles should be avoided. The mathematical basis for that is given by Zlámal's minimum angle condition [Zl8]: if the minimum angle of elements is bounded away from zero, $\phi_{\min} \geq \phi_0 > 0$, then the FE interpolation error tends to zero for the family of meshes with decreasing mesh sizes. Geometrically equivalent to Zlámal's condition is the boundedness of the ratio of the element diameter (maximum element edge $l_{\max}$) to the radius ρ of the inscribed circle.

Zlámal's condition implies that small angles should be avoided. But must they? In mathematical terms, one may wonder if Zlámal's condition is not only sufficient but in some sense necessary for accurate approximation.

If Zlámal's condition were necessary, a right triangle with a small acute angle would be unsuitable. However, on a regular mesh with right triangles, first-order FE discretization of the Laplace equation is easily shown to be identical with the standard 5-point finite difference scheme. But the FD scheme does not have any shape related approximation problems. (The accuracy is limited by the maximum mesh size but not by the aspect ratio.) This observation suggests that Zlámal's condition could be too stringent.

Indeed, a less restrictive shape condition for triangular elements exists. It is sufficient to require that the *maximum* angle of an element be bounded away from π. In particular, according to this condition, right triangles, even with very small acute angles, are acceptable (what matters is the maximum angle that remains equal to $\pi/2$). The maximum angle condition appeared in J.L. Synge's monograph [Syn57] (pp. 209–213) in 1957, before the finite element era. (Synge considered piecewise-linear interpolation on triangles without calling them finite elements.) In 1976, I. Babuška & A.K. Aziz [BA76] published a more detailed analysis of FE interpolation on triangles and showed that the maximum angle condition was not only sufficient, but in a sense essential for the convergence of FEM. In addition, they proved the corresponding W_p^1-norm estimate. In 1992, M. Křižek [Kŷ2] generalized the maximum angle condition to tetrahedral elements: the maximum angle for all triangular faces

and the maximum dihedral angle should be bounded away from π. Other estimates for tetrahedra (and, more generally, simplices in $\mathbb{R}^d$) were given by Yu.N. Subbotin [Sub90] and S. Waldron [Wal98a]. P. Jamet's condition [Jam76] is closest to the result of this section but is more difficult to formulate and apply.

On a more general theoretical level, the study of piecewise-polynomial interpolation in Sobolev spaces, with applications to spline interpolation and FEM, has a long history dating back to the fundamental works of J. Deny & J.L. Lions, J.H. Bramble & S.R. Hilbert [BH70], I. Babuška [Bab71], and the already cited Ciarlet & Raviart paper.

Two general approaches systematically developed by Ciarlet & Raviart have now become classical. The first one is based on the multipoint Taylor formula (P.G. Ciarlet & C. Wagschal [CW71]); the second approach (e.g. Ciarlet [Cia80]) relies on the Deny-Lions and Bramble–Hilbert lemmas. In both cases, under remarkably few assumptions, error estimates for Lagrange and Hermite interpolation on a set of points in $\mathbb{R}^n$ are obtained.

For tetrahedra, the "shape part" of Ciarlet & Raviart's powerful result (p. 125) translates into the ratio of the element diameter (i.e. the maximum edge) to the radius of the inscribed sphere. Boundedness of this ratio ensures convergence of FE interpolation on a family of tetrahedral meshes with decreasing mesh sizes. However, as in the 2D case, such a condition is a little too restrictive. For example, "right tetrahedra" (having three mutually orthogonal edges) are rejected, even though it is intuitively felt, by analogy with right triangles, that there is in fact nothing wrong with them.

A precise characterization of the shape of tetrahedral elements is one of the particular results of the general analysis that follows. An *algebraic*, rather than geometric, source of interpolation errors for arbitrary finite elements is identified and its geometric interpretation for triangular and tetrahedral elements is given.

3.14.2 Algebraic Sources of Shape-Dependent Errors: Eigenvalue and Singular Value Conditions

First, we establish a direct connection between interpolation errors and the maximum eigenvalue (or the trace) of the appropriate FE stiffness matrices. This is different from the more standard consideration of matrices of the affine transformation to/from a reference element (as done e.g. by N. Al Shenk [She94]).

As shown below, the maximum eigenvalue of the stiffness matrix has a simple geometric meaning for first and higher order triangles and tetrahedra. Even without a geometric interpretation, the eigenvalue/trace condition is useful in practical FE computation, as the matrix trace is available at virtually no additional cost. Moreover, the stiffness matrix automatically reflects the chosen energy norm, possibly for inhomogeneous and/or anisotropic parameters.

For the energy-seminorm approximation on first order tetrahedral nodal elements, or equivalently, for L_2-approximation of conservative fields on tetrahedral edge elements (Section 3.12), the maximum eigenvalue analysis leads to a new criterion in terms of the minimum singular value of the "edge shape matrix". The columns of this matrix are the Cartesian representations of the unit edge vectors of the tetrahedron.

The new singular value estimate has a clear algebraic and geometric meaning and proves to be not only sufficient, but in some strong sense necessary for the convergence of FE interpolation on a sequence of meshes. The minimum singular value criterion is a direct generalization of the Synge–Babuška–Aziz maximum angle condition to three (and more) dimensions.

Even though the approach presented here is general, let us start with first order triangular elements to fix ideas. Let $\Omega \subset \mathbb{R}^2$ be a convex polygonal domain. Following the standard definition, we shall call a set M of triangular elements K_i, $M = \{K_1, K_2, \ldots, K_n\}$, a *triangulation* of the domain if
(a) $\bigcup_{i=1}^n K_i = \Omega$;
(b) any two triangles either have no common points, or have exactly one common node, or exactly one common edge.

Let $h_i = \operatorname{diam} K_i$; then the *mesh size* h is the maximum of h_i for all elements in M (i.e. the maximum edge length of all triangles). Let $\mathcal{N}$ be the geometric set of nodes $\{r_i\}$ ($i = 1, 2, \ldots, n$, $r_i \in \bar{\Omega}$) of all triangles in M, and let $P^1(M)$ be the space of functions that are continuous in Ω and linear within each of the triangular elements K_i.[54] Let $P^1(K_i)$ be the restriction of $P^1(M)$ to a specific element K_i. Thus $P^1(K_i)$ is just the (three-dimensional) space of linear functions over the element.

Considering interpolation of functions in $C^2(\bar{\Omega})$ for simplicity, one can define the interpolation operator $\Pi : C^2(\bar{\Omega}) \to P^1(M)$ by

$$(\Pi u)(r_i) \;=\; u(r_i), \quad \forall r_i \in \mathcal{N}, \;\; \forall u \in C^2(\bar{\Omega}) \tag{3.145}$$

We are interested in evaluating the interpolation error $\Pi u - u$ in the energy norm $\|\cdot\|_E$ induced by an inner product $(\cdot\,,\,\cdot)_E$ ("E" for "energy," not to be confused with Euclidean spaces).[55]

Remark 7. In FE applications, u is normally the solution of a certain boundary value problem in Ω. The error bounds for interpolation and for the Galerkin or Ritz projection are closely related (e.g. by Céa's lemma or the LBB condition, Section 3.5). Although this provides an important motivation to study interpolation errors, here u need not be associated with any boundary value problem.

[54] Elsewhere in the book, symbol $\mathcal{N}$ denotes the nodal values of a function. The usage of this symbol for the set of nodes is limited to this section only and should not cause confusion.

[55] The analysis is also applicable to *semi*norms instead of norms if the definition of energy inner product is relaxed to allow $(u, u)_E = 0$ for a nonzero u.

Consider a representative example where the inner product and the energy seminorm in $C^2(\bar{\Omega})$ are introduced as

$$(u,v)_{E,\Omega} \;=\; \int_\Omega \nabla u \cdot \nabla v \, d\Omega \tag{3.146}$$

$$|u|_E \;=\; (u,u)_E^{\frac{1}{2}} \tag{3.147}$$

(If Dirichlet boundary conditions on a nontrivial part of the boundary are incorporated in the definition of the functional space, the seminorm is in fact a norm.)

The element stiffness matrix $A(K_i)$ for a given basis $\{\psi_1, \psi_2, \psi_3\}$ of $P^1(K_i)$ corresponds to the energy inner product (3.146) viewed as a bilinear form on $P^1(K_i) \times P^1(K_i)$:

$$(u,v)_{E,K_i} \;=\; (A(K_i)\,\underline{u}(K_i),\; \underline{v}(K_i)), \quad \forall u,v \in P^1(K_i) \tag{3.148}$$

where vectors of nodal values of a given function are underscored. $\underline{u}(K_i)$ is an E^3 vector of node values on a given element and $\underline{u}$ is an E^n vector of node values on the whole mesh. The standard E^3 inner product is implied in the right hand side of (3.148). Explicitly the entries of the element stiffness matrix are given by

$$A(K_i)_{jl} \;=\; (\psi_j, \psi_l)_{E,K_i} \;=\; \int_{K_i} \nabla\psi_j \cdot \nabla\psi_l \, d\Omega, \quad j,l = 1,2,3 \tag{3.149}$$

To obtain an error estimate over a particular element K_i, we shall use, as an auxiliary function, the first order Taylor approximation $\mathcal{T}u$ of $u \in C^2(\bar{\Omega})$ around an arbitrary point r_0 within that element:

$$(\mathcal{T}u)(r_0, r) \;=\; u(r_0) \;+\; \nabla u(r_0) \cdot (r - r_0)$$

Fig. 3.39 illustrates this in 1D. The difference between the *nodal values* of the Taylor approximation $\mathcal{T}u$ and the exact function u (or its FE interpolant Πu) is "small" (on the order of $\mathcal{O}(h^2)$ for linear approximation) and shape-independent in 2D and 3D. At the same time, the difference between $\mathcal{T}u$ and Πu *in the energy norm* is generally much greater: not only is the order of approximation lower, but also the error can be adversely affected by the element shape. Obviously, somewhere in the transition from the nodal values to the energy norm the precision is lost. Since the energy norm in the FE space is governed by the FE stiffness matrix, the large error in the energy norm indicates the presence of a large eigenvalue of the matrix.

For a more precise analysis, let us write the function u as its Taylor approximation plus a (small) residual term $R(r_0, r)$:

$$u(r) \;=\; (\mathcal{T}u)(r_0, r) \;+\; R(r_0, r), \qquad r \in K_i,$$

where $R(r_0, r)$ can be expressed via the second derivatives of u at an interior point of the segment $[r, r_0]$:

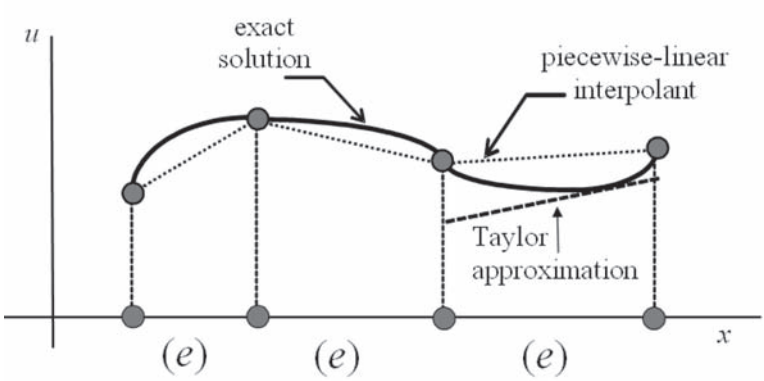


Fig. 3.39. Taylor approximation vs. FE interpolation. Function u (solid line) is approximated by its piecewise-linear node interpolant Πu (dashed line) and by element-wise Taylor approximations $\mathcal{T}u$ (dotted lines). The energy norm difference between Πu and $\mathcal{T}u$ is generally much greater than the difference in their node values.

$$R(r_0, r) \;=\; \sum_{|\alpha|=2} \frac{D^\alpha u(r_0 + \theta(r - r_0))}{\alpha!}\,(r - r_0)^\alpha, \quad 0 \le \theta < 1 \tag{3.150}$$

with the standard shorthand notation for the multi-index $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ (in the current example $d = 2$), $|\alpha| = \alpha_1 + \alpha_2 + \ldots + \alpha_d$, $\alpha! = \alpha_1!\alpha_2!\ldots, \alpha_d!$, and partial derivatives

$$D^\alpha u \;=\; \frac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \ldots \partial x_d^{\alpha_d}}$$

It follows from (3.150) that the residual term is indeed small, in the sense that

$$|R(r_0, r)| \;=\; |(\mathcal{T}u)(r_0, r) - u(r)| \;\le\; \|u\|_{2,\infty,K_i} |r - r_0|^2 \tag{3.151}$$

$$|\nabla R(r_0, r)| \;=\; |\nabla(\mathcal{T}u)(r_0, r) - \nabla u(r)| \;\le\; \|u\|_{2,\infty,K_i} |r - r_0| \tag{3.152}$$

where

$$\|u\|_{m,\infty,K} \;=\; \sum_{|\alpha|=m} |D^\alpha u|_{L_\infty(K)} \tag{3.153}$$

The key observations leading to the maximum eigenvalue condition can be informally summarized as follows:

1. The Taylor approximation is uniformly accurate within the element due to (3.151), (3.152) and is completely independent of the element geometry. Therefore, for the purpose of evaluating the dependence of the interpolation error on shape, $\mathcal{T}u$ can be used in lieu of u, i.e. one can consider the difference $\Pi u - \mathcal{T}u$ instead of $\Pi u - u$.

2. The energy norm of the difference $\Pi u - \mathcal{T} u$ is generally much higher than the nodal values of $\Pi u - \mathcal{T} u$: the nodal values are of the order $\mathcal{O}(h^2)$ *and independent of element shape* due to (3.151), while the energy norm is $\mathcal{O}(h)$ and depends on the shape.
3. The above observations imply that in the transition from node values to the energy norm the accuracy is lost. Since within the element K_i both u and $\mathcal{T} u$ lie in the FE space $P^1(K_i)$, and since in this space the energy norm is induced by the element stiffness matrix $A(K_i)$, a large energy norm can be attributed to the presence of a large eigenvalue in that stiffness matrix.

The first of these statements can be made precise by writing

$$\begin{aligned} \|\Pi u - u\|_{E,K_i} &\leq \|\Pi u - \mathcal{T} u\|_{E,K_i} + \|\mathcal{T} u - u\|_{E,K_i} \\ &\leq \|\Pi u - \mathcal{T} u\|_{E,K_i} + c h_i V_i^{\frac{1}{2}} \|u\|_{2,\infty,K_i} \end{aligned} \tag{3.154}$$

where the second inequality follows from estimate (3.152) of the Taylor residual, $V_i = \text{meas}(K_i)$, and c is an absolute constant independent of the element shape and of u.

We now focus on the term $\|\Pi u - \mathcal{T} u\|_{E,K_i}$ in (3.154). Restrictions of both u and $\mathcal{T} u$ to K_i lie in the FE space $P^1(K_i)$, and therefore

$$\|\Pi u - \mathcal{T} u\|_{E,K_i} = (A(K_i)(\underline{u}(K_i) - \underline{\mathcal{T} u}(K_i)),\ \underline{u}(K_i) - \underline{\mathcal{T} u}(K_i))^{\frac{1}{2}} \tag{3.155}$$

The standard Euclidean inner product in E^3 is implied in the right hand side of (3.155), and we recall that the underscore denotes Euclidean vectors of nodal values.

It follows immediately from (3.155) that

$$\|\Pi u - \mathcal{T} u\|_{E,K_i} \leq \max_{x \neq 0, x \in \mathbb{R}^3} \left(\frac{(A(K_i)\underline{x},\ \underline{x})}{(\underline{x},\ \underline{x})} \right)^{\frac{1}{2}} \|\underline{u}(K_i) - \underline{\mathcal{T} u}(K_i)\|_{E^3} \tag{3.156}$$

that is,

$$\|\Pi u - \mathcal{T} u\|_{E,K_i} \leq \lambda_{\max}^{\frac{1}{2}}(A(K_i)) \, \|\underline{u}(K_i) - \underline{\mathcal{T} u}(K_i)\|_{E^3} \tag{3.157}$$

In the right hand side of (3.157), $\lambda_{\max}$ is the maximum eigenvalue of the element stiffness matrix (3.148), (3.149). The difference $\underline{u}(K_i) - \underline{\mathcal{T} u}(K_i)$ is the error vector for the Taylor expansion at element nodes, and due to the uniformity (3.151), (3.152) of the Taylor approximation, we have

$$\|\underline{u}(K_i) - \underline{\mathcal{T} u}(K_i)\|_{E^3} \leq c h_i^2 \, |u|_{2,\infty,K_i} \tag{3.158}$$

(the generic constant c is not necessarily the same in all occurrences). Combining (3.157) and (3.158), we obtain the element-wise estimate

$$\|\Pi u - \mathcal{T} u\|_{E,K_i} \leq c h_i^2 \, \lambda_{\max}^{\frac{1}{2}}(A(K_i)) |u|_{2,\infty,K_i} \tag{3.159}$$

or, taking into account the triangle inequality (3.154),

$$\|\Pi u - u\|_{E,K_i} \;\le\; c\left(h_i^2\lambda_{\max}^{\frac{1}{2}}(A(K_i)) + h_i V_i^{\frac{1}{2}}\right)\,|u|_{2,\infty,K_i} \tag{3.160}$$

The corresponding global estimate is

$$\|\Pi u - u\|_{E,\Omega} \;\le\; c|u|_{2,\infty,\Omega}\left[\sum_{K_i\in M}\left(h_i^4\lambda_{\max}(A(K_i)) + h_i^2 V_i\right)\right]^{\frac{1}{2}} \tag{3.161}$$

This result can be simplified by noting that $\lambda_{\max}(A(K_i)) \le \mathrm{tr}A(K_i)$, $\sum_{K_i}\mathrm{tr}A(K_i) = \mathrm{tr}A$, $\sum_{K_i} V_i = V$, where A is the global stiffness matrix and $V = \mathrm{meas}(\Omega)$:

$$\|\Pi u - u\|_{E,\Omega} \;\le\; c|u|_{2,\infty,\Omega}\left[h^2(\mathrm{tr}^{\frac{1}{2}}A + hV^{\frac{1}{2}})\right] \tag{3.162}$$

Alternatively, one can factor out the element area V_i in (3.160) to obtain

$$\|\Pi u - u\|_{E,K_i} \;\le\; cV_i^{\frac{1}{2}}|u|_{2,\infty,K_i}\left(h_i^2(\lambda_{\max}^{\frac{1}{2}}(\hat{A}(K_i)) + h_i\right) \tag{3.163}$$

where the hat denotes the scaled element stiffness matrix $\hat{A}(K_i) = A(K_i)/V_i$. Then the global error estimate simplifies to

$$\|\Pi u - u\|_{E,\Omega} \;\le\; cV^{\frac{1}{2}}\max_{K_i\in M}\left[\left(h_i^2\lambda_{\max}^{\frac{1}{2}}(\hat{A}(K_i)) + h_i\right)\,|u|_{2,\infty,K_i}\right] \tag{3.164}$$

The maximum eigenvalue can again be replaced with the (easily computable) matrix trace.

Remark 8. The trace- and max-terms in estimates (3.162), (3.164) are not of the order $\mathcal{O}(h^2)$ as it might appear, but $\mathcal{O}(h)$, since both $\mathrm{tr}A$ and $\lambda_{\max}(\hat{A}(K_i))$ are $\mathcal{O}(h^{-2})$.

The analysis above can be generalized, without any substantial changes, to elements of any geometric shape and order:

Theorem 5. *Let M be a finite element mesh in a bounded domain $\Omega \in \mathbb{R}^d$ ($d \ge 1$) and let the following assumptions hold for any (scalar or vector) function $u \in (C^{m+1})^s(\bar{\Omega})$: $\Omega \to \mathbb{R}^s$, with some nonnegative integers m and s.*
(A.1) A given energy (semi)norm is bounded as

$$|u|_{E,K_i}^2 \;\le\; \sum_{j=0}^{\nu} c_j^2\,|u|_{j,\infty,K_i}^2, \quad c_\nu > 0, \quad V_i = \mathrm{meas}(K_i) \tag{3.165}$$

for any element K_i, with constants c_j independent of the element.
(A.2) The FE approximation space over K_i contains all polynomials of degree $\le m$.

(A.3) The FE degrees of freedom – linear functionals ψ_j over the FE space – are bounded as

$$|\psi_j(K_i)u| \;\le\; \sum_{l=0}^{\mu} \tilde{c}_l^2\, |u|_{l,\infty,K_i}, \quad \tilde{c}_\mu > 0 \tag{3.166}$$

for a certain $\mu \ge 0$, with some absolute constants $\tilde{c}_l$. Then

$$\|\Pi u - u\|_{E,\Omega} \;\le\; c|u|_{m+1,\infty,\Omega} \left(h^\kappa \operatorname{tr}^{\frac{1}{2}} A \;+\; h^\tau V^{\frac{1}{2}} \right) \tag{3.167}$$

where $V = meas(\Omega)$, $\kappa = m+1-\mu$, $\tau = m+1-\nu$, and the global stiffness matrix A is given by (3.148), (3.149). Alternatively,

$$\|\Pi u - u\|_{E,\Omega} \;\le\; c|u|_{m+1,\infty,\Omega}\, V^{\frac{1}{2}} \max_{K_i} \left(h_i^\kappa \lambda_{\max}^{\frac{1}{2}}(\hat{A}(K_i)) \;+\; h^\tau \right) \tag{3.168}$$

where $\hat{A}(K_i) = A(K_i)/V_i$.

The meaning of the parameters in the theorem is as follows: m characterizes the level of smoothness of the function that is being approximated; $s = 1$ for scalar functions and $s > 1$ for vector functions with s components, approximated component-wise; ν is the highest derivative "contained" in the energy (semi)norm; μ is the highest derivative in the degrees of freedom.

Example 7. First order tetrahedral node elements satisfy assumptions (A.1–A.3). Indeed, for the energy norm (3.147), condition (3.165) holds with $\nu = 1$, $c_0 = 0$, $c_1 = 1$. (A.2) is satisfied with $m = 1$, and (A.3) is valid because of the uniformity (3.151) of the Taylor approximation within a sufficiently small circle.

More generally, (A.3) is satisfied if FE degrees of freedom are represented by a linear combination of values of the function and its derivatives at some specified points of the finite element.

Example 8. First order triangular nodal elements.

Let the seminorm be (3.147), (3.146). Then the trace of the scaled element stiffness matrix has a simple geometric interpretation. The diagonal elements of the matrix are equal to d_j^{-2} $(j = 1, 2, 3)$, where the d_js are the altitudes of the triangle (Fig. 3.40). Therefore, denoting interior angles of the triangle with ϕ_j and its sides with l_j, and assuming $h_i = \operatorname{diam}(K_i) = l_1 \ge l_2 \ge l_3$, one obtains

$$\lambda_{\max}(\hat{A}(K_i)) \;\le\; \operatorname{Tr} \hat{A}(K_i) \;=\; \sum_{j=1}^{3} d_j^{-2} \;=\; h_i^{-2} \sum_{j=1}^{3} \left(\frac{l_1}{d_j} \right)^2$$

$$< \; h_i^{-2} \left[\left(\frac{l_2 + l_3}{d_1} \right)^2 + \left(\frac{l_1}{d_2} \right)^2 + \left(\frac{l_1}{d_3} \right)^2 \right] \;\le\; 3h_i^{-2} (\sin^{-2}\phi_2 + \sin^{-2}\phi_3) \tag{3.169}$$

which leads to Zlámal's *minimum* angle condition. This result is reasonable but not optimal, which shows that the maximum eigenvalue criterion does not generally guarantee the sharpest estimates. Nevertheless the optimal condition for first order elements – the maximum angle condition – will be obtained below by applying the maximum eigenvalue criterion *to the Nédélec–Whitney–Bossavit edge elements.*

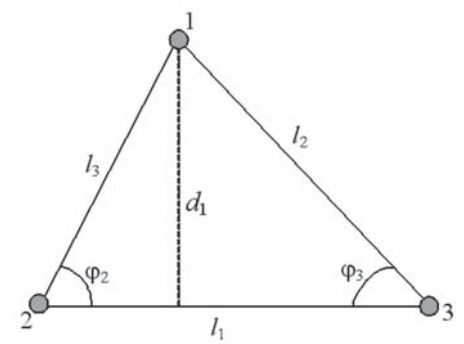


Fig. 3.40. Geometric parameters of a triangular element K_i.

Example 9. For *first order tetrahedral elements*, the trace of the scaled nodal stiffness matrix can also be interpreted geometrically. A simple transformation similar to (3.169) [Tsu98b] yields the minimum–maximum angle condition for angles ϕ_{jl} between edges j and faces l: ϕ_{jl} are to be bounded away from both zero and π to ensure that the interpolation error tends to zero as the element size decreases.

For higher order scalar elements on triangles and tetrahedra, the matrix trace is evaluated in an analogous but lengthier way, and the estimate is similar, except for an additional factor that depends on the order of the element.[56]

Example 10. L_2-approximation of scalar functions on tetrahedral or triangular node elements. Suppose that Ω is a two- or three-dimensional polygonal (polyhedral) domain and that continuous and discrete spaces are taken as $L_2(\Omega)$ and $P^1(M)$, respectively, for a given triangular/tetrahedral mesh. Assume that the energy inner product and norm are the standard L_2 ones. This energy norm in the FE space is induced by the "mass matrix"

$$A(K_i)_{jl} \;=\; \int_{K_i} \phi_i \phi_l \, d\Omega; \quad \hat{A}(K_i)_{jl} \;=\; V_i^{-1} \int_{K_i} \phi_i \phi_l \, d\Omega \tag{3.170}$$

[56] Here we are discussing shape dependence only, as the factor related to the dependence of the approximation error on the element size is obvious.

For first order tetrahedral elements, this matrix is given by (3.102) on p. 123, repeated here for convenience:

$$\hat{A}(K_i) \;=\; \frac{1}{20}\begin{pmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 \end{pmatrix} \tag{3.171}$$

The maximum eigenvalue of $\hat{A}$ is equal to $1/4$ and does not depend on the element shape. Assumptions (A.1–A.3) of Theorem 5 hold with $m = 1$, $\mu = \nu = 0$, $c_0 = \tilde{c_0} = 1$, and therefore approximation of the potential is shape-independent due to (3.168). This known result is obtained here directly from the maximum eigenvalue condition.

Analysis for first order triangular elements is completely similar, and the conclusion is the same.

Example 11. $(L_2)^3$-approximation of conservative vector fields on tetrahedral or triangular meshes. In lieu of the piecewise-linear approximation of u on a triangular or tetrahedral mesh, one may consider the equivalent piecewise-constant approximation of ∇u on the same mesh. Despite the equivalence of the two approximations, the corresponding error estimates are not necessarily the same, since the maximum eigenvalue criterion is not guaranteed to give optimal results in all cases.

It therefore makes sense to apply the maximum eigenvalue condition to interpolation errors in $L_2^3(\Omega)$ for a conservative field $\mathbf{q} = \nabla u$ on a tetrahedral mesh. To this end, a version of the first order edge element on a tetrahedron K may be defined by the Whitney–Nédélec–Bossavit space (see Section 3.12, p. 139) spanned by functions w_{jl}, $1 \le j < k \le 4$

$$w_{jk} \;=\; l_{jk}(\lambda_j \nabla \lambda_k \;-\; \lambda_k \nabla \lambda_j) \tag{3.172}$$

where the λs are the barycentric coordinates of the tetrahedron. (They also are the nodal basis functions of the first order scalar element.) The scaling factor l_{jk}, introduced for convenience of further analysis, is the length of edge jk.

As a reminder, the dimension of the Whitney–Nédélec–Bossavit space over one element is equal to the number of element edges, i.e. three for triangles and six for tetrahedra. There is the corresponding global FE space $W(M)$ (W for "Whitney") over the whole mesh M. It is a subspace of $H(\mathrm{curl}, \Omega) = \{\mathbf{q} : \mathbf{q} \in L_2^3(\Omega), \nabla \times \mathbf{q} \in L_2^3(\Omega)\}$.

The "exactness property" (see A. Bossavit [Bos88b, Bos88a]) of this space is critical for the analysis of this section: if the computational domain is simply connected, a vector field in $W(M)$ is conservative if and only if it is the gradient of a continuous piecewise-linear scalar field $u \in P^1(\Omega)$ on the same mesh. The exactness property remains valid if the definitions of functional spaces are amended in a natural way to include Dirichlet conditions for the tangential components of the field on part of the domain boundary.

The degrees of freedom are defined as the average values of the tangential components of the field along the edges:

$$\psi_{jk}(\mathbf{q}) \;=\; l_{jk}^{-1} \int_{\text{edge } jk} \mathbf{q} \cdot d\tau \tag{3.173}$$

The maximum eigenvalue estimate could now be directly applied to interpolation in $W(M)$. However, a more accurate result is obtained by taking the maximum in the right hand side of the generic expression (3.157) in a *subspace* of $\mathbb{R}^6$. This subspace corresponds to $\mathbb{R}^6$-vectors $\underline{q}$ of edge d.o.f.'s for vector fields $\mathbf{q} \in \nabla P^1(K)$. Within a given element, such vector fields are in fact constant and can therefore be treated as vectors in $\mathbb{R}^3$. The subspace maximization of (3.157) yields

$$\max_{\mathbf{q}\in\mathbb{R}^3} \frac{(A(K_i)\,\underline{q}, \underline{q})_{E^6}}{\|\underline{q}\|^2_{E^6}} \;=\; \max_{\mathbf{q}\in\mathbb{R}^3} \frac{\int_{K_i} |\mathbf{q}|^2\, d\Omega}{\|\underline{q}\|^2_{E^6}} \;=\; \text{meas}(K_i) \max_{\mathbf{q}\in\mathbb{R}^3} \frac{|\mathbf{q}|^2}{\|\underline{q}\|^2_{E^6}} \tag{3.174}$$

To evaluate the ratio in the right hand side, note that the $\mathbb{R}^6$-vector $\underline{q}$ of the edge projections of $\mathbf{q}$ is related to the column vector of the Cartesian components $\underline{q}_C = (q_x, q_y, q_z)^T$ of $\mathbf{q}$ as

$$\underline{q} \;=\; E^T(K_i)\,\underline{q}_C \tag{3.175}$$

Here $E^T(K_i)$ is the 3×6 "edge shape matrix" whose columns are the unit vectors e_α $(1 \le \alpha \le 6)$ directed along the tetrahedral edges (in either of the two directions):

$$E(K) \;=\; [e_1 \,|\, e_2 \,|\, e_3 \,|\, e_4 \,|\, e_5 \,|\, e_6] \tag{3.176}$$

The element index i has been dropped for simplicity of notation. Singular value decomposition (G.H. Golub & C.F. Van Loan [GL96]) of this matrix is the key to further analysis:

$$E(K) \;=\; L\Sigma Q^T \tag{3.177}$$

where L is a 3×3 orthogonal matrix, Q is a 6×6 orthogonal matrix, and

$$\Sigma \;=\; \begin{pmatrix} \sigma_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_3 & 0 & 0 & 0 \end{pmatrix} \tag{3.178}$$

is the matrix containing the singular values $\sigma_1 \ge \sigma_2 \ge \sigma_3 \ge 0$ of the edge shape matrix E. Hence

$$\|\underline{q}\|^2_{E^6} \;=\; (E^T(K)\underline{q}_C,\, E^T(K)\underline{q}_C) \;=\; (E(K)\,E^T(K)\underline{q}_C,\, \underline{q}_C) \tag{3.179}$$

and the last maximum in (3.174) is

$$\max_{\mathbf{q}\in\mathbb{R}^3} \frac{|\mathbf{q}|^2}{\|\underline{q}\|^2_{E^6}} \;=\; \max_{q_C\in\mathbb{R}^3} \frac{(q_C,\, q_C)}{(E(K)\,E^T(K)q_C,\, q_C)}$$

$$= \lambda_{\min}^{-1}(E(K)\,E^T(K)) \;=\; \sigma_{\min}^{-2}(E(K)) \tag{3.180}$$

where $\lambda_{\min}$ is the minimum eigenvalue, and $\sigma_{\min}$ is the minimum singular value (if $\underline{q} = 0$, $\sigma_{\min} = 0$ is implied in (3.180)). The last equality of (3.180) is based on the well known fact (G.H. Golub & C.F. Van Loan [GL96]) that

$$\sigma_j^2(E(K)) \;=\; \lambda_j(E(K)E^T(K)) \;=\; \lambda_j(E^T(K)E(K)) \tag{3.181}$$

where for $E^T E$ only the nonzero eigenvalues are considered.

The minimum singular value $\sigma_{\min}(E(K))$ is zero if and only if there exists a vector $\mathbf{q}$ orthogonal to all six edge vectors e_j (so that $\underline{q} = 0$), that is, if and only if all edges are coplanar (and the tetrahedron is thus degenerate). In general, the minimum singular value characterizes the "level of degeneracy," or "flatness" of a tetrahedron.

In the maximum eigenvalue condition, parameters now have the following values:
$m = 0$ (the pertinent Taylor approximation is just a vector constant);
$\nu = 0$ (L_2-norm);
$\mu = 0$ (the d.o.f.'s are the tangential field components along the edges, with no derivatives involved).

Hence $\kappa = \tau = 1$ in (3.167), (3.168), and, with (3.174), (3.180) in mind, one has

$$\|\Pi u - u\|_{E,\Omega} \;\le\; c\left[\sum_{K_i \in M} h_i^2\,\left(\sigma_{\min}^{-2}(E(K_i)) + 1\right)\,V_i\,|u|_{2,\infty,K_i}\right]^{\frac{1}{2}} \tag{3.182}$$

This is a global error estimate, but each individual term in the sum represents a (squared) element-wise error.

It is not hard to establish an upper bound for $\sigma_{\min}(E(K_i))$. Indeed,

$$\sigma_{\min}(E) \;\le\; \frac{1}{3}\sum_{j=1}^{3}\sigma_j^2(E) \;=\; \frac{1}{3}\,\mathrm{tr}(E^T E) \;=\; 2 \tag{3.183}$$

so (3.182) can be simplified:

$$\|\Pi u - u\|_{E,\Omega} \;\le\; c\left[\sum_{K_i \in M} h_i^2\,\sigma_{\min}^{-2}(E(K_i))\,V_i\,|u|_{2,\infty,K_i}\right]^{\frac{1}{2}} \tag{3.184}$$

Analysis for triangular elements is quite similar, and the final result is the same. In addition, for triangular elements the following proposition holds:

Proposition 6. *The* minimum singular value *criterion for the* 2×3 *edge shape matrix of a first order triangular element is equivalent to the Synge–Babuška–Aziz* maximum angle *condition.*

Proof. The minimum singular value can be explicitly evaluated in this case. Letting the x-axis run along one of the edges of the element (Fig. 3.41), one has the edge shape matrix in the form

$$E = \begin{pmatrix} 1 & \cos\phi_1 & -\cos\phi_2 \\ 0 & \cos\phi_1 & -\cos\phi_2 \end{pmatrix} \tag{3.185}$$

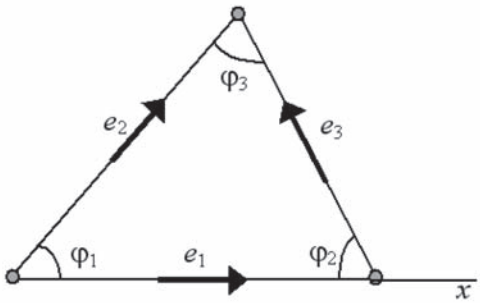


Fig. 3.41. Three unit edge vectors for a triangular element.

The trace of $(EE^T)^{-1}$ is found to be (with some help of symbolic algebra)

$$\mathrm{Tr}(EE^T)^{-1} = \frac{3}{\sin^2\phi_1 + \sin^2\phi_2 + \sin^2\phi_3} \tag{3.186}$$

Since $\mathrm{tr}(EE^T)^{-1} = \lambda_1(EE^T)^{-1} + \lambda_2(EE^T)^{-1} = \lambda_1^{-1}(EE^T) + \lambda_2^{-1}(EE^T) = \sigma_1^{-2}(E) + \sigma_2^{-2}(E)$, one has

$$\frac{1}{3}(\sin^2\phi_1 + \sin^2\phi_2 + \sin^2\phi_3) \le \sigma_{\min}^2(E) \le \frac{2}{3}(\sin^2\phi_1 + \sin^2\phi_2 + \sin^2\phi_3) \tag{3.187}$$

It can immediately be seen from these inequalities that the minimum singular value can be arbitrarily close to zero if and only if the maximum angle approaches π (and the other two angles approach zero).

3.14.3 Geometric Implications of the Singular Value Condition

The Minimum Singular Value vs. the Inscribed Sphere Criterion

The most common geometric characteristic of a tetrahedral finite element K is the ratio of radius r of the inscribed sphere to the maximum edge $l_{\max}$. The following inequality shows that the singular value criterion is less stringent than the $r/l_{\max}$ ratio.

Proposition 7. *[Tsu98a]*

$$\sigma_{\min}(E) \;\geq\; \frac{r}{l_{\max}} \tag{3.188}$$

Proof. We appeal to a geometric interpretation of the minimum singular value. For vector $\underline{q} \in \mathbb{R}^6$ of edge projections of an arbitrary nonzero Cartesian vector $q_C \in \mathbb{R}^3$, we have

$$\sigma_{\min}(E) \;\leq\; \frac{\|\underline{q}\|_{E^6}}{\|\underline{q}_C\|_{E^3}} \tag{3.189}$$

where the exact equality is achieved when (and only when) q_C is an eigenvector corresponding to the minimum eigenvalue of EE^T. Thus

$$\sigma_{\min}(E) \;=\; \min_{\|\underline{q}_C\|_{E^3}=1} \|\underline{q}\|_{E^6} \tag{3.190}$$

We can assume without loss of generality that the first node of the tetrahedron is placed at the origin and that the tetrahedron is scaled to $l_{\max} = 1$ and rotated to have the unit eigenvector $\mathbf{v}$ corresponding to the minimum eigenvalue of EE^T run along the z-axis. Let z_β ($\beta = 1, 2, 3, 4$) be the z-coordinates of the nodes. According to (3.190),

$$\begin{aligned}\sigma^2_{\min}(E) \;&=\; \|\underline{v}\|_{E^6} \;=\; \sum_{1\leq\alpha<\beta\leq 4} (\mathbf{v}\cdot e_{\alpha\beta})^2 \;\geq\; \sum_{2\leq\alpha<\beta\leq 4} (\mathbf{v}\cdot e_{1\beta})^2 \\ &=\; \sum_{2\leq\alpha<\beta\leq 4} \left(\frac{z_\beta}{l_{1\beta}}\right)^2 \;\geq\; \sum_{2\leq\alpha<\beta\leq 4} z_\beta^2 \end{aligned} \tag{3.191}$$

where each edge is now labeled by its two end nodes; $l_{1\beta}$ is the length of the edge connecting nodes 1 and β, $l_{1\beta} \leq l_{\max} = 1$. The first summation in (3.191) is over all six edges $\alpha\beta$, while the subsequent summations are over three nodes $\beta = 2, 3, 4$ and the corresponding edges 1β.

It immediately follows from (3.191) that for all nodes $|z_\beta| \leq \sigma_{\min}$. The scaled tetrahedron therefore lies entirely between the planes $z = \pm\sigma_{\min}$; hence $r \leq \sigma_{\min}$ □

Remark 9. The converse statement that $\sigma_{\min}(E) \leq cr/l_{\max}$ is not true. Consider a sequence of tetrahedra with three mutually orthogonal edges, two of these edges being of unit length and the third one tending to zero. Then the radius of the inscribed sphere tends to zero, while the minimum singular value remains equal to one [TP98].

Jamet's Condition

P. Jamet [Jam76] obtained accurate interpolation error estimates under quite general assumptions. For tetrahedral elements, the governing factor in Jamet's estimate is $\cos\theta$, where θ is defined as

$$\theta = \max_{\xi} \min_{i} \theta_i, \quad i = 1, \ldots, 6 \tag{3.192}$$

Here θ_i is the angle between an arbitrary unit vector $\xi \in \mathbb{R}^3$ and the unit edge vector e_i; the minimum is taken over all edges, and the maximum is taken over all unit vectors ξ. (Jamet's angle characterizes, geometrically, how far the edges are from being perpendicular to a certain vector ξ.) It turns out that Jamet's measure is very closely related to the minimum singular value criterion. Indeed, one can rewrite (3.192) as

$$\cos\theta = \min_{\xi} \max_{i} \cos\theta_i = \min_{\xi} \|E^T\xi\|_{\infty,E^6} \tag{3.193}$$

versus

$$\sigma_{\min}(E) = \min_{\xi} \|E^T\xi\|_{2,E^6}$$

That is, the only theoretical difference between Jamet's $\cos\theta$ and the minimum singular value of the edge shape matrix is in the matrix norm employed. This adds further credence to the analysis and results that involve eigenvalues and singular values of FE matrices.

Jamet's condition is more general than the present formulation of the minimum singular value estimate (in particular, Jamet's analysis applies to any Sobolev norms in W_p^m). On the other hand, computational algorithms (SVD) for the minimum singular value, unlike for Jamet's angle, are well established and readily available.

The Minimum Singular Value vs. Angle Conditions

The minimum singular value of the edge shape matrix can be computed and used as an *a priori algebraic* measure of the interpolation error; alternatively, $\sigma_{\min}^{-2}$ can be replaced with $\operatorname{tr}(EE^T)^{-1}$. At the same time, given that $\sigma_{\min}$ characterizes the level of linear independence of the element edges and the overall "flatness" of the element, geometric implications of the minimum singular value condition are worth investigating. The following proposition shows that asymptotically the singular value criterion is equally or less restrictive than criteria based on solid angles.

Proposition 8. *Let $\{K_i\}_{i=1}^{\infty}$ be a sequence of tetrahedra with their diameters h_i tending to zero, and let E_i be the edge shape matrix (3.176) of K_i. Then, if the minimum singular value condition is violated, i.e. if $\sigma_{\min}(E_i) \to 0$ as $i \to \infty$, then there exists a subsequence of $\{K_i\}$ for which all solid (trihedral) angles tend to either zero or 2π.*

Proof. As before, without loss of generality, each tetrahedron K_i can be assumed to have one of its nodes at the origin of a Cartesian system and to be rotated to have the minimum eigenvector of EE^T run along the z axis.

Let S be the unit sphere in $\mathbb{R}^3$. To each tetrahedron K_i in the sequence there corresponds a point $P_i \equiv (e_1^{(i)}, \ldots, e_6^{(i)}) \in S^6$ representing the six unit

edge vectors $e_l^{(i)}$ of K_i. Since S^6 is compact, one can select a subsequence of $\{K_i\}$, again denoted $\{K_i\}$, with the respective points P_i converging to a point $P_\infty \equiv (e_1^{(\infty)}, \ldots, e_6^{(\infty)}) \in S^6$. Since

$$\sigma_{\min}(E_i) \;=\; \sum_{1 \le l \le 6} \left(e_l^{(i)} \cdot \hat{z} \right)^2 , \tag{3.194}$$

all six unit vectors $e_l^{(\infty)}$ must lie in the xy-plane, and consequently the trihedral angle formed by any three of these vectors is zero or 2π. Since the trihedral angles depend continuously on P_i, the proposition follows.

Remark 10. If a solid angle tends to zero, it does not necessarily imply that the minimum singular value does, too. A counterexample is the same as in Remark 9.

A valid asymptotic condition is for the maximum solid angle to be bounded away from 2π. Indeed, if this condition were violated, the three edges forming the largest trihedral angle would tend to three distinct coplanar vectors. Hence all six edges would in the limit be coplanar, which corresponds to a zero singular value.

M. Križek [Kŷ2] introduced a sufficient convergence condition requiring that all dihedral angles, *as well as all face angles*, be bounded away from π. The Proposition below shows that the minimum singular value criterion is equally or less restrictive than the Křížek condition.

Proposition 9. *Let γ_{dj} ($j = 1, 2, \ldots, 6$) be the dihedral angles of a tetrahedron K and γ_{fl}^{β} ($l = 1, 2, 3$; $1 \le \beta \le 4$) be the angles of each triangular face β. Let γ_{d0} be the angle with the* maximum *sine of all dihedral angles:* $\sin\gamma_{d0} = \max(\sin\gamma_{dj})$. *Similarly, for each face β, let $\sin\gamma_{f0}^{\beta}$ be the* maximum *of all $\sin\gamma_{fl}^{\beta}$ for face β. Finally, let $\sin\gamma_{f0}$ be the* minimum *of $\sin\gamma_{f0}^{\beta}$ over all faces β; i.e.*

$$\sin\gamma_{f0} \;=\; \min_{1 \le \beta \le 4} \max_{1 \le l \le 3} \sin\gamma_{fl}^{\beta}$$

Then

$$\sigma_{\min}(E(K)) \;\ge\; \left(\frac{2}{3}\right)^{\frac{1}{2}} \sin\frac{\gamma_{d0}}{2} \, \sin\gamma_{f0} \tag{3.195}$$

Proof. Consider the two faces forming the dihedral angle γ_{d0} with the maximum sine of all $\sin\gamma_{dj}$. Let one of these faces lie in the xz-plane and let their common edge be on the z axis, with one node at the origin as shown in Fig. 3.42.

Further, consider an arbitrary unit vector

$$v \;=\; \hat{x}\sin\theta\cos\phi \;+\; \hat{y}\sin\theta\sin\phi \;+\; \hat{z}\cos\theta$$

in $\mathbb{R}^3$, and let v_1 and v_2 be its projections on faces (1) and (2), respectively (Fig. 3.42). Then

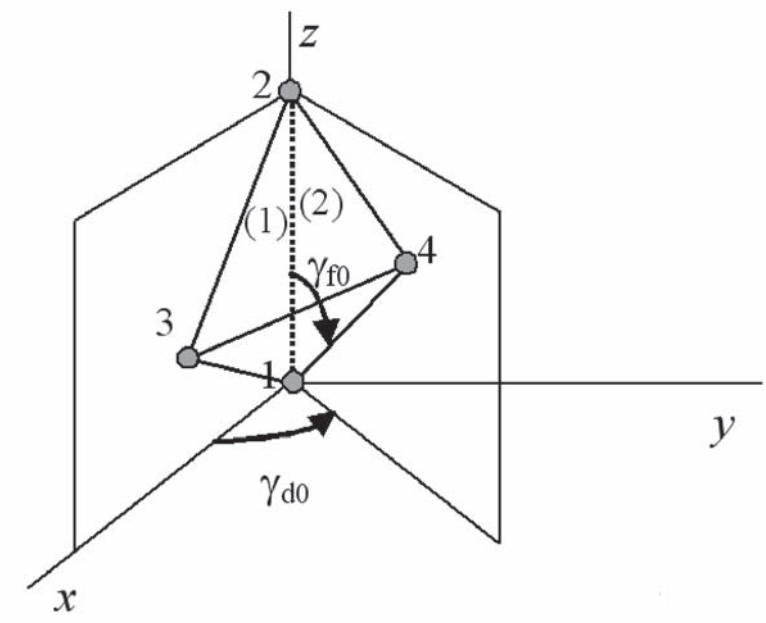


Fig. 3.42. Tetrahedral nodes and critical angles. 1, 2, 3 are the nodes of face (1); 1, 4, 2 are the nodes of face (2).

$$v_1^2 \;=\; \sin^2\theta\cos^2\phi \;+\; \cos^2\theta, \qquad v_2^2 \;=\; \sin^2\theta\cos^2(\phi-\gamma_{d0}) \;+\; \cos^2\theta$$

Further projecting v_1 and v_2 on each of the three edges of the respective faces (1) and (2) and using expression (3.187) for the minimum singular value of the edge shape matrix of a triangle, one obtains:

$$2\sum_{1\le j\le 5} v_{ej}^2 \;\ge\; (\sin^2\theta\cos^2\phi \;+\; \cos^2\theta)\,\frac{1}{3}(\sin^2\gamma_{f1}^{(1)}+\sin^2\gamma_{f2}^{(1)}+\sin^2\gamma_{f3}^{(1)})$$

$$+\;(\sin^2\theta\cos^2(\phi-\gamma_{d0}) \;+\; \cos^2\theta)\,\frac{1}{3}(\sin^2\gamma_{f1}^{(2)}+\sin^2\gamma_{f2}^{(2)}+\sin^2\gamma_{f3}^{(2)})$$

$$\ge\;(\sin^2\theta\cos^2\phi+\cos^2\theta)\,\frac{1}{3}\,\sin^2\gamma_{f0}^{(1)}+(\sin^2\theta\cos^2(\phi-\gamma_{d0})+\cos^2\theta)\,\frac{1}{3}\sin^2\gamma_{f0}^{(2)}$$

$$\ge\;\left[\sin^2\theta(\cos^2\phi \;+\; \cos^2(\phi-\gamma_{d0}))+2\cos^2\theta\right]\,\frac{1}{3}\sin^2\gamma_{f0}$$

$$\ge\;\left[\sin^2\theta\cdot 2\sin^2\frac{\gamma_{d0}}{2} \;+\; 2\cos^2\theta\right]\,\frac{1}{3}\,\sin^2\gamma_{f0} \;\ge\; \frac{2}{3}\sin^2\frac{\gamma_{d0}}{2}\,\sin^2\gamma_{f0}$$

(The factor of two in the left hand side is due to the fact that the projection on edge 1-2 is counted twice in the right hand side. Summation over $1 \le j \le 5$ excludes edge 3-4.) □

Conversely, let the Křížek condition be violated for some sequence of tetrahedra. Suppose first that a dihedral angle tends to π in that sequence. Then all six edges tend to positions in one fixed plane (after a possible rotation of each tetrahedron in the sequence). The edge projections of a unit vector perpendicular to that plane will tend to zero, and so will the minimum singular value of the edge shape matrix. Similarly, if one of the face angles tends to π, then all the edges of that face tend to positions on one straight line, and consequently all six edges again tend to positions in one plane, and $\sigma_{\min}(E(K_i)) \to 0$.

It follows that the minimum singular value and Křížek conditions are equivalent as asymptotic criteria of convergence of piecewise-linear interpolation on a family of tetrahedral meshes.

The Minimum Singular Value vs. Trihedral Volume

Consider first three unit edge vectors corresponding to a common tetrahedral node. There is a 3×3 submatrix E_3 of E associated in the obvious way with these three edges. The volume of the parallelepiped based on the three unit vectors is

$$V_3 \;=\; |\det E_3| \tag{3.196}$$

Both $\sigma_{\min}(E_3)$ and V_3 characterize the level of linear independence of the three unit vectors, suggesting a connection between these two measures. Since the product of the eigenvalues is equal to the determinant, and the sum of the eigenvalues is equal to the trace, one has

$$\begin{aligned}[\sigma_1(E_3)\,\sigma_2(E_3)\,\sigma_3(E_3)]^2 \;&=\; \lambda_1(E_3^T E_3)\,\lambda_2(E_3^T E_3)\,\lambda_3(E_3^T E_3)\\ &=\; \det(E_3^T E_3) \;=\; \det{}^2(E_3) \;=\; V_3^2\end{aligned}$$

that is,

$$\sigma_1(E_3)\,\sigma_2(E_3)\,\sigma_3(E_3) \;=\; V_3 \tag{3.197}$$

Similarly,

$$\begin{aligned}\sigma_1^2(E_3) \,+\, \sigma_2^2(E_3) \,+\, \sigma_3^2(E_3) \;&=\; \lambda_1(E_3^T E_3) \,+\, \lambda_2(E_3^T E_3) \,+\, \lambda_3(E_3^T E_3)\\ &=\; \operatorname{tr} E_3^T E_3 \;=\; 1+1+1 \;=\; 3\end{aligned} \tag{3.198}$$

From (3.198), one immediately obtains

$$1 \;\le\; \sigma_{\max}^2(E_3) \;\le\; 3$$

and therefore it follows from (3.197), with the convention $\sigma_{\max} = \sigma_1 \ge \sigma_2 \ge \sigma_3 = \sigma_{\min}$, that

$$\sigma_{\min}^4(E_3) \;\le\; \sigma_1^2(E_3)\,\sigma_2^2(E_3) \;\le\; \sigma_1^2(E_3)\,\sigma_2^2(E_3)\,\sigma_3^2(E_3) \;=\; V_3^2 \;\le\; 9\sigma_{\min}^2(E_3)$$

Hence

$$\frac{V_3}{3} \;\leq\; \sigma_{\min}(E_3) \;\leq\; V_3^{\frac{1}{2}} \tag{3.199}$$

The right inequality indicates that $\sigma_{\min}$ and V_3 could be of different "orders of magnitude". Examples given in [TP98] demonstrate that the inequalities in (3.199) cannot be asymptotically improved.

The maximum "trihedral volume" V_3 based on three unit edge vectors[57] may serve as a sufficient convergence condition for FE interpolation. However, due to a "nonlinear" relationship (3.199) between V_3 and $\sigma_{\min}$, volume V_3 is expected to be a less accurate *a priori* error measure than $\sigma_{\min}$.

Necessity of the minimum singular value condition

There are several, and not equivalent, definitions of a shape condition being "essential" for the convergence of FE approximation. These definitions can be subdivided into the following broad categories:

(a): if a shape condition is violated, the interpolation error *may fail to* tend to zero for some families $\{K_i\}$ of elements (of a given type) with $h_i = \operatorname{diam}(K_i) \to 0$ and for some admissible functions;

(b): if a shape condition is violated for *any* family of elements $\{K_i\}$ of a given type, the interpolation error *will not* tend to zero for some admissible functions;

(c),(d): same as (a) and (b), respectively, but for the error of the numerical *solution* (the Galerkin projection) instead of the interpolation error.

Clearly, (b) is stronger than (a). Categories (c)–(d) are much more difficult to establish than (a)–(b). For first order triangular elements the minimum and the maximum angle conditions are both "essential" in the sense of (a), but only the maximum angle condition (equivalent in this case to the minimum singular value criterion) is "essential" in the sense of (b). M. Křížek [Kŷ2] proved that his condition is essential in the (a)-sense. Babuka and Aziz [BA76] showed that the maximum angle condition for triangles is essential in the (c) sense.

It is easy to demonstrate that the minimum singular value condition is essential in the (a) sense; in fact, either of the two examples given by Křížek [Kŷ2] suffices for this (the minimum singular value condition is violated, and there is no convergence). Establishing the necessity of the minimum singular value condition in a stronger (b) sense is more difficult. To this end, we need a definition that allows for freedom of solid rotation and translation of tetrahedral elements.

[57] Strictly speaking, the maximum should be taken over all triples of edges, not necessarily having a common node.

Definition 8. *For a given tetrahedron K, the equivalence class of tetrahedra obtained from K by rigid rotations and/or translations is denoted with $\hat{K}$. Any energy norm $\|u\|_{E,K}$ on K is extended to the equivalence class $\hat{K}$ by*

$$\|u\|_{E,\hat{K},\Omega} \;=\; \sup\{\|u\|_{E,K}, \quad K \in \hat{K}, \quad K \subset \Omega\} \tag{3.200}$$

The necessity of the minimum singular value condition in the (b)-sense can then be stated as follows.

Proposition 10. *Let $\{K_i\}_{i=1}^{\infty}$ be an arbitrary sequence of tetrahedral elements such that $h_i \equiv \mathrm{diam}(K_i) \to 0$ and $\sigma_{\min}(E(K_i)) \to 0$ as $i \to \infty$. In addition, assume that the ratio of the maximum edge $h_i \equiv l_{\max}(K_i)$ to the minimum edge $l_{\min}(K_i)$ is uniformly bounded on $\{K_i\}_{i=1}^{\infty}$. Then there exists a function $u \in C^2(\bar{\Omega})$ for which the H^1-error of linear interpolation tends to infinity:*

$$\|\Pi_1(K_i)u \;-\; u\|_{H^1,\hat{K}_i,\Omega} \;\to\; \infty \tag{3.201}$$

Proof. The starting point is exactly as in the proofs of Proposition 7 and Proposition 8. Since arbitrary translations and rotations are allowed by Definition 8 in the norm used in (3.200), the minimum eigenvector of EE^T may be assumed to run along the z-axis. Then, for the elements in the sequence, all edges will tend to the xy-plane.

Hence one can select a subsequence of elements, again denoted as $\{K_i\}$, with their nodes $r_1^{(i)}$, $r_2^{(i)}$, $r_3^{(i)}$, $r_4^{(i)}$ converging to four points r_{1-4} in the xy-plane, with $r_1^{(i)} = r_1 = 0$ for all i. Due to the assumed boundedness of $l_{\max}/l_{\min}$, the four points r_{1-4} must be distinct.

Consider first the case when no three of the points $r_j \equiv (x_j, y_j)$ ($j = 1$, 2, 3, 4) lie on a straight line. Introduce a Cartesian system with point r_1 at the origin, point r_2 on the x axis at $(x_2, 0)$, point r_3 at (x_3, y_3), point r_4 at (x_4, y_4), and points $r_j^{(i)} \equiv (x_j^{(i)}, y_j^{(i)}, z_j^{(i)})$. Since by assumption points r_1, r_2, r_3 do not lie on the same line,

$$y_3 \;\neq\; 0 \tag{3.202}$$

For each K_i, there exists a quadratic function of x, y

$$u_{\mathrm{quadr}}^{(i)}(a^{(i)}; x, y) \;=\; \frac{1}{2}\, a_1^{(i)} x^2 \;+\; a_2^{(i)} xy \;+\; a_3^{(i)} y^2 \tag{3.203}$$

with a coefficient vector $a^{(i)} = (a_1^{(i)}, a_2^{(i)}, a_3^{(i)})^T$ such that

$$u_{\mathrm{quadr}}^{(i)}(a^{(i)}; x, y) \;=\; \frac{z^{(i)}}{\|z^{(i)}\|_2} \tag{3.204}$$

Indeed, the suitable $a^{(i)}$ is given by

$$a^{(i)} \;=\; \frac{1}{\|z^{(i)}\|_2}\, (Q^{(i)})^{-1} z^{(i)} \tag{3.205}$$

where the matrix

$$Q^{(i)} \;=\; \begin{pmatrix} \frac{1}{2}\,x_2^{(i)2} & x_2^{(i)}\,y_2^{(i)} & \frac{1}{2}\,y_2^{(i)2} \\ \frac{1}{2}\,x_3^{(i)2} & x_3^{(i)}\,y_3^{(i)} & \frac{1}{2}\,y_3^{(i)2} \\ \frac{1}{2}\,x_4^{(i)2} & x_4^{(i)}\,y_4^{(i)} & \frac{1}{2}\,y_4^{(i)2} \end{pmatrix}$$

can easily be verified to be nonsingular if no three points r_{1-4} lie on one straight line. Moreover, since $Q^{(i)}$ is a continuous function of coordinates $x_j^{(i)}$, $(Q^{(i)})^{-1}$ exists and is uniformly bounded for the sequence[58] and $\lim_{i\to\infty}(Q^{(i)})^{-1} = Q^{-1}$ exists. Therefore the sequence of coefficient vectors $a^{(i)}$ defined by (3.205) is bounded, and one can select a converging subsequence $a^{(i)} \to a^{(\infty)}$, with the corresponding function $u_{\text{quadr}}^{(\infty)} = u_{\text{quadr}}(a^{(\infty)}; x, y)$.

According to (3.204), the coefficients of $u_{\text{quadr}}^{(i)}$ are chosen in such a way that its linear interpolant over K_i is simply

$$u_{\text{lin}}^{(i)} \;\equiv\; \Pi_1(K_i)\, u_{\text{quadr}}^{(i)} \;=\; \frac{z^{(i)}}{\|z^{(i)}\|_2}$$

Therefore the z-derivative of the interpolation error

$$\partial_z(u_{\text{lin}}^{(i)} - u_{\text{quadr}}^{(i)}) \;=\; \frac{1}{\|z^{(i)}\|_2} \;\to\; \infty$$

because $\|z^{(i)}\|_2 \to 0$ as $\sigma_{\min}(E(K_i)) \to 0$. This implies that the interpolation error for the limiting function $u_{\text{quadr}}^{(\infty)}$ also tends to infinity, despite the boundedness of the seminorm $|u_{\text{quadr}}|_{2,\infty,K_i} = \|a^{(\infty)}\|_1$.

If three of the points r_{1-4} lie on one straight line, the corresponding face is degenerate, and the proof can be essentially repeated in two dimensions in the plane of this face.

3.14.4 Condition Number and Approximation

Practical experience has shown (see e.g. F.-X. Zgainski *et al.* [ZMC+97]) that the condition number of the FE stiffness matrix is a useful measure of mesh quality. Since the condition number strongly affects the performance of iterative systems solvers, it is not surprising that slow convergence of the solvers and poor accuracy of the solution (due to poor quality of the FE mesh) typically go hand in hand.

Based on the results of this section, it can be argued that poor approximation and poor conditioning of the system are related to each other *indirectly*: both of these quantities stem from the maximum eigenvalue of the global stiffness matrix. This connection is schematically illustrated in Fig. 3.43. (The

[58] With a possible exception of a finite set of indices.

minimum eigenvalue has no bearing on interpolation accuracy and can typically be viewed as a fixed parameter associated with the size of the computational domain.[59])

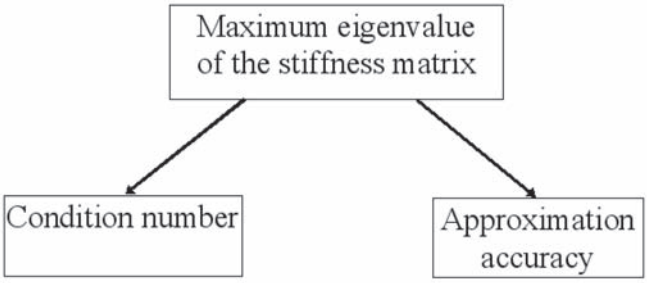


Fig. 3.43. A large eigenvalue of the FE stiffness matrix is a common source of both ill-conditioning of the FE system and poor accuracy of the solution.

3.14.5 Discussion of Algebraic and Geometric *a priori* Estimates

We have explored the dual algebraic/geometric nature of finite element interpolation errors. From the algebraic perspective, the error was shown to be governed by the maximum eigenvalue of the FE stiffness matrix. When the maximum eigenvalue estimate is applied to triangular and tetrahedral elements, several known geometric conditions and several nonstandard results are obtained. For triangular elements in particular, Zlámal's minimum angle condition and the Synge–Babuška–Aziz maximum angle condition are recovered.

For tetrahedral elements, the maximum eigenvalue estimate leads to an interesting result. The shape of tetrahedral elements turns out to be accurately represented, in the FE context, by the minimum singular value of the "edge shape matrix". This singular value characterizes, on the one hand, the "flatness" of the element and, on the other hand, the accuracy of the FE interpolation.

There are several links between the minimum singular value and some geometric parameters of the tetrahedron, but the minimum singular value is, in some well-defined sense, one of the most precise measures. (Jamet's condition is another one.)

Due to its generality, the maximum eigenvalue condition can be applied in cases where no other shape criteria are immediately available. For example,

[59] Strictly speaking, the ratio of maximum/minimum eigenvalues is in general a suitable measure of conditioning for symmetric positive definite matrices only. This case is implicitly assumed, to avoid further complications.

anisotropy of material parameters should result, intuitively, in some "scaling" of the coordinate axes before any geometric accuracy criteria can be considered. In contrast, the maximum eigenvalue criterion accommodates anisotropy automatically, since material parameters are built into the stiffness matrix.

This criterion can be applied to elements of any shape and order but is not without limitations. First, it provides *a priori* estimates only; it remains to be seen whether similar ideas can be used to enhance a posteriori estimates critical for adaptive mesh refinement (Section 3.13).

Second, the maximum eigenvalue criterion is a sufficient but not generally a necessary condition; it does not guarantee the best error estimate. This is well illustrated by two cases considered in this section: (a) for conservative fields on Whitney edge elements, the result (expressed via the minimum singular value of the edge shape matrix) *is* optimal; (b) at the same time, for triangular node elements the maximum eigenvalue criterion leads to Zlámal's minimum angle condition rather than to the more accurate Synge–Babuška–Aziz maximum angle condition.

The theoretical results provide general and easy-to-implement *a priori* criteria of FE accuracy. The computational overhead in the overall FE procedure is negligible. For tetrahedral elements in particular, the precise characterization of shape via the minimum singular value of the element "edge shape matrix" can be recommended for engineering practice. Experimental results reported by M. Dorica & D.D. Giannacopoulos [DG05] and by A. Plaks & myself [TP98] support this conclusion.

3.15 Special Topic: Generalized FEM

3.15.1 Description of the Method

A detailed explanation and analysis of Generalized FEM proposed originally by I. Babuška & J.M. Melenk [MB96, BM97] is widely available (e.g. T. Strouboulis *et al.* [SBC00]). Of all interesting features of GFEM, the most salient one is its ability to employ a variety of special non-polynomial approximating functions. In particular, jumps of the normal derivatives of the potential at interface boundaries can be represented by special basis functions. Strouboulis *et al.* [SBC00] present an extensive set of application examples with special functions for material inclusions in stress analysis. Babuška *et al.* [BCO94] applied Generalized FEM (before the method was referred to as such) to problems with "rough" coefficients – discontinuities at material interfaces. A. Plaks *et al.* [PTFY03] implemented GFEM for problems with magnetized particles.

In GFEM the computational domain Ω is covered with overlapping subdomains ("patches") $\Omega^{(i)}$, and different local approximations are merged by Partition of Unity (PU) $\{\Omega_i\}_{i=1}^{n_{\text{patches}}}$ on this system of patches. More precisely, a set of PU functions $\{\varphi^{(i)}\}$, $1 \le i \le n_{\text{patches}}$ is constructed to satisfy

$$\sum_{i=1}^{n_{\text{patches}}} \varphi^{(i)} \equiv 1 \quad \text{in } \Omega, \quad \operatorname{supp} \varphi^{(i)} = \Omega^{(i)} \tag{3.206}$$

That is, each function $\varphi^{(i)}$ is associated with the respective patch $\Omega^{(i)}$ and vanishes outside that patch.

Then the global solution u can be decomposed into its "patch components" $u^{(i)}$

$$u = u \sum_{i=1}^{n_{\text{patches}}} \varphi^{(i)} = \sum_{i=1}^{n_{\text{patches}}} u\varphi^{(i)} = \sum_{i=1}^{n_{\text{patches}}} u^{(i)}, \text{ with } u^{(i)} \equiv u\varphi^{(i)} \tag{3.207}$$

Fig. 3.44 gives a simple 1D illustration of the PU principle, with just two overlapping patches. A seamless transition from the solution in the first patch to the solution in the second patch is achieved by multiplying these individual solutions by the weighting functions $\varphi^{(1)}$ and $\varphi^{(2)}$, respectively. As a reference point moves from left to right, the weight of the first solution gradually decreases, while simultaneously the weight of the second solution increases.

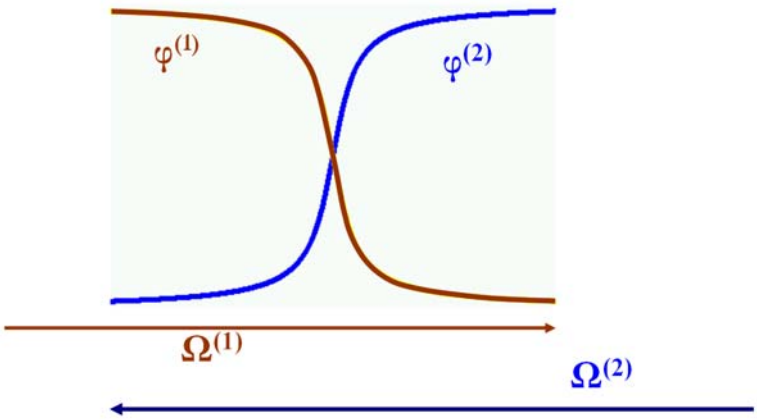


Fig. 3.44. The idea of partition of unity illustrated in 1D: weighting functions $\varphi^{(1)}$ and $\varphi^{(2)}$ are used to merge two solutions in the overlapping subdomains. The sum of the weighting functions is unity everywhere. (Reprinted by permission from [Tsu06] ©2006 Elsevier.)

Decomposition (3.207) is valid for the exact solution but can equally well be used for assembling a global *approximate* solution from the local ones. Suppose that locally, within each patch $\Omega^{(i)}$, the exact solution u can be approximated by a linear combination $u_h^{(i)}$ of some approximating functions $g_\alpha^{(i)}$:

$$u_h^{(i)} = \sum_\alpha c_\alpha^{(i)} g_\alpha^{(i)} \tag{3.208}$$

$c_\alpha^{(i)}$ being some (real- or complex-valued) coefficients. The final system of approximating functions $\psi_\alpha^{(i)}$ is built with $\varphi^{(i)}$ as weight functions:

$$\psi_\alpha^{(i)} = g_\alpha^{(i)} \varphi^{(i)} \tag{3.209}$$

The global approximation error is guaranteed to be bounded by the local (patch-wise) errors [BM97], [SBC00], [BBO03], with rigorously provable estimates of the global error in terms of local errors and the norms of the PU functions ϕ.

3.15.2 Trade-offs

The multiplication by $\varphi^{(i)}$ in (3.209) guarantees seamless merging of patch-wise approximations, with rigorously provable estimates of the global error in terms of local errors and the norms of the PU functions φ [BM97]. On the negative side, however, this multiplication complicates the set of approximating functions and tends to make it more ill-conditioned (in some cases even linearly dependent, see [BM97]). For positive definite problems, the linear dependence can be tolerated because the resultant algebraic system remains consistent and positive-semidefinite and can be handled by clever linear algebra algorithms (see T. Strouboulis *et al.* [SBC00] for further information).

The "no free lunch" cliché applies fully to GFEM. While the rigid requirements on mesh structure and the approximating functions are greatly relaxed, the computational burden is shifted toward numerical quadratures that need to be computed in the Galerkin method over the intersections of overlapping patches. This complex task can be accomplished in general only by adaptive numerical integration. The efficiency of this integration is critical for the overall performance of the algorithm.

In addition, GFEM-PU may lead to a combinatorial increase in the number of degrees of freedom. For illustration, consider a regular hexahedral mesh where a "patch" is defined as a set of eight hexahedra around a common node. In the presence of material boundaries, it is sensible to replace the usual eight trilinear basis functions with eight special functions satisfying the derivative jump condition at the interface (see also Chapter 4). In GFEM-PU, each of these special functions gets multiplied by the "shape function" φ of the patch. As each hexahedral element of the mesh is an intersection of eight patches (centered at its eight respective nodes) and each of these patches contributes eight approximating functions, the stiffness matrix for elements close to material interfaces is 64×64 instead of the usual 8×8. For all of the above reasons, alternative approaches may be worth exploring. One such approach that generalizes finite *difference*, rather than finite element, analysis is discussed in Chapter 4.

3.16 Summary and Further Reading

The Finite Element Method is arguably the most powerful computational tool ever invented. Its solid variational foundation makes the method remarkably robust – often beyond the areas where a complete mathematical analysis is available.

FEM is well established in traditional branches of engineering such as stress analysis, heat transfer, electromagnetic fields in machines and microwave circuits, etc. However, FEM has not yet been taken full advantage of in some areas of nanoscale simulation. Examples include nano-photonics and nano-optics – more specifically, plasmonic field enhancement by particle clusters, scattering of light by optical tips in near-field microscopy, and wave propagation in photonic crystal devices. These and other cases presented in Chapter 7 will hopefully stimulate further applications of FEM in nanoscale science and technology.

The present chapter explains the fundamentals of FEM (the underlying variational principles, finite elements and spaces, FE matrices, algorithmic implementation) and provides an overview of state-of-the-art techniques of FE analysis (adaptive mesh refinement and multigrid algorithms). The chapter also covers more advanced topics: edge elements, *a priori* estimates of numerical accuracy as a function of element shape, and Generalized FEM.

Adaptive *hp*-refinement aims at the most effective use of the computational resources by constructing quasi-optimal meshes: the density of elements is higher in regions where the solution is less smooth and changes more rapidly; the density is lower in regions of smooth variation of the solution. Adaptive techniques are now an integral part of the commercial FE packages. The same is true for edge elements in electromagnetic applications: the gap between the elegant mathematical theory and practical utility was bridged in the 1990s, especially after it became clear that many families of edge elements, in contrast with the nodal ones, do not produce nonphysical eigensolutions known as "spurious modes".

Generalized FEM occupies a niche in practical applications. This will most likely continue to be the case, although the niche may grow to some extent. The power of GFEM lies in its ability to use a wide selection of approximations not limited to element-wise polynomials as in the standard FEM. This could be a great advantage in many cases where particular features of the physical field or potential, such as singularities, boundary layers, dipole-like behavior, etc., are known *a priori* and can therefore be accurately represented by special approximating functions. However, there is a substantial price to be paid for this advantage: complex numerical quadratures, increased number of unknowns, and possible ill-conditioning or in some cases even linear dependence of the system of approximating functions.

The special section on *a priori* error estimates in this chapter examines the links between algebraic and geometric accuracy measures. While it is well known that "flat" elements provide poor numerical approximation of the

solution, it is argued in Section 3.14 that the "true source" of the error is of algebraic nature. This source can be traced to the maximum eigenvalue of the FE stiffness matrix and, in the case of triangular and tetrahedral elements, to the minimum singular value of the "edge shape matrix". It is shown that the latter measure is, in some sense, a precise one, and its connection with various geometric parameters is examined.

The reader who would like to learn more about Finite Element analysis is in an enviable position. There are many excellent books and papers on all aspects of FEM, written from the engineering, mathematical and computational perspectives. Researchers and developers of engineering applications cannot go wrong with the books by O.C. Zienkiewicz *et al.* [ZTZ05, ZT05]. In engineering electromagnetics, P.P. Silvester's group was the first to apply finite elements; his book with R.L. Ferrari [SF90] is still valuable. J. Jin's more recent monograph [Jin02] is a very good source of information on FEM in electromagnetics and includes, in addition to standard subjects, chapters on vector finite elements, absorbing boundary conditions, finite element – boundary integral methods and on time-domain analysis. The book by J.L. Volakis *et al.* [VCK98] also covers vector elements, as well as applications to radiation and scattering and hybrid finite element – boundary integral methods. Several books are focused on the applications of FEM to low-frequency electromagnetic fields in electric machines and devices: J.P. A. Bastos & N. Sadowski [aPABS03], S. Salon & M. V.K. Chari [SC99], G. Meunier (ed.) [Meu07].

On the mathematical side, there are several magnificent books as well. The works by G. Strang & G.J. Fix [SF73] and I. Babuška, A.K. Aziz & B. Szabó [BA72, SB91] are classical. The main reference on the mathematical treatment of FEM in electromagnetism is P. Monk's monograph [Mon03].

The monograph by L. Demkowicz [Dem06] bridges mathematical theory and applications, with the emphasis on *hp*-adaptivity. The book deals with elliptic and wave problems and includes 1D and 2D codes developed by Demkowicz & co-workers.

Finally, A. Bossavit's book [Bos98] is in a category of its own due to its unconventional approach and style. The focus of this book is on the mathematical principles and structures underlying FE methods in electromagnetism – in particular, concepts of variational analysis, differential geometry and algebraic topology. While the content is mostly mathematical, Bossavit's style of writing makes the material accessible to non-experts (still, the reader will need enough patience and perseverance to understand the book).

In the coming years, I look forward to seeing further applications of FEM in the simulation of micro- and nanoscale systems. Electromagnetic field analysis in optics and photonics seems particularly interesting, as it may well lead to the development of new devices and materials with completely unconventional properties and behavior (Chapter 7).

3.17 Appendix: Generalized Curl and Divergence

This section is an extension of Appendix 6.15 (p. 343) on generalized functions (distributions) and their derivatives.

The conventional representation of the divergence and curl operators – say, in Cartesian coordinates – requires differentiability:

$$\nabla \cdot \mathbf{A} = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z}; \quad (\nabla \times \mathbf{A})_x = \frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}; \quad \text{etc.}$$

However, derivatives in these expressions can be treated in the generalized sense of distributions (see Appendix 6.15), thereby extending the notion of divergence and curl to functions that are not differentiable in the standard sense of differential calculus.

Example 12. The $\mathbf{A}$ field with a step-like x-component, $A_x = 0$ for $x < 0$ and $A_x = 1$ for $x \geq 0$, and zero y- and z-components, has generalized divergence $\nabla \cdot \mathbf{A} = \delta(x)$. For the electric field, this Dirac-delta divergence corresponds to a surface charge.

Example 13. The $\mathbf{A}$ field with a step-like z-component, $A_z = 0$ for $y < 0$ and $A_z = 1$ for $y \geq 0$, and zero x- and y-components, has generalized curl $\nabla \times \mathbf{A} = \delta(y)\hat{x}$. For the magnetic field, this Dirac-delta curl corresponds to a surface current.

Instead of appealing to the Cartesian representation of divergence and curl with generalized derivatives, one can give an equivalent but coordinate-free definition via integration-by-parts identities. For divergence,

$$(\nabla \cdot \mathbf{A}, \phi) = -(\mathbf{A}, \nabla\phi) \tag{3.210}$$

where the inner product is that of $\mathbf{L}_2$. This identity, in the regular calculus sense, follows from the calculus formula

$$\nabla \cdot (\mathbf{A}\phi) = \phi\nabla \cdot \mathbf{A} + \mathbf{A} \cdot \nabla\phi$$

if fields $\mathbf{A}$, ϕ are continuously differentiable and ϕ has a compact support. (The latter requirement ensures that the surface integral term in the integration by parts vanishes). One can then extend the notion of divergence to non-differentiable fields and *define* generalized divergence as the linear functional

$$\langle \nabla \cdot \mathbf{A}, \phi \rangle \equiv -(\mathbf{A}, \nabla\phi) \tag{3.211}$$

over smooth scalar functions ϕ with a compact support. Equation (3.210) ensures that the extended definition is consistent with the regular calculus version of divergence as long as the vector field is smooth.

For a vector field that has a jump of its normal component across a surface S, but is otherwise smooth, the generalized divergence is

$$\langle \nabla \cdot \mathbf{A},\, \phi \rangle \;\equiv\; -\,(\mathbf{A},\, \nabla\phi) \;=\; \int_S [A_n]\phi\, dS \;+\; (\{\nabla \cdot \mathbf{A}\},\, \phi)$$

where integration by parts was applied. Here $[A_n] = A_{n+} - A_{n-}$ is the jump of the normal component of the vector field across S ($n+$ referring to the region into which the normal to S is pointing). Thus

$$\nabla \cdot \mathbf{A} \;=\; [A_n]\, \delta_S \;+\; \{\nabla \cdot \mathbf{A}\} \tag{3.212}$$

where generalized divergence is implied in the left hand side and divergence in its regular calculus sense is specified by the curly brackets in the right hand side. This is V.S. Vladimirov's notation; see Appendix 6.15 on p. 343 and also footnote 18 on p. 320 and 44 on p. 347.

The curl operator is generalized in a similar fashion:

$$(\nabla \times \mathbf{A},\, \mathbf{B}) \;=\; (\mathbf{A},\, \nabla \times \mathbf{B}) \tag{3.213}$$

where the inner product is again that of $\mathbf{L}_2$. This identity, in the regular calculus sense, follows from (3.128) if fields $\mathbf{A}$, $\mathbf{B}$ are continuously differentiable and $\mathbf{B}$ has a compact support. Generalized curl is defined as

$$\langle \nabla \times \mathbf{A}, \mathbf{B} \rangle \;\equiv\; (\mathbf{A}, \nabla \times \mathbf{B}) \tag{3.214}$$

over smooth vector functions $\mathbf{B}$ with a compact support. For vector fields with a discontinuous tangential component across a surface S, but smooth otherwise, the generalized curl is

$$\nabla \times \mathbf{A} \;=\; [\mathbf{A} \times \hat{n}]\, \delta_S \;+\; \{\nabla \times \mathbf{A}\} \tag{3.215}$$

This formula is analogous, and obtained in a similar way, to expression (3.212) for generalized divergence. A key observation in the context of edge elements is that a jump of the tangential component of a vector field across a surface leads to the Dirac-delta term for the generalized curl on this surface; Example 13 on p. 186 is a simple but representative illustration of this property that is not difficult to verify in general. The tangential component is continuous if and only if the generalized curl exists as a regular function, not only a distribution.

4

Flexible Local Approximation MEthods (FLAME)

This chapter is based to a large extent upon my papers [Tsu05a, Tsu05b, Tsu06].

4.1 A Preview

Although the Finite Element Method (FEM) described in Chapter 3 is one of the most powerful and general analysis techniques, in some cases the complicated FE meshes, data structures and solvers can become computationally expensive or even impractical.

Finite Difference (FD) algorithms (Chapter 2), on the other hand, operate on geometrically simple grids and the data structures associated with them are much simpler than those of FEM. The system solvers also tend to be more efficient. The downside, in comparison with FEM, is relatively poor numerical accuracy at material interfaces not conforming to the simple FD grid.

This leads to a legitimate question: given a regular grid not geometrically conforming to material interfaces, what is – in some sense – "the best" one can do? The answer, in general, is *not* the classical FD schemes. This chapter argues in favor of a new FD calculus referred to by the acronym "FLAME": Flexible Local Approximation MEthods. The word "Flexible" implies that any desired approximation of the solution (exponentials, spherical harmonics, plane waves, generic or special polynomials, etc.) can be incorporated directly into the FD scheme. This is in contrast with Taylor polynomial expansions that form the basis of standard FD.

In FLAME, approximation is always treated as *local*, with the intention to represent *local* features of the solution that in many cases may qualitatively be known *a priori* (for example, the behavior of the potential near a material interface; see also Section 4.5 on p. 219).

As a preview, consider a simple 2D test problem: a cylindrical magnetic particle, with relative permeability $\mu_{\mathrm{p}} = 100$, immersed in a uniform external

field. A contour plot and a grayscale plot of the magnetic scalar potential u (the magnetic field $\mathbf{H} = -\nabla u$) are shown in Fig. 4.1 for illustration.[1]

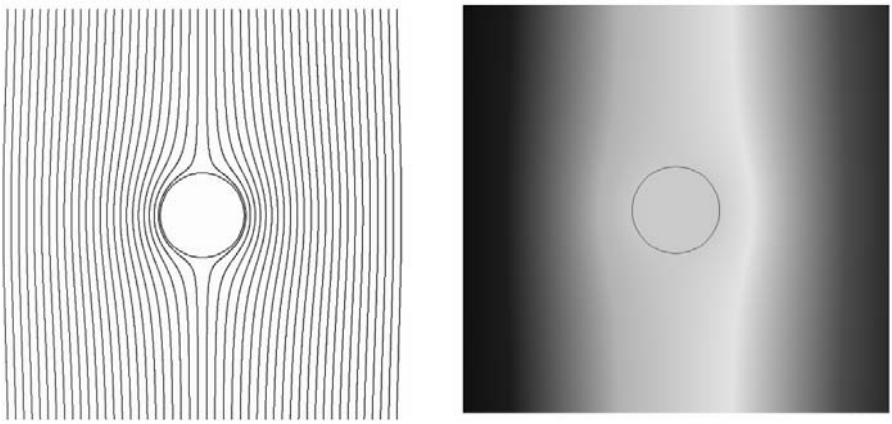


Fig. 4.1. A contour plot and a grayscale plot of the magnetostatic potential for a cylindrical particle in a uniform external field.

Fig. 4.2 compares two meshes that give about the same level of numerical accuracy for this problem. The Finite Element mesh has 31,537 nodes, 62,592 second order triangular elements and 125,665 degrees of freedom (d.o.f.); the relative error in the potential at the nodes is 2.07×10^{-8}. The FLAME grid has 900 d.o.f. (30×30), and the relative error in the potential at the nodes is 2.77×10^{-8} if 9-point (3×3) stencils are used. The high accuracy of FLAME schemes is due to the approximating functions employed in FLAME. For the particle problem, these functions are cylindrical harmonics that represent the behavior of the potential in the vicinity of the particle much better than the Taylor polynomials do in standard FD. This chapter explains how FLAME schemes are constructed.

First, Section 4.2 provides an introduction to FLAME and highlights the main ideas behind it. Some of these ideas, such as Trefftz basis functions *in the finite-difference context*, multivalued approximation, trade-off between conformity and flexibility of approximation, are nonstandard.

As a preliminary example, FLAME is developed for the (trivial) case of the 1D Laplace equation in Section 4.2.6 to fix ideas. General construction of Trefftz–FLAME is presented in Section 4.3, where case studies in 1D, 2D and 3D are provided. In Trefftz–FLAME, the approximating functions are chosen as *local* solutions of the underlying differential equation. In a number of practically interesting cases, the local solutions are not difficult to derive analytically; in addition to this chapter, computational examples are given

[1] The electrostatic problem for a dielectric particle is completely analogous.

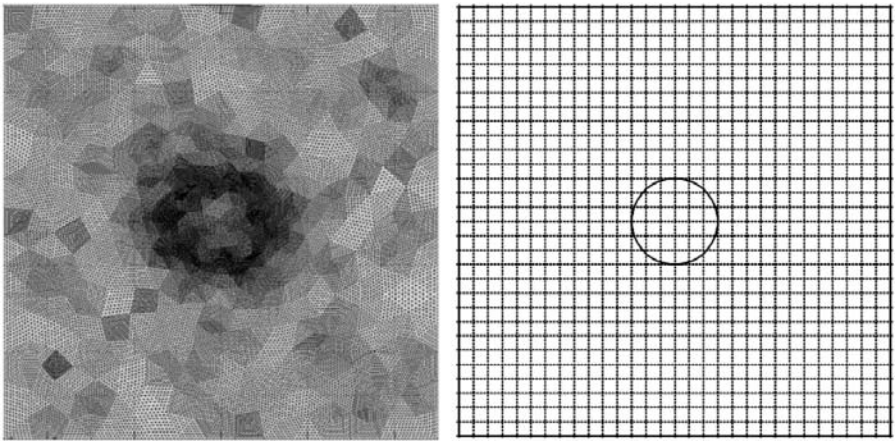


Fig. 4.2. Two meshes yielding about the same level of accuracy for the particle problem. The FE mesh has 31,537 nodes, 62,592 second order triangular elements and 125,665 degrees of freedom. The FLAME grid has 900 degrees of freedom. (Reprinted by permission from [Tsu06] ©2006 Elsevier.)

in Chapter 6 (electrostatic interactions of colloidal particles) and Chapter 7 (electromagnetic field enhancement by plasmonic particles and waves in photonic crystals).

Section 4.5 reviews existing classes of methods with nontraditional approximation: Generalized FEM (GFEM), variational homogenization, pseudospectral methods, and others. FLAME borrows some features of these methods (most notably, flexible approximation from Generalized FEM) but is not a particular case of any of them. *Some* existing methods turn out to be particular cases of FLAME: the exact schemes by R.E. Mickens [Mic94, Mic00]; the Hadley schemes for electromagnetic wave propagation [Had02a]; the "Measured Equation of Invariance" [MPC+94] by K.K. Mei *et al.*

The chapter concludes with a discussion (Section 4.6) and appendices on the variaitonal version of FLAME, the 9-point 2D FLAME for the wave equation, and the Fréchet derivative.

4.2 Perspectives on Generalized FD Schemes

4.2.1 Perspective #1: Basis Functions Not Limited to Polynomials

Taylor polynomials are generic and may be the best option when no *a priori* information about the solution is available. When the local behavior of the solution is known, more effective approximations can usually be generated.

For example, if the solution exhibits boundary layers, wave-like behavior, dipole components, etc., in certain regions, as schematically shown in

Fig. 4.3, then it may be appropriate to use exponentials, sinusoids, dipole harmonics, and so on, as approximating functions in the respective regions. The subsequent sections of this chapter show how this can be accomplished in a generalized finite-difference framework.

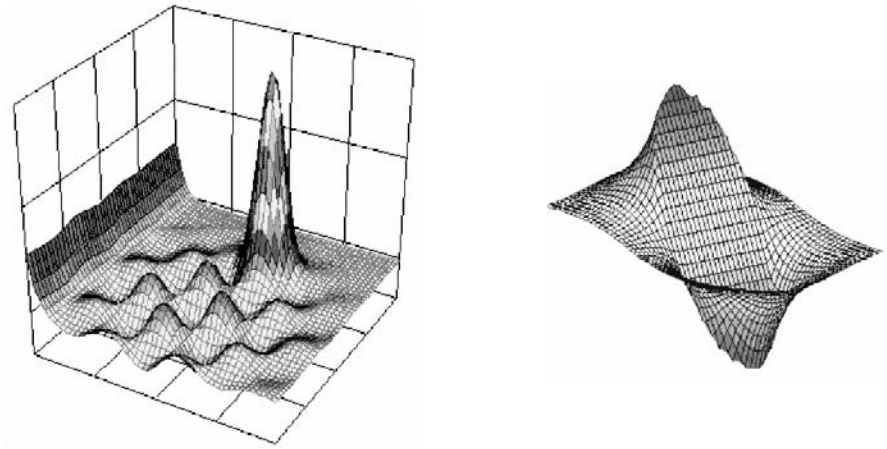


Fig. 4.3. Physical fields or potentials often have salient local features: boundary layers, wave-like behavior, peaks (left picture), dipole components (right picture), etc. Numerical accuracy can be improved significantly if such local behavior is taken into account.

4.2.2 Perspective #2: Approximating the *Solution*, Not the Equation

In classic Taylor-based FD schemes, one approximates the underlying differential equation – i.e. the operator and the right hand side. For instance, on a three-point stencil in 1D one can expect a second order approximation of the Poisson equation. There is, however, substantial redundancy built into this approach. Indeed, the scheme covers all sufficiently smooth functions for which the Taylor approximation is valid. Yet it is only the *solution* of the problem that is of direct interest; it is, in a sense, wasteful to approximate other functions.

To highlight this point, imagine for a moment that the exact solution u^* is known. It is then trivial to find a three-point scheme that is itself *exact*, e.g.:

$$\frac{u_{k-1}}{u^*_{k-1}} - 2\,\frac{u_k}{u^*_k} + \frac{u_{k+1}}{u^*_{k+1}} = 0 \tag{4.1}$$

It is easy to dismiss this example as frivolous, as it requires knowledge of the exact solution. The message, however, is that as more information about the

solution is utilized, higher accuracy can be achieved; equation (4.1) is just an extreme example of this principle.

One practical illustration is the use of *harmonic* polynomials to approximate harmonic functions (Sections 4.4.4, 4.4.5). More generally, the "Trefftz" version of FLAME calculus employs basis functions that *satisfy the differential equation* being solved. No effort is wasted on trying to approximate functions that do not satisfy the equation. This "Trefftz" approximation is purely *local* and therefore relatively easy to construct.

4.2.3 Perspective #3: Multivalued Approximation

In FD analysis, interpolation between the nodes is usually viewed just as a postprocessing tool not inherent in the FD method itself. However, approximation between the nodes *is* in fact an integral part of the derivation of classical FD schemes. Indeed, this approximation involves Taylor expansions around grid nodes (Fig. 4.4). Each of these expansions "lives" in a neighborhood of its node. The disparate Taylor expansions *coexist* in the overlap region of two or more such neighborhoods. This is precisely the viewpoint

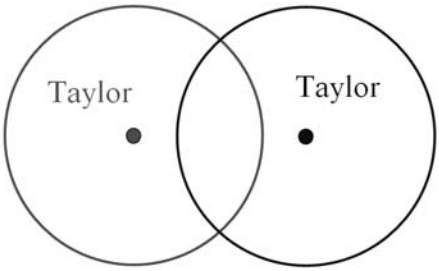


Fig. 4.4. Taylor approximations around two grid nodes coexist in the overlap area.

taken in FLAME, except that any desirable approximating functions are allowed rather than just the Taylor polynomials. Each of these approximations is purely local and valid in the vicinity of a given grid stencil; as in classic FD, two or more such approximations may coexist at any given point. The discrepancies between these approximations are expected to tend to zero if the method converges as the grid is refined. At the same time, these discrepancies may prove useful as an *a posteriori* error indicator in practical computation (J. Dai & I. Tsukerman [DT07]).

4.2.4 Perspective #4: Conformity vs. Flexibility

The following schematic chart (Fig. 4.5) puts various methods into a "flexibility vs. conformity" perspective. "Conformity" is a common jargon term for

(loosely speaking) a sufficient level of smoothness of the solution. More formally, in "fully conforming" methods the numerical solution belongs to the appropriate Sobolev space over the whole computational domain.[2] Various methods shown in the chart are reviewed in Section 4.5 (p. 219).

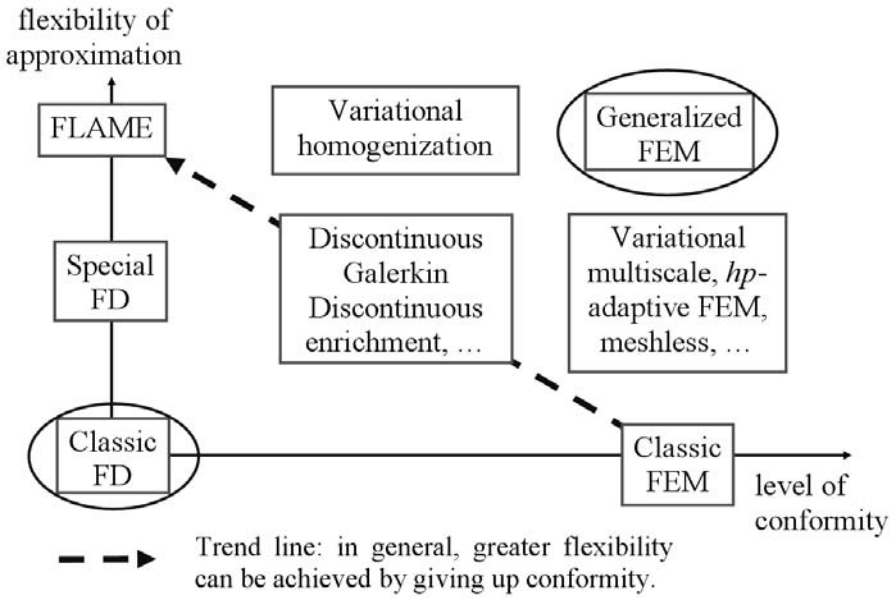


Fig. 4.5. A schematic "conformity vs. flexibility" view of various numerical methods. One can gain flexibility of approximation by giving up conformity. This general trend is indicated by the dashed arrow. GFEM *outperforms* this trend, at a high computational and algorithmic cost. Classic FD schemes *under*perform. FLAME schemes fill the existing void. (Reprinted by permission from [Tsu06] ©2006 Elsevier.)

The dashed arrow in the figure shows the general trend: flexibility of approximation can be gained by giving up some conformity of the method. Two methods stand out of that trend: Generalized FEM (Section 3.15, p. 181) and classic FD (Chapter 2).

GFEM *outperforms* the trend: it is fully conforming (i.e. operating in a globally defined subspace of the relevant Sobolev space) and yet allows any desirable approximating functions to be used. However, this advantage is achieved at a high computational and algorithmic cost. Classic FD schemes

[2] In vector field problems, divergence-conforming and especially curl-conforming spaces $H(\mathrm{div};\Omega)$ and $H(\mathrm{curl};\Omega)$ are widely used; see A. Bossavit's & P. Monk's monographs [Bos98, Mon03].

*under*perform relative to the general trend: they are fully nonconforming and yet make use only of local polynomial (i.e. Taylor) expansions.

FLAME schemes fill the existing void in the upper-left corner of the chart: they are fully nonconforming and admit arbitrary approximations.

Clearly, it would be somewhat simplistic to ask which side of this chart is "better". No one would question the wonderful success of conventional FE analysis lying at the "conformal" end. However, the conformity requirements do impose significant limitations in a number of practical cases. This was understood early on in the development of FEM – hence the notion of "variational crimes" (G. Strang [Str72]), the Crouzeix–Raviart elements (M. Crouzeix & P.A. Raviart [CR73]), etc. The advantages of the *non*conforming end of the spectrum are clear for problems with multiple moving particles, where finite element mesh generation may be inefficient or impractical.

4.2.5 Why Flexible Approximation?

As already noted, in many physical problems some salient features of the solution are qualitatively known *a priori*. Such features include singularities at point sources, edge and corners; boundary layers; derivative jumps at material interfaces; strong dipole field components near polarized spherical particles; cusps of electronic wavefunctions at the nuclei; electrostatic double layers around colloidal particles – and countless other examples. Such "special" behavior of physical fields is arguably a rule rather than an exception. Clearly, taking this behavior into account in numerical simulation will tend to produce more accurate and physically meaningful results.

The special features of the field are typically local, and in numerical modeling it is therefore desirable to employ various *local* approximations of the field. The focus of this chapter is precisely on "Flexible Local Approximation" and on methods capable of providing it – that is, employing a variety of approximating functions not limited to polynomials.

One motivation for developing this class of methods is to minimize the notorious "staircase" effect at curved and slanted interface boundaries on regular Cartesian grids. In the spirit of "Flexible Local Approximation," the behavior of the solution at the interfaces is represented *algebraically*, by suitable basis functions on simple grids, rather than *geometrically* on conforming meshes. More specifically, fields around spherical particles can be approximated by several spherical harmonics; fields scattered from cylinders by Bessel functions, and so on. Such analytical approximations are incorporated directly into the difference scheme.

This approach can be contrasted with very well known, and very powerful, Finite Element (FE) methodology, where the geometric features of the problem are represented on complex conforming meshes. The flexibility of approximation in FEM is achieved through adaptive mesh refinement: changing

the mesh size (h-refinement) or the order of approximation (p-refinement). Still, approximation remains piecewise-polynomial.

FEM is indispensable in many problems where the geometries are complex and material parameters vary. In addition to mechanical, thermal and electromagnetic modeling of traditional devices and machines, FEM has recently penetrated new areas of macromolecular simulation. Molecular interface surfaces can be viewed as intersections of hundreds or thousands of spheres and consequently are geometrically extremely complex. These interfaces separate the interior of the molecule, that can be approximated by an equivalent relative dielectric constant on the order of 1 to 4, from the solvent that in "implicit" models is considered as a continuum with equivalent dielectric and Debye parameters ([BSS+01, GPN01, HN95, CF97, FEVM01, RAH01, Sim03, DTRS07], references therein, and Chapter 6). The computational cost of finite element macromolecular simulation can be enormous. N.A. Baker *et al.* [BSS+01] used a massively parallel supercomputer with 1152 processors to simulate cell structures with 88,000 to 1.25 million atoms; the Poisson–Boltzmann model was used (see Chapter 6).

The computational overhead of mesh generation and matrix assembly in FEM is significant, and for geometrically simple problems FEM may not be competitive with Finite Difference (FD) schemes and other methods operating on simple Cartesian grids. One extreme example of geometric simplicity comes from molecular dynamics simulations, where charges or dipoles are typically considered in a cubic box with periodic boundary conditions. The Ewald algorithm (taking advantage of Fast Fourier Transforms) is then usually the method of choice (Chapter 5).

Problems with multiple moving particles also call for development and application of new techniques. Generation of geometrically conforming FE meshes is obviously quite complicated or impractical when the particles move and their number is large (say, on the order of a hundred or more). Parallel adaptive Generalized FEM has been developed [GS00, GS02a, GS02b], but the procedure is quite complicated both algorithmically and computationally. Standard FD schemes would require unreasonably fine meshes to resolve the shapes of all particles. An alternative approach is to use two types of grid: spherical meshes around the particles and a global Cartesian grid [Fus92, DHM+04]. The electrostatic potential then has to be interpolated back and forth between the grids, which reduces the numerical accuracy.

The celebrated Fast Multipole Method (FMM) has clear advantages for systems with a large number of known charges or dipoles in free space (or a homogeneous medium). For inhomogeneous media (e.g. a dielectric substrate, or finite size particles with dielectric or magnetic parameters different for those of free space) FMM can still be used as a fast matrix-vector multiplication algorithm embedded in an iterative process for the unknown distribution of volume sources. However, the benefits of FMM in this case are much less clear. An even stronger case in favor of difference schemes (as compared to FMM)

can be made if the problem is nonlinear (for example, the Poisson–Boltzmann equation). FMM will remain outside the scope of this chapter.

The proposed new FLAME schemes provide a practical alternative that is both uncomplicated and accurate (Section 4.3). In addition to multiparticle simulations, FLAME techniques can be applied to a variety of other problems. As a peculiar example, super high-order 3-point schemes are derived for the 1D Schrödinger equation in Sections 4.4.6, 4.4.7 and for a 1D singular equation in Section 4.4.8. With the 20th-order 3-point scheme as an illustration, the solution of the harmonic oscillator problem is found almost to machine precision with 10–20 grid nodes. The system matrix remains tridiagonal.

4.2.6 A Preliminary Example: the 1D Laplace Equation

The 1D Laplace equation is trivial and is used here only to provide the simplest possible example of the Trefftz–FLAME schemes. For convenience, consider a uniform grid with size h, choose a 3-point stencil and place the origin at the middle node of the stencil.

The key step in Trefftz–FLAME schemes is to approximate the solution – *locally, over the stencil* – by a linear combination of basis functions satisfying the underlying differential equation. The 1D Laplace equation is so simple that the two independent local solutions

$$\psi_1 \;=\; 1; \qquad \psi_2 \;=\; x$$

also happen to be *global* solutions of the equation (disregarding the boundary conditions), but this circumstance is irrelevant for FLAME. The numerical solution over the stencil is

$$u_h \;=\; c_1\psi_1 \;+\; c_2\psi_2 \tag{4.2}$$

In general, all the variables in this equation may be different for different grid stencils, although for the 1D Laplace equation $c_{1,2}$ happen to be the same throughout the domain. In the future, if there is any possibility of confusion, the stencil number will be indicated with a superscript, but for now it is omitted for simplicity.

We are looking for a difference scheme with some coefficient vector $\underline{s} \equiv (s_1, s_2, s_3)^T \in \mathbb{R}^3$ (s – mnemonic for "scheme") that would relate the nodal values $u_{h1,2,3}$ of the numerical solution on the stencil:

$$s_1 u_{h1} \;+\; s_2 u_{h2} \;+\; s_3 u_{h3} \;=\; 0 \tag{4.3}$$

Since u_h (4.2) contains only two independent parameters ($c_{1,2}$), it is clear that the three nodal values must be linearly related and thus (4.3) must hold for some s. Finding a suitable coefficient vector $\underline{s}$ is easy, and we shall do so in a way that will be straightforward to generalize.

The nodal values that figure in (4.3) are

$$\begin{array}{rclcl} u_{h1} & \equiv & u_h(x_1) & = & c_1\psi_1(x_1) + c_2\psi_2(x_1) \\ u_{h2} & \equiv & u_h(x_2) & = & c_1\psi_1(x_2) + c_2\psi_2(x_2) \\ u_{h3} & \equiv & u_h(x_3) & = & c_1\psi_1(x_3) + c_2\psi_2(x_3) \end{array} \tag{4.4}$$

The matrix-vector form of equations (4.3) and (4.4) is

$$\underline{s}^T \underline{u}_h = 0 \tag{4.5}$$

and

$$\underline{u}_h = N\underline{c} \tag{4.6}$$

where $\underline{u}_h = (u_{h1}, u_{h2}, u_{h3})^T$ is the $\mathbb{R}^3$-vector of nodal values, $\underline{c} = (c_1, c_2)^T$ is the $\mathbb{R}^2$-vector of coefficients, and n is the 3×2 matrix of nodal values of the basis functions:

$$N = \begin{pmatrix} \psi_1(x_1) & \psi_2(x_1) \\ \psi_1(x_2) & \psi_2(x_2) \\ \psi_1(x_3) & \psi_2(x_3) \end{pmatrix} \tag{4.7}$$

Combining (4.5) and (4.6), one obtains

$$\underline{s}^T N\underline{c} = 0 \tag{4.8}$$

For this identity to be valid for any $\underline{c}$, we must have, from basic linear algebra,

$$\underline{s} \in \operatorname{Null} N^T \tag{4.9}$$

Let us spell this out for the 1D Laplace equation. With $\psi_1 = 1$, $\psi_2 = x$ and the coordinates of the nodes $(-h, 0, h)$, the (transposed) nodal matrix (4.7) is

$$N^T = \begin{pmatrix} 1 & 1 & 1 \\ -h & 0 & h \end{pmatrix}$$

The Trefftz–FLAME difference scheme then is

$$\underline{s} = \operatorname{Null} N^T = (1, -2, 1)^T \quad \text{(times an arbitrary coefficient)}$$

which coincides with the standard 3-point scheme for the Laplace equation.

In the remainder of this chapter, we shall see that the definition (4.9) of the scheme has a great deal of generality and is applicable to a variety of equations (Section 4.3). First, however, we need to discuss a general setup for local, finite-difference-like, approximation.

4.3 Trefftz Schemes with Flexible Local Approximation

4.3.1 Overlapping Patches

An important element of the setup, to be used in the remainder of this chapter, is a set of overlapping patches $\Omega^{(i)}$ covering the computational domain $\Omega =$

$\cup\Omega^{(i)}$, $i = 1, 2, \ldots n$. This cover of the domain is the same as in Generalized FEM (see Sections 3.15, p. 181, and 4.5.2, p. 221); however, FLAME differs from GFEM in many critical respects as we shall see.

The domain cover is needed to define a local, patch-wise, approximation of the solution. More precisely, within each patch $\Omega^{(i)}$ we introduce a local approximation space

$$\Psi^{(i)} = \text{span}\{\psi_\alpha^{(i)}, \quad \alpha = 1, 2, \ldots, m(i)\} \tag{4.10}$$

Note that no *global* approximation space will be considered. Instead, the following notion of *multivalued approximation* is introduced:

For a given domain cover $\{\cup\Omega^{(i)}\}$ with corresponding local spaces $\Psi^{(i)}$, a multivalued approximation $u_h\{\cup\Omega^{(i)}\}$ of a given potential u is just a collection of patch-wise approximations:

$$u_h\{\cup\Omega^{(i)}\} \;\equiv\; \{u_h^{(i)} \in \Psi^{(i)}\} \tag{4.11}$$

In regions where two or more patches overlap (Fig. 4.6), several local approximations coexist and do not have to be the same. This situation in fact is inherent in the FD methodology but is almost never stated explicitly.[3]

The second ingredient of FLAME is a set of n nodes (the number of nodes is equal to the number of patches). Although a meshless setup is possible, we shall for maximum simplicity assume a regular grid with a mesh size h. The i-th stencil is defined as a set of $m(i)$ nodes within $\Omega^{(i)}$: stencil $\#i \equiv \{\text{nodes} \in \Omega^{(i)}\}$. For any continuous potential u, $\mathcal{N}u$ will denote the set of its values at all grid nodes (viewed as a Euclidean vector in $\mathbb{R}^n$), and $\mathcal{N}^{(i)}u$ – the set of nodal values on stencil $\#i$. Although the FLAME solution may be multivalued between the nodes, its values *at the nodes* are required to be unique.

Within each patch, the approximate solution $u_h^{(i)}$ is sought as a linear combination of $m(i)$ basis functions $\{\psi_\alpha^{(i)}\}$:

$$u_h^{(i)} = \sum\nolimits_{\alpha=1}^{m(i)} c_\alpha^{(i)} \psi_\alpha^{(i)} \tag{4.12}$$

Here we are following the same line of reasoning as in the preliminary example of Section 4.2.6 on p. 197, but in a more general setting. We need to relate the coefficient vector $\underline{c}^{(i)} \equiv \{c_\alpha^{(i)}\} \in \mathbb{R}^m$ of expansion (4.12) to the vector $\underline{u}^{(i)} \in \mathbb{R}^M$ of the nodal values of $u_h^{(i)}$ on stencil $\#i$. (Both M and m can be different for different patches (i); this is understood but not explicitly indicated for simplicity of notation.) The relevant transformation matrix $N^{(i)}$,

[3] One might argue that in FD methods approximation between the grid nodes is not multivalued but simply undefined. This point of view is not incorrect but ignores the fact that the very derivation of FD schemes typically relies upon disparate Taylor expansions in the neighborhoods of each grid point.

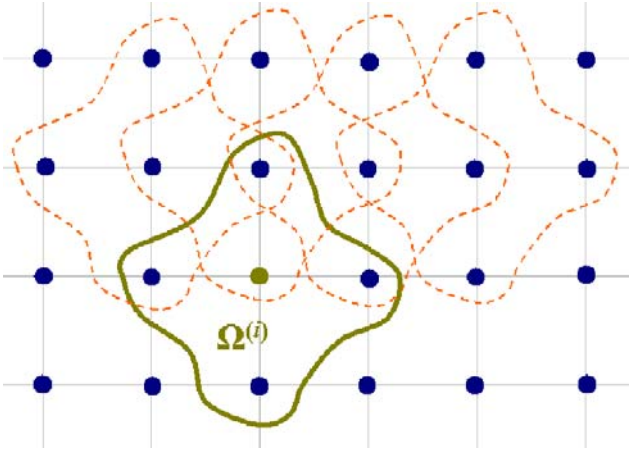


Fig. 4.6. Overlapping patches with 5-point stencils. (Reprinted by permission from [Tsu04b] ©2004 IEEE.)

$$\underline{u}^{(i)} = N^{(i)}\underline{c}^{(i)} \tag{4.13}$$

contains the nodal values of the basis functions on the stencil; if r_k is the position vector of node k, then

$$N^{(i)} = \begin{pmatrix} \psi_1^{(i)}(r_1) & \psi_2^{(i)}(r_1) & \dots & \psi_m^{(i)}(r_1) \\ \psi_1^{(i)}(r_2) & \psi_2^{(i)}(r_2) & \dots & \psi_m^{(i)}(r_2) \\ \dots & \dots & \dots & \dots \\ \psi_1^{(i)}(r_M) & \psi_2^{(i)}(r_M) & \dots & \psi_m^{(i)}(r_M) \end{pmatrix} \tag{4.14}$$

4.3.2 Construction of the Schemes

In the remainder, except for Appendix 4.7.3, the focus will be on the *Trefftz* version of FLAME, where the approximating functions $\psi^{(i)}$ *satisfy the underlying differential equation* (4.15) *exactly.* Trefftz methods are well known in the variational context (I. Herrera [Her00]); in contrast, here a purely *finite-difference* approach is taken and will prove to be attractive in a variety of cases.[4] Trefftz–FLAME is simpler and at the same time usually more effective than the more general variational version of FLAME considered in Appendix 4.7.3 on p. 232.

Since the basis functions by construction already satisfy the underlying differential equation, so does the approximate solution $u_h^{(i)}$, automatically. As we shall see, there will typically be fewer approximating functions than

[4] The starting point for this development of Trefftz–FLAME schemes was Gary Friedman's non-variational version of FLAME for unbounded problems [Fri05], [HFT04].

nodes within the patch – most frequently, m functions for $M = m + 1$ stencil nodes. The nodal matrix $N^{(i)}$ is thus in general rectangular.[5] The number of approximating functions may be different for different patches, but for brevity of notation this is not explicitly indicated.

Let us initially assume that the underlying differential equation within a patch $\Omega^{(i)}$ has a zero right hand side:

$$Lu = 0 \quad \text{in } \Omega^{(i)} \tag{4.15}$$

where L is a differential operator (one may want to have in mind, say, the Laplace operator as one of the simplest examples).

Within each patch, the approximate solution $u_h^{(i)}$ is sought as a linear combination (4.12) of $m(i)$ basis functions $\{\psi_\alpha^{(i)}\}$. Identity (4.13) relates the vector of coefficients $\underline{c}^{(i)}$ to the nodal values:

$$N^{(i)}\underline{c}^{(i)} \quad = \quad \underline{u}^{(i)} \tag{4.16}$$

In the simplest 1D example, with $m = 2$ basis functions $\psi_{1,2}$ at three grid points x_{i-1}, x_i, x_{i+1}, matrix $N^{(i)}$ (4.14) is

$$N^{(i)} \quad = \quad \begin{pmatrix} \psi_1(x_{i-1}) & \psi_2(x_{i-1}) \\ \psi_1(x_i) & \psi_2(x_i) \\ \psi_1(x_{i+1}) & \psi_2(x_{i+1}) \end{pmatrix} \tag{4.17}$$

We have already seen this for the 1D Laplace equation and the three-point stencil in Section 4.2.6. More generally for an M-point stencil, a vector of coefficients $\underline{s}^{(i)} \in \mathbb{R}^M$ of the difference scheme is sought to yield

$$\underline{s}^{(i)T}\underline{u}^{(i)} = 0 \tag{4.18}$$

for the nodal values $\underline{u}^{(i)}$ of *any* function $u_h^{(i)}$ of form (4.12). Due to (4.13) and (4.18),

$$\underline{s}^{(i)T}N^{(i)}\underline{c}^{(i)} = 0 \tag{4.19}$$

For this to hold for any set of coefficients $\underline{c}^{(i)}$, the null-space condition already familiar to us must hold:

$$\underline{s}^{(i)} \in \operatorname{Null} N^{(i)T} \tag{4.20}$$

If the null space is of dimension one, $\underline{s}^{(i)}$ represents the desired scheme (up to an arbitrary factor), and (4.20) is the principal expression of this Trefftz–FLAME scheme. The meaning of (4.20) is simple: each equation in the system $N^{(i)T}\underline{s}^{(i)} = 0$ implies that the respective basis function satisfies the difference equation with coefficients $\underline{s}^{(i)}$. There is thus an elegant duality feature between the continuous and discrete problems: any linear combination of the basis

[5] However, in the *variational*-difference formulation (Appendix 4.7.3), the number of basis functions is typically equal to the number of nodes.

functions satisfies both the differential equation (due to the choice of the "Trefftz" basis) *and* the difference equation with coefficients $\underline{s}^{(i)}$.

An alternative interpretation of (4.20) is that $\underline{s}^{(i)}$ is orthogonal to the image of $N^{(i)}$ due to (4.19), hence $\underline{s}^{(i)}$ is in the null space of $N^{(i)T}$. In the complex case, though, orthogonality should not be understood in terms of the standard complex inner product which, unlike (4.19), includes conjugates.

While there is no obvious way to determine the dimension of the null space *a priori*, for several classes of problems considered later the dimension is indeed one. If the null space is empty, the construction of the Trefftz–FLAME scheme fails, and one may want to either increase the size of the stencil or reduce the basis set. If the dimension of the null space is greater than one, there are two general options. First, the stencil and/or the basis can be changed. Second, one may use the additional freedom in the choice of the coefficients $\underline{s}^{(i)}$ to seek an "optimal" (in some sense) scheme as a linear combination of the independent null space vectors. For example, it may be desirable to find a diagonally dominant scheme.

Once the basis and the stencil are chosen, the Trefftz–FLAME scheme is generated in a very simple way:

- Form matrix $N^{(i)}$ of the nodal values of the basis functions.
- Find the null space of $N^{(i)T}$.

Proposition 11. *The Trefftz–FLAME scheme defined by (4.20) is invariant with respect to the choice of the basis in the local space* $\Psi^{(i)} \equiv \mathrm{span}\{\psi_\alpha^{(i)}\}$.

Proof. A linear transformation of the ψ-basis replaces N^T with QN^T, where Q is a nonsingular matrix, which does not affect the null space. □

The algorithm can be sketched as a "machine" for generating Trefftz–FLAME schemes (Fig. 4.7).

It should be stressed that the algorithm is heuristic and no blanket claim of convergence can be made. The schemes need to be considered on a case-by-case basis, which is done for a variety of problems in Section 4.3. However, consistency can be proven (Section 4.3.5) in general, and convergence then follows for the subclass of schemes with a monotone difference operator [Tsu05a].

As we shall see in Section 4.4, definition (4.20), despite its simplicity, is surprisingly rich. For different choices of basis functions and stencils it gives rise to a variety of difference schemes.

4.3.3 The Treatment of Boundary Conditions

Note that in the FLAME framework approximations over different stencils are completely independent from one another. Therefore, if the domain boundary conditions are of standard types and no special behavior of the solution at

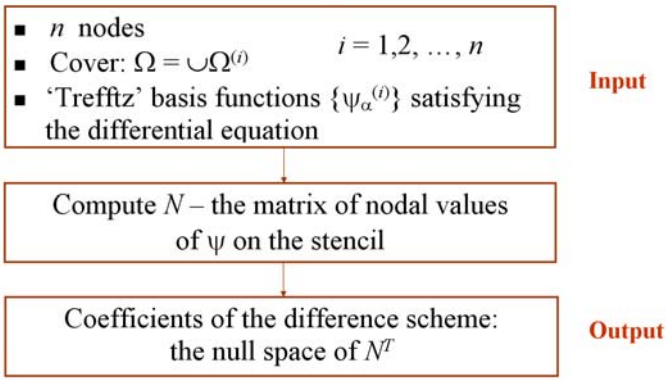


Fig. 4.7. A "machine" for Trefftz–FLAME schemes. (Reprinted by permission from [Tsu05a] ©2005 IEEE.)

the boundaries is manifest, one can simply employ any standard FD scheme at the boundary.[6]

If the solution is known to exhibit some special features at the boundary, it may be possible to incorporate these features into FLAME. One example – Perfectly Matched Layers (PML) for electromagnetic and acoustic wave propagation – is considered briefly in Section 4.4.11 and in [Tsu05a].

4.3.4 Trefftz–FLAME Schemes for Inhomogeneous and Nonlinear Equations

So far we considered Trefftz–FLAME schemes only for homogeneous equations (i.e. with the zero right hand side within a given patch). For inhomogeneous equations of the form

$$Lu = f \quad \text{in } \Omega^{(i)} \tag{4.21}$$

a natural approach is to split the solution up into a particular solution $u_f^{(i)}$ of the inhomogeneous equation and the remainder $u_0^{(i)}$ satisfying the homogeneous one:

$$u = u_0^{(i)} + u_f^{(i)} \tag{4.22}$$

$$Lu_0^{(i)} = 0; \qquad Lu_f^{(i)} = f \tag{4.23}$$

Superscript (i) emphasizes that the splitting is *local*, i.e. needs to be introduced only within its respective patch $\Omega^{(i)}$ containing the grid stencil around node

[6] Since most Taylor-based schemes are particular cases of FLAME (with polynomial basis functions), it would be technically correct to say that the whole set of difference equations, including the treatment of boundary conditions, is based on FLAME.

i. Since $u_f^{(i)}$ is local (and in particular need not satisfy any exterior boundary conditions), it is usually relatively easy to construct.

Let a Trefftz–FLAME scheme $\underline{s}^{(i)}$ be generated for a given set of basis functions and assume that the consistency error ϵ for this scheme tends to zero as $h \to 0$; that is,

$$\underline{s}^{(i)T}\mathcal{N}^{(i)}u_0^{(i)} \;=\; \epsilon \;\equiv\; \epsilon\left(h,\, u_0^{(i)}\right) \to 0 \;\text{ as grid size } h \to 0 \tag{4.24}$$

where $\mathcal{N}^{(i)}$, as before, denotes the nodal values of a function on stencil (i). Then clearly

$$\underline{s}^{(i)T}\mathcal{N}^{(i)}u \;=\; \underline{s}^{(i)T}\mathcal{N}^{(i)}u_0 \;+\; \underline{s}^{(i)T}\mathcal{N}^{(i)}u_f \;=\; \underline{s}^{(i)T}\mathcal{N}^{(i)}u_f \;+\; \epsilon$$

This immediately implies that the consistency error of the difference scheme

$$\underline{s}^{(i)T}\underline{u}_h \;=\; \underline{s}^{(i)T}\mathcal{N}^{(i)}u_f \tag{4.25}$$

is ϵ, i.e. exactly the same as for the homogeneous case. (The Euclidean vector $\underline{u}_h$ of nodal values does not need the superscript because the nodal values are unique and do not depend on the patch.) Note that there are absolutely no constraints on the smoothness of $u_f^{(i)}$, provided that it has valid nodal values. The particular solution $u_f^{(i)}$ can even be singular as long as the singularity point does not coincide with a grid node. In [Tsu04a] difference schemes of this kind were constructed for the Coulomb potential of point charges. An electrostatic problem with a line charge source is solved in a similar way in [Tsu05a].

For *nonlinear* problems, the Newton–Raphson method is traditionally used for the *discrete* system of equations. In connection with FLAME schemes, Newton–Raphson–Kantorovich iterations are applied to the original continuous problem rather than the discrete one. Let the equation be

$$Lu \;=\; f \tag{4.26}$$

where L is a differentiable operator. The $(k+1)$-th approximation u_{k+1} to the exact solution is obtained from the k-th approximation u_k by linearization in the following way. If $u = u_k + \delta u$,

$$Lu \;=\; L(u_k + \delta u) \;=\; Lu_k \;+\; L'(u_k)\delta u + o(\|\delta u\|) \tag{4.27}$$

where L' is the Fréchet derivative of L (Appendix 4.9). Ignoring higher-order terms, one gets an approximation δu_k for δu by solving the linear system

$$L'(u_k)\,\delta u_k \;=\; f \;-\; Lu_k \tag{4.28}$$

and then updates the solution:

$$u_{k+1} \;=\; u_k \;+\; \delta u_k \tag{4.29}$$

Equivalently,

$$u_{k+1} \;=\; u_k \;+\; (L'(u_k))^{-1}(f \;-\; Lu_k) \tag{4.30}$$

Along with an initial guess u_0, iterative process (4.28), (4.29) – or just (4.30) – defines the Newton–Raphson–Kantorovich algorithm. Trefftz–FLAME schemes can then be applied to L' (which of course is a linear operator by definition), provided that a suitable set of local approximating functions can be found.

Further analysis of the N–R–K iterations for FLAME schemes in colloidal simulation (the Poisson–Boltzmann equation) can be found in Section 6.8 on p. 319.

4.3.5 Consistency and Convergence of the Schemes

Let us rewrite the patch-wise difference equation (4.25) in matrix form as a global system of difference equations for the underlying differential equation $Lu \;=\; f$:

$$L_h \underline{u}_h \;=\; \underline{f}_h, \quad \text{with } \underline{f}_{hi} \;=\; \underline{s}^{(i)T} \mathcal{N}^{(i)} u_f^{(i)} \tag{4.31}$$

(if the differential equation is homogeneous within the patch, then $u_f^{(i)} = 0$). Note that the i-th row of matrix L_h contains the coefficients of scheme $\underline{s}^{(i)T}$ and, in addition, a (large) number of zero entries.[7] We shall assume that the equations can be scaled in such a way that

$$c_1 f(r) \;\le\; \underline{f}_{hi} \;\le\; c_2 f(r), \quad \forall r \in \Omega^{(i)}, \quad c_{1,2} > 0 \tag{4.32}$$

where $c_{1,2}$ do not depend on i and h. This scaling is important because otherwise e.g. the meaningless scheme $h^{100} u_i = 0$ would technically be consistent (as defined below) for *any* differential equation.

The consistency error of scheme (4.31) is, by definition, obtained by substituting the nodal values of the exact solution u^* into the difference equation. We shall call this scheme consistent if, with scaling (4.32), the following condition holds:

$$\begin{aligned} \text{consistency error} \;\equiv\; \epsilon_c(h) \;&=\; \max_i \left| \underline{s}^{(i)T} \mathcal{N}^{(i)} u^* \;-\; \underline{f}_{hi} \right| \\ &=\; \max_i \left| \underline{s}^{(i)T} (\mathcal{N}^{(i)} u^* \;-\; \underline{u}_{hi}) \right| \;\to\; 0 \;\text{ as } h \to 0 \end{aligned} \tag{4.33}$$

For FLAME schemes, consistency follows directly from the approximation properties of the basis set as long as (4.32) holds. Indeed, let $\epsilon_a(h)$ be the approximation error of the "homogeneous part" $u_0^{(i)}$ of the exact solution u^* in a patch $\Omega^{(i)}$:

[7] Our notation would perhaps be more consistent if the matrix were denoted with L_h and the scheme with $l^{(i)}$ or, alternatively, if the scheme were $\underline{s}^{(i)}$ and the matrix were S_h. However, throughout the book the usual symbol L is adopted for differential and difference operators, and s is used as a mnemonic symbol for "scheme".

$$\epsilon_a(h) \;=\; \min_{c^{(i)}\in\mathbb{R}^m} \left\| u^* - u_f^{(i)} - \sum\nolimits_{\alpha=1}^{m} c_\alpha^{(i)}\psi_\alpha^{(i)} \right\|_\infty \tag{4.34}$$

Equivalently, there exists a coefficient vector $\underline{c}^{(i)} \in \mathbb{R}^m$ such that

$$u^* \;=\; u_f^{(i)} \;+\; \sum\nolimits_{\alpha=1}^{m} c_\alpha^{(i)}\psi_\alpha^{(i)} + \eta, \qquad \|\eta\|_\infty \;=\; \epsilon_a(h) \tag{4.35}$$

For the nodal values, one then has due to (4.16)

$$\mathcal{N}^{(i)}u^* \;=\; \mathcal{N}^{(i)}u_f^{(i)} \;+\; N^{(i)}\underline{c}^{(i)} \;+\; \underline{\eta} \tag{4.36}$$

where $\underline{\eta} = \mathcal{N}^{(i)}\eta$ is the vector of nodal values of η on stencil i and $N^{(i)}$ is (as always) the matrix of nodal values of the basis functions. Due to (4.35),

$$\|\underline{\eta}\|_\infty \;\le\; \epsilon_a(h)$$

and due to (4.36), the consistency error for scheme (4.31) with coefficients (4.20) is

$$|\epsilon_c(h)| \;=\; \max_i \left| \underline{s}^{(i)T}\mathcal{N}^{(i)}u^* \;-\; \underline{s}^{(i)T}\mathcal{N}^{(i)}u_f^{(i)} \right| \;=\; \max_i \left| \underline{s}^{(i)T}(N^{(i)}\underline{c}^{(i)} \;+\; \underline{\eta}) \right|$$

$$= \max_i \left| \underline{s}^{(i)T}N^{(i)}\underline{c}^{(i)} \;+\; \underline{s}^{(i)T}\underline{\eta} \right| \;=\; \max_i \left| \underline{s}^{(i)T}\underline{\eta} \right| \;\le\; M\epsilon_a(h) \tag{4.37}$$

which shows that the consistency error is bounded by the approximation error.

Theoretical results relevant to the convergence of the schemes were summarized in Chapter 2. Estimate (2.148) (p. 62) of the solution error is the ratio of approximation and stability parameters. The approximation accuracy ϵ_a is key. In fact, the "Trefftz" bases are effective not just because they (by definition) satisfy the underlying differential equation, but because they happen to have superior approximation properties in many cases (see e.g. Sections 4.4.4, 4.4.5).

4.4 Trefftz–FLAME Schemes: Case Studies

4.4.1 1D Laplace, Helmholtz and Convection-Diffusion Equations

The 1D Laplace equation was already considered as a preliminary example in Section 4.2.6 of this chapter (p. 197). A less trivial case is the 1D Helmholtz equation

$$\frac{d^2u}{dx^2} \;-\; \kappa^2 u \;=\; 0$$

with any complex κ. Two basis functions satisfying the Helmholtz equation are

$$\psi_1 \;=\; \exp(\kappa x); \;\; \psi_2 \;=\; \exp(-\kappa x)$$

For a three-point stencil with the coordinates of the nodes $(-h, 0, h)$ (the middle node is placed at the origin for simplicity), the matrix of nodal values (4.14) is

$$N^T = \begin{pmatrix} \exp(-\kappa h) & 1 & \exp(\kappa h) \\ \exp(\kappa h) & 1 & \exp(-\kappa h) \end{pmatrix}$$

and the resultant difference scheme is

$$\underline{s} = \operatorname{Null} N^T = (1, \ -2\cosh(\kappa h), \ 1)^T \tag{4.38}$$

Since the theoretical solution in this 1D case is exactly representable as a linear combination of the chosen basis functions, the difference scheme yields the exact solution (in practice, up to the round-off error). This scheme is known and has been derived in a different way by R.E. Mickens [Mic94]; see also C. Farhat *et al.* [FHF01] and I. Harari & E. Turkel [HT95].

Quite similarly, for the 1D convection-diffusion equation

$$D\,\frac{d^2u}{dx^2} - b\,\frac{du}{dx} = 0, \quad D > 0$$

with constant coefficients D and b, one has two Trefftz basis functions:

$$\psi_1 = 1; \ \ \psi_2 = \exp(qx), \quad q = b/D$$

For the 3-point stencil $(-h, 0, h)$, the (transposed) matrix of nodal values (4.14) is

$$N^T = \begin{pmatrix} 1 & 1 & 1 \\ \exp(-qh) & 1 & \exp(qh) \end{pmatrix}$$

and the Trefftz–FLAME difference scheme is

$$\underline{s} = \operatorname{Null} N^T = \left[\frac{\exp(qh)}{\exp(qh)-1}, \ -\frac{\exp(qh)+1}{\exp(qh)-1}, \ \frac{1}{\exp(qh)-1} \right] \tag{4.39}$$

(up to an arbitrary factor). This coincides (in the case of the homogeneous convection-diffusion equation with constant coefficients) with the well-known exponentially fitted scheme (see e.g. D.B. Spalding [Spa72], G.D. Raithby & K.E. Torrance [RT74], S.V. Patankar [Pat80]).

4.4.2 The 1D Heat Equation with Variable Material Parameter

Consider the 1D homogeneous heat conduction equation:

$$\frac{d}{dx}\left(\lambda(x)\frac{du}{dx}\right) = 0 \tag{4.40}$$

where $\lambda(x)$ is the material parameter. Two approximating functions for the FLAME-Trefftz scheme can be chosen as linearly independent solutions of this equation on the interval $[x_{k-1}, x_{k+1}]$:

$$\psi_1 = 1, \quad \psi_2 = \int_{x_k}^{x} \lambda^{-1}(\xi) d\xi$$

With this basis, the transposed nodal matrix (4.14) for the stencil (x_{k-1}, x_k, x_{k+1}) is

$$N^T = \begin{pmatrix} 1 & 1 & 1 \\ -\Sigma_{k-1} & 0 & \Sigma_{k+1} \end{pmatrix}$$

where $\Sigma_{k-1} = \int_{x_{k-1}}^{x_k} \lambda^{-1}(\xi)\, d\xi$, $\Sigma_{k+1} = \int_{x_k}^{x_{k+1}} \lambda^{-1}(\xi)\, d\xi$ have the physical meaning of thermal resistances of the respective segments. The difference scheme is, up to an arbitrary factor,

$$\underline{s} = \operatorname{Null} N^T = (-\Sigma_{k-1}^{-1}, \ \Sigma_{k-1}^{-1} + \Sigma_{k+1}^{-1}, \ -\Sigma_{k+1}^{-1})^T \tag{4.41}$$

which has a clear interpretation as a flux balance equation:

$$\Sigma_{k-1}^{-1}(u_k - u_{k-1}) + \Sigma_{k+1}^{-1}(u_k - u_{k+1}) = 0$$

Such schemes are indeed typically derived from flux balance considerations (see e.g. the "homogeneous schemes" in [Sam01]) but, as we can now see, emerge as a natural particular case of Trefftz–FLAME.

If the integrals in the expressions for thermal resistances Σ can be calculated exactly, the scheme is itself exact, i.e. the consistency error is zero (the theoretical solution satisfies the FD equation). This holds even if the material parameter λ is discontinuous. A very similar analysis applies to the 1D linear electrostatic equation with a variable (and possibly discontinuous) permittivity ϵ.

4.4.3 The 2D and 3D Laplace Equation

Consider a regular rectangular grid, for simplicity with spacing h the same in both directions, and the standard 5-point stencil. The origin of the coordinate system is placed for convenience at the central node of the stencil. With four basis functions $[1, x, y, x^2 - y^2]$ satisfying the Laplace equation, the nodal matrix (4.14) becomes

$$N^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & -h & 0 & h & 0 \\ h & 0 & 0 & 0 & -h \\ -h^2 & h^2 & 0 & h^2 & -h^2 \end{pmatrix}$$

The difference scheme is then $\operatorname{Null} N^T = (-1, -1, 4, -1, -1)^T$ (times an arbitrary constant), which coincides with the standard 5-point scheme for the Laplace equation. A more general case with different mesh sizes in the x- and y- directions is handled similarly.

The 3D case is also fully analogous. With six basis functions $\{1, x, y, z, x^2 - y^2, x^2 - z^2\}$ and the standard 7-point stencil on a uniform grid, one

arrives, after computing the null space of the respective 6×7 matrix N^T, at the standard 7-point scheme with the coefficients $(-1, -1, -1, 6, -1, -1, -1)^T$. As in 2D, the case of different mesh sizes in the x-, y- and z-directions does not present any difficulty.

4.4.4 The Fourth Order 9-point Mehrstellen Scheme for the Laplace Equation in 2D

The solution is, by definition, a harmonic function. Harmonic polynomials are known to provide an excellent (in some sense, even optimal [BM97]) approximation of harmonic functions [And87, BM97, Ber66, Mel99]. Indeed, for a fixed polynomial order p, the FEM and harmonic approximation errors are similar [BM97]; however, the FEM approximation is realized in a much wider space containing *all* polynomials up to order p, not just the harmonic ones. For solving the Laplace equation, the standard FE basis set can thus be viewed as having substantial redundancy that is eliminated by using the harmonic basis. The following result is cited in [BM97]:

Theorem 6. *(Szegö). Let $\Omega \subset \mathbb{R}^2$ be a simply connected bounded Lipschitz domain. Let $\tilde{\Omega} \supset\supset \Omega$ and assume that $u \in L^2(\tilde{\Omega})$ is harmonic on $\tilde{\Omega}$. Then there is a sequence $(u_p)_{p=0}^{\infty}$ of harmonic polynomials of degree p such that*

$$\begin{aligned} \|u - u_p\|_{L^\infty(\Omega)} &\leq c \exp(-\gamma p) \|u\|_{L^2(\tilde{\Omega})} \\ \|\nabla(u - u_p)\|_{L^\infty(\Omega)} &\leq c \exp(-\gamma p) \|u\|_{L^2(\tilde{\Omega})} \end{aligned} \tag{4.42}$$

where γ, $c > 0$ depend only on Ω, $\tilde{\Omega}$.

For comparison, the H^1-norm error estimate in the standard FEM is

Theorem 7. *(P.G. Ciarlet & P.A. Raviart, I. Babuska & M. Suri [CR72], [Cia80], [BS94]). For a family of quasiuniform meshes with elements of order p and maximum diameter h, the approximation error in the corresponding finite element space V^n is*

$$\inf_{v \in V^n} \|u - v\|_{H^1(\Omega)} = Ch^{\mu-1} p^{-(k-1)} \|u\|_{H^k(\Omega)}$$

where $\mu = \min(p+1, k)$ and c is a constant independent of h, p, and u.

For a fixed polynomial order p, the FEM and harmonic polynomial estimates are similar (factor $\mathcal{O}(h^p)$ vs. $\mathcal{O}([\exp(-\gamma)]^p)$ if the solution is sufficiently smooth. However, the FEM approximation is realized in a much wider space containing *all* polynomials up to order p, not just the harmonic ones. For solving the Laplace equation, the standard FE basis set can thus be viewed as having substantial redundancy that is eliminated by using the harmonic basis.

With these observations in mind, one may choose the basis functions as harmonic polynomials in x, y up to order 4, namely, $\{1, x, y, xy, x^2-y^2, x(x^2-3y^2), y(3x^2-y^2), (x^2-y^2)xy, (x^2-2xy-y^2)(x^2+2xy-y^2)\}$. Then for a 3×3 stencil of adjacent nodes of a uniform Cartesian grid, the computation of the nodal matrix (4.14) (transposed) and its null space is simple with any symbolic algebra package. If the mesh size is equal in both x- and y- directions, the resultant scheme has order 6. Its coefficients are 20 for the central node, -4 for the four mid-edge nodes, and -1 for the four corner nodes of the stencil. In the standard texts (L. Collatz [Col66], A.A. Samarskii [Sam01]), this scheme is derived by manipulating the Taylor expansions for the solution and its derivatives.

4.4.5 The Fourth Order 19-point Mehrstellen Scheme for the Laplace Equation in 3D

Construction of the scheme is analogous to the 2D case. The 19-point stencil is obtained by considering a $3 \times 3 \times 3$ cluster of adjacent nodes and then discarding the eight corner nodes. The basis functions are chosen as the 25 independent harmonic polynomials in x, y, z up to order 4. Computation of the matrix of nodal values (4.14) and of the null space of its transpose is straightforward by symbolic algebra.

The result is the 19-point fourth-order "Mehrstellen" scheme by L. Collatz [Col66] (see also A.A. Samarskii [Sam01]) already discussed in Chapter 2 (Section 2.8.5, p. 58). In that chapter, as well as in the Collatz and Samarskii books, the scheme is derived from completely different considerations.[8] We can now see, however, that in the Trefftz–FLAME framework Mehrstellen schemes and classic Taylor-based schemes for the Laplace equation stem from the same root – namely, the nullspace equation (4.20). The scheme is defined by the chosen stencil and a harmonic polynomial basis.

As a side note, the 19-point Mehrstellen scheme, due to its geometrically compact stencil, reduces processor communication in parallel solvers and therefore has gained popularity in computationally intensive applications of physical chemistry and quantum chemistry: electrostatic fields of multiple charges, the Poisson–Boltzmann equation in colloidal and protein simulation, and the Kohn–Sham equation of Density Functional Theory (E.L. Briggs *et al.* [BSB96]).

4.4.6 The 1D Schrödinger Equation. FLAME Schemes by Variation of Parameters

This test problem is borrowed from the comparison study by R. Chen *et al.* [CXS93] of several FD schemes for the boundary value (rather than eigenvalue) problem for the 1D Schrödinger equation over a given interval $[a, b]$:

[8] A generalization of the Mehrstellen schemes, known as the HODIE schemes by R.E. Lynch & J.R. Rice [LR80], will not be considered here.

$$-u'' + (V(x) - E)u \;=\; 0, \quad u(a) = u_a, \quad u(b) = u_b \tag{4.43}$$

The specific numerical example is the 5th energy level of the harmonic oscillator, with $V(x) = x^2$ and $E = 11$ $(= 2 \times 5 + 1)$. For testing and verification, boundary conditions are taken from the analytical solution, and as in [CXS93] the interval $[a, b]$ is $[-2, 2]$. The exact solution is

$$u_{\text{exact}} \;=\; (15x - 20x^3 + 4x^5)\exp(-x^2/2) \tag{4.44}$$

To construct a Trefftz–FLAME scheme for (4.43) on a stencil $[x_{i-1},\, x_i,\, x_{i+1}]$ (where $x_{i\pm1} = x_i \pm h$), one would need to take two independent local solutions of the Schrödinger equation as the FLAME basis functions. The exact solution in our example is reserved exclusively for verification and error analysis. We shall construct Trefftz–FLAME scheme *pretending* that the theoretical solution is not known, as would be the case in general for an arbitrary potential $V(x)$.

Thus in lieu of the exact solutions the basis set will contain their approximations. There are at least two ways to construct such approximations. This subsection uses a perturbation technique that produces a fourth-order scheme. The next subsection employs the Taylor expansion that leads to 3-point schemes of arbitrarily high order.

At an arbitrary point x_0 let

$$V(x) \;=\; \kappa^2 \;+\; \delta V, \quad \text{where } \kappa^2 \;\equiv\; V(x_0) \tag{4.45}$$

$$u(x) \;=\; u_0(x) \;+\; \delta u(x) \tag{4.46}$$

$$u_0(x) \;=\; c_+ \exp(\kappa x) \;+\; c_- \exp(-\kappa x), \quad \text{with arbitrary } c_+, c_- \tag{4.47}$$

Substituting these expressions into the Schrödinger equation and ignoring the higher order term, one gets the perturbation equation

$$\delta u'' \;-\; \kappa^2 \delta u \;=\; \delta V u_0 \tag{4.48}$$

Solving this equation by variation of parameters, one obtains after some algebra

$$u(x) \;=\; u_0(x) \;+\; \delta u(x) \;=\; u_0(x) \;+\; \frac{1}{2}\exp(\kappa x)\int_{x_0}^{x} u_0(\xi)\exp(-\kappa\xi)\delta V(\xi)d\xi$$
$$-\frac{1}{2}\exp(-\kappa x)\int_{x_0}^{x} u_0(\xi)\exp(\kappa\xi)\delta V(\xi)d\xi \tag{4.49}$$

Two independent sets of values for c_+, c_- then yield two basis functions for FLAME.

Fig. 4.8 compares convergence of several schemes: the well-known Numerov scheme, the "Numerov–Mickens scheme" [CXS93], Trefftz–FLAME, and the Mickens scheme [Mic94, CXS93]. The first three schemes are all of order four, but the FLAME errors are much smaller. In the following section, the FLAME error is further reduced, in many cases to machine precision.

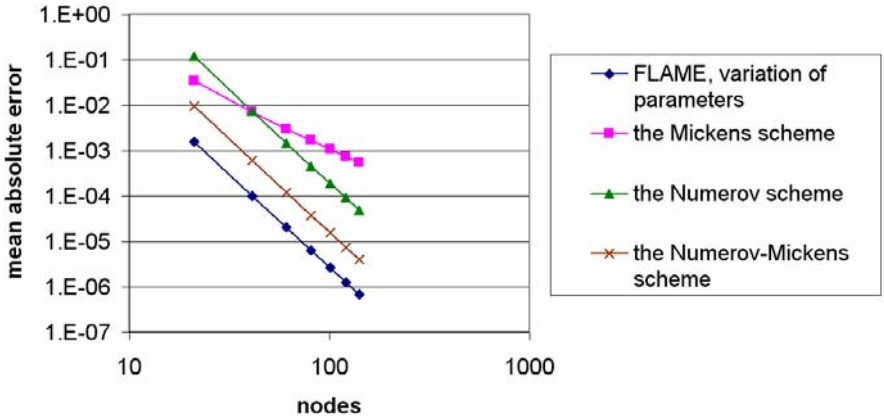


Fig. 4.8. Convergence of the variation of parameters – FLAME scheme for the Schrödinger equation. Comparison with other schemes described in [CXS93] is very favorable (note the logarithmic scale). As the Numerov and Numerov-Mickens schemes, the FLAME scheme is of fourth order but its error is much smaller. The Taylor version of FLAME (see below) performs much better still. (Reprinted by permission from [Tsu06] ©2006 Elsevier.)

4.4.7 Super-high-order FLAME Schemes for the 1D Schrödinger Equation

For sufficiently smooth potentials $V(x)$, as in our example of the harmonic oscillator, one can expand the potential and the solution into a Taylor series around the central stencil node x_i to obtain two local independent solutions with any desired order of accuracy. Consequently, the order of the FLAME scheme can also be arbitrarily high, even though the stencil still has only three points.

For the 20th-order scheme as an example, the roundoff level of the numerical error is reached for the uniform grid with just 10–15 nodes (Table 4.4.7). For a fixed grid size and varying order of the scheme, the error falls off very rapidly as the order is increased and quickly saturates at the roundoff level (Fig. 4.9).

Table 4.1. Errors for the 3-point FLAME scheme of order 20

Number of nodes	Mean absolute error
7	2.14E-10
11	2.06E-14
15	1.75E-15

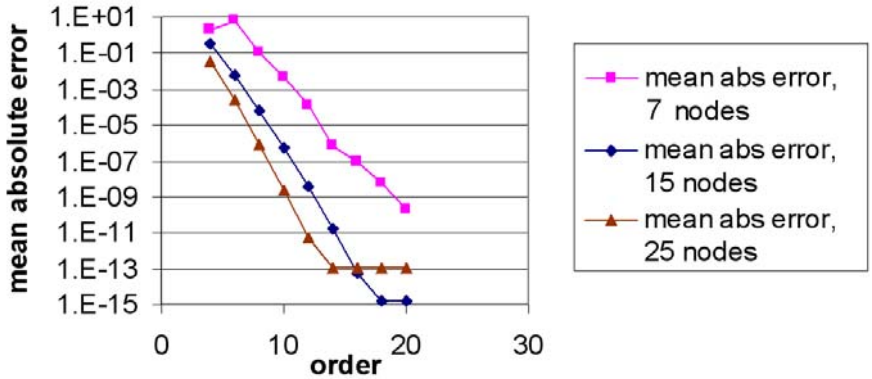


Fig. 4.9. Error vs. order of the Trefftz–FLAME scheme for the model Schrödinger equation. (Reprinted by permission from [Tsu06] ©2006 Elsevier.)

4.4.8 A Singular Equation

G.W. Reddien & L.L. Schumaker [RS76] (RS) proposed a spline-based collocation method for 1D singular boundary value problems and use the following example:[9]

$$(x^{0.5}u')' \;-\; x^{0.5}u \;=\; 0, \quad 0 < x < 1, \quad u(0) = 1, \quad u(1) = 0 \tag{4.50}$$

Here we apply the non-variational FLAME method to the same example and compare the results. A 3-point stencil on a uniform grid is used for FLAME. The two basis functions for FLAME are constructed separately for stencil points in the vicinity of the singularity point $x = 0$ and away from zero.

1) Let the midpoint x_i of the i-th stencil be sufficiently far away from zero (the singularity point of the differential equation): $x_i > \delta$, where δ is a chosen threshold. Expanding u over the i-th stencil into the Taylor series with respect to $\xi = x - x_i$,

$$u \;=\; \sum\nolimits_{k=0}^{\infty} c_k \xi^k \tag{4.51}$$

one obtains, by straightforward calculation, the following recursion:

$$c_{k+2} \;=\; \frac{c_k x_i + c_{k-1} - c_{k+1}(k+1)(k+\frac{1}{2})}{x_i(k+1)(k+2)}, \quad k = 0, 1, \ldots \tag{4.52}$$

where the coefficients with negative indices are understood to be zero. Two basis functions are obtained by choosing two independent sets of starting values for $c_{0,1}$ for the recursion and by retaining a finite number of terms, $k = K$, in series (4.51).

[9] This example is as a result of my short communication with Larry L. Schumaker and Douglas N. Arnold.

2) For $x_i < \delta$, the approach is similar but the series expansion is different:

$$u = \sum_{k=0}^{\infty} b_k x^{k/2} \tag{4.53}$$

Straightforward algebra again yields

$$\forall b_0, b_1; \quad b_2 = b_3 = 0$$

$$b_{k+2} = \frac{4b_{k-2}}{(k+1)(k+2)}, \quad k = 0, 1, \ldots \tag{4.54}$$

Two independent basis functions are then obtained in the same manner as above, with terms $k \leq 2K$ retained in (4.53).

Numerical values of the solution at $x = 0.5$ are given in [RS76, CR72] and serve as a basis for accuracy comparison. As Tables 4.2 and 4.3 show, Trefftz–FLAME gives orders of magnitude higher accuracy than the methods of [RS76, CR72]. The price for this accuracy gain is the analytical work needed for "preprocessing," i.e. for deriving the FLAME basis functions.

This example is intended to serve as an illustration of the capabilities of FLAME and its possible applications; it does not imply that FLAME is necessarily better than all methods designed for singular equations. Many other effective techniques have been developed (e.g. M. Kumar [Kum03]).

n	FLAME, $K = 6$	FLAME, $K = 12$	RS [RS76]	Jamet [Jam70]
8	0.25204513942296	0.252041978171219	0.25305	0.29038
16	0.252044597187729	0.252041977565477	0.25223	0.27826
8192	0.252042091673094	0.252041976551393		0.25310

Table 4.2. Numerical values of the solution at $x = 0.5$: FLAME vs. other methods. The number of grid subdivisions and the order of the scheme in FLAME varied.

n	FLAME, $K = 6$	FLAME, $K = 12$	RS [RS76]	Jamet [Jam70]
8	3.16E-06	1.68E-09	1.01E-03	3.83E-02
16	2.62E-06	1.07E-09	1.88E-04	2.62E-02
8192	1.15E-07	5.80E-11		1.06E-03

Table 4.3. Numerical errors of the solution at $x = 0.5$: FLAME vs. other methods. The result for the FLAME scheme of order 40 with 8192 grid subdivisions was treated as "exact" for the purposes of error evaluation.

4.4.9 A Polarized Elliptic Particle

This subsection gives an example of FLAME in two dimensions. A dielectric cylinder, with an elliptic cross-section, is immersed in a uniform external field. An analytical solution using complex variables is developed, for example, by W.B. Smythe [Smy89].

If $l_x > l_y$ are the two semiaxes of the ellipse and the applied external field is in the x-direction, then the solution in the first quadrant of the plane can be described by the following sequence of expressions [Smy89], with $z = x + \mathrm{i}y$:

$$\begin{aligned} \alpha^2 &= l_x^2 - l_y^2 \\ z_1 &= \frac{\alpha}{z - \sqrt{z^2 - \alpha^2}} \\ A'' &= \frac{(l_x + l_y)(l_x - \epsilon l_y)}{(l_x - l_y)(l_x + \epsilon l_y)} \\ B'' &= \frac{l_x + l_y}{l_x + \epsilon l_y} \end{aligned}$$

Potential outside the ellipse:

$$u = \mathrm{Re}\left[\frac{\alpha}{2}\left(z_1 + \frac{A''}{z_1}\right)\right]$$

Potential inside the ellipse:

$$u = \mathrm{Re}\left[\frac{\alpha}{2}B''\left(z_1 + \frac{1}{z_1}\right)\right]$$

Similar expressions hold in other quadrants and for the y-direction of the applied field.

In the numerical example below, the computational domain Ω is taken to be the unit square $[0,1] \times [0,1]$. To eliminate the numerical errors associated with the finite size of this domain, the analytical solution (for the x-direction of the external field) is imposed, for testing and verification purposes, as the Dirichlet condition on the exterior boundary of Ω.

For the usual 5-point stencil in 2D, four basis functions would normally be needed to yield the null space of dimension one in Trefftz–FLAME. The choice of *three* basis functions is clear: $\psi_1 = 1$, and $\psi_{2,3}$ are the theoretical solutions for two perpendicular directions of the applied external field (along each axis of the ellipse). Deriving a fourth Trefftz function is not worth the effort. Instead, Trefftz–FLAME is applied with the three basis functions. This yields a two-dimensional null space, with two independent 5-point difference schemes $\underline{s}_{1,2} \in \mathbb{R}^5$. It then turns out to be possible to find a linear combination

of these two schemes with a dominant diagonal entry, so that the convergence conditions of Section 4.3.5 are satisfied.[10]

The particular results below are for the material parameter $\epsilon_{\text{in}} = 10$ within the ellipse, for $\epsilon_{\text{out}} = 1$ outside the ellipse, and for the main axis of the ellipse aligned with the external field. The semiaxes are $l_x = 0.22$ and $l_y = 0.12$. The FLAME basis functions $\psi_{1,2,3}$ are introduced for all stencils having at least one node inside the ellipse and, in addition in some experiments, in several layers around the ellipse. These additional layers are such that $\xi_{\text{midpoint}} < \xi_{\text{cutoff}}$, where $\xi = (x/l_x)^2 + (y/l_y)^2 - 1$ (with x, y measured from the center of the ellipse), ξ_{midpoint} is the value of ξ for the midpoint of the stencil, and ξ_{cutoff} is an adjustable threshold. For $\xi_{\text{cutoff}} = 0$ no additional layers with the special basis are introduced. For $\xi_{\text{cutoff}} \gg 1$ the special bases are used throughout the domain, which yields the solution with machine precision.[11] Outside the cutoff, the standard 5-point scheme for the Laplace equation is applied, which asymptotically produces an $\mathcal{O}(h^2)$ bottleneck for the convergence rate.

Fig. 4.4.9 compares the relative errors in the potential (nodal 2-norm) for the standard flux balance scheme and the FLAME scheme. The errors are plotted vs. grid size h. For $\xi_{\text{cutoff}} = 0$, no additional layers with special bases are introduced in FLAME around the elliptic particle; for $\xi_{\text{cutoff}} = 3$, three such layers are introduced. It is evident that Trefftz–FLAME exhibits much more rapid convergence than the standard flux-balance scheme. The rate of convergence for FLAME is formally $\mathcal{O}(h^2)$, but only due to the above-mentioned bottleneck of the standard 5-point scheme away from the ellipse.

4.4.10 A Line Charge Near a Slanted Boundary

This problem was chosen in [Tsu05a] to illustrate how FLAME schemes can rectify the notorious "staircase" effect that occurs when slanted or curved boundaries are rendered on Cartesian grids. The electrostatic field is generated by a line charge located near a slanted material interface boundary between air (relative dielectric constant $\epsilon = 1$) and water ($\epsilon = 80$). This can be viewed as a drastically simplified 2D version of electrostatic problems in macro- and biomolecular simulation [Sim03, RAH01, GPN01].

Four basis functions on a 5-point stencil at the interface boundary were obtained by matching polynomial approximations in the two media via the boundary conditions. As demonstrated in [Tsu05a], the Trefftz–FLAME result is substantially more accurate than solutions obtained with the standard flux-balance scheme.

[10] Diagonal dominance has been monitored and verified in numerical simulations but has not been shown analytically. Therefore, convergence of the scheme is not proven rigorously, but the numerical evidence for it is very strong.

[11] Because in this example the exact solution happens to lie in the FLAME space.

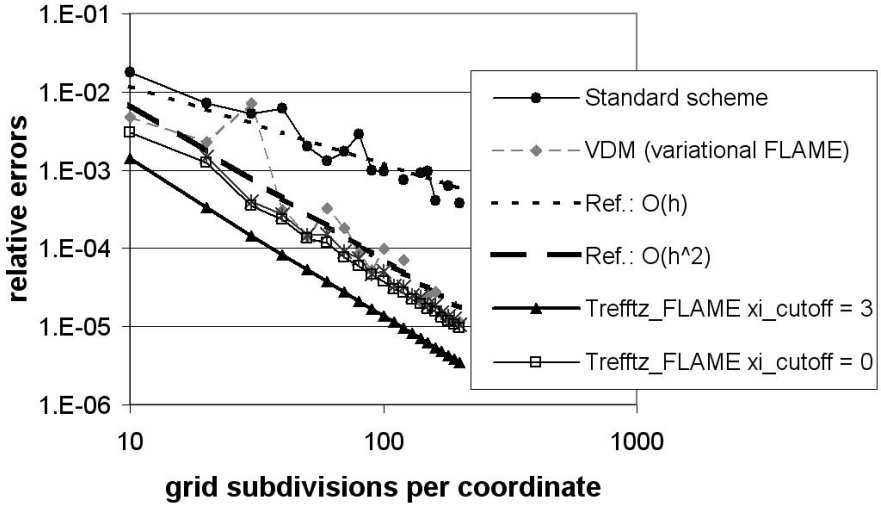


Fig. 4.10. The 5-point Trefftz–FLAME scheme yields much faster convergence than the standard 5-point flux-balance scheme. The numerical error in FLAME is reduced if special bases are introduced in several additional layers of nodes outside the particle. (Reprinted by permission from [Tsu05a] ©2005 IEEE.)

4.4.11 Scattering from a Dielectric Cylinder

In this classic example, a monochromatic plane wave impinges on a dielectric circular cylinder and gets scattered. The analytical solution is available via cylindrical harmonics (R.F. Harrington [Har01]) and can be used for verification and error analysis. The basis functions in FLAME are cylindrical harmonics in the vicinity of the cylinder and plane waves away from the cylinder. The 9-point (3×3) stencil is used throughout the domain (with the obvious truncation to 6 and 4 nodes at the edges and corners, respectively). A Perfectly Matched Layer is introduced in some test cases [Tsu05a] using FLAME. Very rapid 6th-order convergence of the nodal values of the field was experimentally observed when the Dirichlet conditions were imposed on the exterior boundary of the computational domain. It would be quite difficult to construct a conventional difference scheme with comparable accuracy in the presence of such material interfaces.

In this section and the following one, we consider the E-mode (one-component E field and a TM field) governed by the standard 2D equation

$$\nabla \cdot (\mu^{-1} \nabla E) \; + \; \omega^2 \epsilon E \; = \; 0 \tag{4.55}$$

with some radiation boundary conditions for the scattered field. The analytical solution is available via cylindrical harmonics [Har01] and can be used for verification and error analysis.

We consider Trefftz–FLAME schemes on a 9-point (3×3) stencil. It is natural to choose the basis functions as cylindrical harmonics in the vicinity of each particle and as plane waves away from the particles. "Vicinity" is defined by an adjustable threshold: $r \leq r_{\text{cutoff}}$, where r is the distance from the midpoint of the stencil to the center of the nearest particle, and the threshold r_{cutoff} is typically chosen as the radius of the particle plus a few grid layers.

Away from the cylinder, eight basis functions are chosen as plane waves propagating toward the central node of the 9-point stencil from each of the other eight nodes. As usual in FLAME, the 9×8 nodal matrix N (4.14) of FLAME comprises the values of the chosen basis functions at the stencil nodes. The Trefftz–FLAME scheme (4.20) is $\underline{s} = \operatorname{Null} N^T$. Straightforward symbolic algebra computation shows that this null space is indeed of dimension one, so that a single valid Trefftz–FLAME scheme exists. Expressions for the coefficients $\underline{s}$ are given in Appendix 4.8, and the scheme turns out to be of order six with respect to the grid size. The scheme is used in several nanophptonics applications in Chapter 7.

Obviously, nodes at the domain boundary are treated differently. At the edges of the domain, the stencil is truncated in a natural way to six points: "ghost" nodes outside the domain are eliminated, and the respective *incoming* plane waves associated with them are likewise eliminated from the basis set. The basis thus consists of five plane waves: three strictly outgoing and two sliding along the edge.

A similar procedure is applied at the corner nodes: a four-node stencil is obtained, and only three plane wave remain in the basis. The elimination of incoming waves from the basis thus leads, in a very natural way, to a FLAME-style Perfectly Matched Layer (PML).

In the vicinity of the cylinder, the basis functions are chosen as cylindrical harmonics:

$$\begin{aligned} \psi_\alpha^{(i)} &= a_n J_n(k_{\text{cyl}} r) \exp(\mathrm{i} n \phi), \quad r \leq r_0 \\ \psi_\alpha^{(i)} &= [b_n J_n(k_{\text{air}} r) + H_n^{(2)}(k_{\text{air}} r)] \exp(\mathrm{i} n \phi), \quad r > r_0 \end{aligned}$$

where J_n is the Bessel function, $H_n^{(2)}$ is the Hankel function of the second kind [Har01], and a_n, b_n are coefficients to be determined. These coefficients are found via the standard conditions on the boundary of the cylinder; the actual expressions for these coefficients are too lengthy to be worth reproducing here but are easily usable in computer codes.

Eight basis functions are obtained by retaining the monopole harmonic ($n = 0$), two harmonics of orders $n = 1, 2, 3$ (i.e. dipole, quadrupole and octupole), and one of harmonics of order $n = 4$. Numerical experiments for scattering from a *single* cylinder, where the analytical solution is available for comparison and verification, show convergence (not just consistency error!) of order six for this scheme [Tsu05a].

Fig. 4.11 shows the relative nodal error in the electric field as a function of the mesh size. Without the PML, convergence of the scheme is of 6th

order; no standard method has comparable performance. The test problem

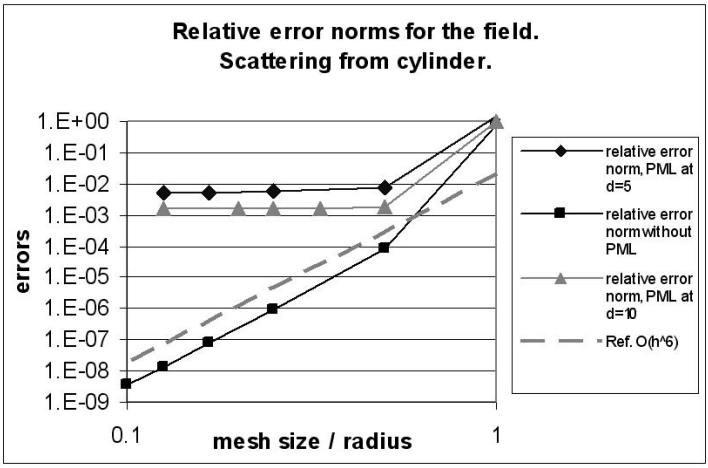


Fig. 4.11. Relative error norms for the electric field. Scattering from a dielectric cylinder. FLAME, 9-point scheme. (Reprinted by permission from [Tsu05a] ©2005 IEEE.)

has the following parameters: the radius of the cylindrical rod is normalized to unity; its index of refraction is 4; the wavenumbers in air and the rod are 1 and 4, respectively. Simulations *without* the PML were run with the exact analytical value of the electric field on the outer boundary imposed as a Dirichlet condition. The field error with the PML is of course higher than with this ideal Dirichlet condition[12] but still only on the order of 10^{-3} even when the PML is close to the scatterer (1 – 1.5 wavelengths). For the exact boundary conditions (and no PML), very high accuracy is achievable.

4.5 Existing Methods Featuring Flexible or Nonstandard Approximation

FLAME schemes are conceptually related to many other methods:

1. Generalized FEM by Partition of Unity [MB96, BM97, DBO00, SBC00, BBO03, PTFY03, PT02, BT05] and "*hp*-cloud" methods [DO96].
2. Homogenization schemes based on variational principles [MDH+99].
3. Spectral and pseudospectral methods [Boy01, DECB98, Ors80, PR04] (and references therein).

[12] It goes without saying that the exact field condition can only be imposed in test problems with known analytical solutions.

4. Meshless methods [BLG94, BKO+96, CM96, DB01, KB98, LJZ95, BBO03, Liu02], and especially the "Meshless Local Petrov–Galerkin" version [AZ98, AS02].
5. Heuristic homogenization schemes, particularly in Finite Difference Time Domain methods [DM99, TH05, YM01].
6. Discontinuous Galerkin (DG) methods [ABCM02, BMS02, CBPS00, CKS00, OBB98].
7. Finite Integration Techniques (FIT) with extensions and enhancements [CW02, SW04].
8. Special FD schemes such as "exact" and "nonstandard" schemes by Mickens and others [Mic94, Mic00]; the Harari–Turkel [HT95] and Singer–Turkel schemes [ST98] for the Helmholtz equation; the Hadley schemes [Had02a, Had02b] for waveguide analysis; Cole schemes for wave propagation [Col97, Col04]; the Lambe–Luczak–Nehrbass schemes for the Helmholtz equation [LLN03].
9. Special finite elements, for example elements with holes [SL00] or inclusions [MZ95].
10. The "Measured Equation of Invariance" (MEI) [MPC+94]).
11. The "immersed surface" methodology [WB00] that modifies the Taylor expansions to account for derivative jumps at material boundaries but leads to rather unwieldy expressions.

This selection of related methods is to some extent subjective and definitely not exhaustive. Most methods and references above are included because they influenced my own research in a significant way.

Even though the methods listed above share some level of "flexible approximation" as one of their features, the term "Flexible Local Approximation MEthods" (FLAME) will refer exclusively to the approach developed in Sections 4.3 and 4.7. The new FLAME schemes are not intended to absorb or supplant any of the methods 1–11. These other methods, while related to FLAME, are not, generally speaking, its particular cases; nor is FLAME a particular case of any of these methods.

Consider, for example, a connection between FLAME on the one hand and variational homogenization (item 2 on the list above) and GFEM (item 1) on the other. The development of FLAME schemes was motivated to a large extent by the need to reduce the computational and algorithmic complexity of Generalized FEM and variational homogenization (especially the volume quadratures inherent in these methods). However, FLAME is emphatically *not* a version of GFEM or variational homogenization of [MDH+99]. Indeed, GFEM is a Galerkin method in the functional space constructed by partition of unity; the variational homogenization is, as argued in [Tsu04c], a Galerkin method in broken Sobolev spaces. In contrast, FLAME is in most cases a non-Galerkin, purely finite-difference method.

The variational version of FLAME is described in [Tsu04b] in a condensed manner; see also Appendix 4.7.3 on p. 232. The crux of this chapter,

however, is the non-variational "Trefftz" version of FLAME (Section 4.3) [Tsu05a, Tsu06]. In this version, the basis functions satisfy the underlying differential equation and the variational testing is therefore redundant. Numerical quadratures – the main bottleneck of Generalized FEM, variational homogenization, meshless and other methods – are completely absent. Despite their relative simplicity, the Trefftz–FLAME schemes are in many cases more accurate than their variational counterparts. This chapter, following [Tsu05a, Tsu06], presents a variety of examples for Trefftz–FLAME, including the 1D Schrödinger equation, a singular 1D equation, 2D and 3D Collatz "Mehrstellen" schemes, and others. Applications to heterogeneous electrostatic problems for colloidal systems are considered in Chapter 6, and to problems in photonics in Chapter 7.

4.5.1 The Treatment of Singularities in Standard FEM

The treatment of singularities was historically one of the first cases where special approximating functions were used in the FE context. In their 1973 paper [FGW73], G.J. Fix *et al.* considered 2D problems with singularities $r^\gamma \sin \beta\phi$, where r, ϕ are the polar coordinates with respect to the singularity point, and β, γ, are known parameters ($\gamma < 0$). The standard FEM bases were enriched with functions of the form $p(r)\, r^\gamma \sin \beta\phi$, where the piecewise-polynomial cutoff function $p(r)$ is unity within a disk $0 \le r \le r_0$, gradually decays to zero in the ring $r_0 \le r \le r_1$ and is zero outside that ring (r_0 and r_1 are adjustable parameters). The cutoff function is needed to maintain the sparsity of the stiffness matrix.

There is clearly a tradeoff between the computational cost and accuracy: if the cutoff radius r_1 is too small, the singular component of the solution is not adequately represented; but if it is too large, the support of the additional basis function overlaps with a large number of elements and the matrix becomes less sparse.

The Generalized FEM (GFEM) briefly described in the following subsection preserves, at least in principle, both accuracy and sparsity. Unfortunately, this major advantage is tainted by additional algorithmic and computational complexity.

4.5.2 Generalized FEM by Partition of Unity

In the Generalized FEM computational domain Ω is covered by overlapping subdomains ("patches") $\Omega^{(i)}$. The solution is approximated locally over each patch. These individual local approximations are independent from one another and are seamlessly merged by Partition of Unity (PU). Details of the method are widely available (see J.M. Melenk & I. Babuška [MB96, BM97], C.A. Duarte *et al.* [DBO00], T. Strouboulis *et al.* [SBC00], I. Babuška *et al.* [BBO03]), and additional information can also be found in the chapter on FEM (Section 3.15 on p. 181).

The main advantage of GFEM is that the approximating functions can in principle be arbitrary and are not limited to polynomials. Thus GFEM definitely qualifies as a method with the kind of flexible local approximation we seek.

On the negative side, however, multiplication by the partition of unity functions makes the system of approximating functions more complicated, and possibly ill-conditioned or even linearly dependent [BM97]. The computation of gradients and implementation of the Dirichlet conditions also get more complicated. In addition, GFEM-PU may lead to a combinatorial increase in the number of degrees of freedom [PTFY03, Tsu04c]. An even greater difficulty in GFEM-PU is the high cost of the Galerkin quadratures that need to be computed numerically in geometrically complex 3D regions (intersections of overlapping patches).

In summary, there is a high algorithmic and computational price to be paid for all the flexibility that GFEM provides.

4.5.3 Homogenization Schemes Based on Variational Principles

S. Moskow *et al.* [MDH$^+$99] improve the approximation of the electrostatic potential near slanted boundaries and narrow sheets on regular Cartesian grids by employing special approximating functions constructed by a coordinate mapping [BCO94]. Within each grid cell, Moskow *et al.* seek a tensor representation of the material parameter such that the discrete and continuous energy inner products are the same over the chosen discrete space. The overall construction in [MDH$^+$99] relies on a special partitioning of the grid ("red-black" numbering, or the "Lebedev grid") and on a specific, central difference, representation of the gradient. As shown in [Tsu04c], this variational homogenization can be interpreted as a Galerkin method in a broken Sobolev space.

The variational method described in Section 4.7 can be viewed as an extension of the variational-difference approach of [MDH$^+$99] – the special "Lebedev" grids and the specific approximation of gradients by central differences adopted in [MDH$^+$99] turn out not to be really essential for the algorithm.

4.5.4 Discontinuous Galerkin Methods

The idea to relax the interelement continuity requirements of the standard FEM and to use nonconforming elements was put forward at the early stages of FE research. For example, in the Crouzeix–Raviart elements [CR73] the continuity of piecewise-linear functions is imposed only at midpoints of the edges.

Over recent years, a substantial amount of work has been devoted to Discontinuous Galerkin Methods (DGM) [BMS02, CBPS00, CKS00, OBB98]; a consolidated view with an extensive bibliography is presented in [ABCM02]. Many of the approaches start with the "mixed" formulation that includes

additional unknown functions for the fluxes on element edges (2D) or faces (3D). However, these additional unknowns can be replaced with their numerical approximations, thereby producing a "primal" variational formulation in terms of the scalar potential alone. In DGM, the interelement continuity is ensured, at least in the weak sense, by retaining the surface integrals of the jumps, generally leading to saddle-point problems even if the original equation is elliptic.

In electromagnetic field computation, DGM was applied by P. Alotto *et al.* to moving meshes in the air gap of machines [ABPS02].

4.5.5 Homogenization Schemes in FDTD

In applied electromagnetics, Finite Difference Time Domain (FDTD) methods (A. Taflove & S.C. Hagness [TH05]) and Finite Integration Techniques (FIT, T. Weiland, M. Clemens & R. Schuhmann [CW02, SW04]) typically require very extensive computational work due to a large number of time steps for numerical wave propagation and large meshes. Therefore simple Cartesian grids are strongly preferred and the need to avoid "staircase" approximations of curved or slanted boundaries is quite acute. Due to the wave nature of the problem, any local numerical error, including the errors due to the staircase effect, tend to propagate in space and time and pollute the solution overall.

A great variety of approaches to reduce or eliminate the staircase effect in FDTD have been proposed [DM99, TH05, YM01, ZSW03]. Each case is a trade-off between the simplicity of the original Yee scheme on staggered grids (K.S. Yee [Yee66]) and the ability to represent the interface boundary conditions accurately. On one side of this spectrum lie various adjustments to the Yee scheme: changes in the time-stepping formulas for the magnetic field or heuristic homogenization of material parameters based on volume or edge length ratios [DM99, TH05, YM01]. A similar homogenization approach (albeit not for time domain simulation) was applied by R.D. Meade and coworkers to compute the bandgap structure of photonic crystals[13] [MRB+93]. In some cases, the second order of the FDTD scheme is maintained by including additional geometric parameters or by using partially filled cells, as done by I.A. Zagorodnov *et al.* [ZSW03] in the framework of "Finite Integration Techniques".

On the other side of the spectrum are Finite Volume–Time Domain methods (FVTD) [PRM98, TH05, YC97] with their historic origin in computational fluid mechanics, and the Finite Element Method (FEM). Tetrahedral meshes are typically used, and material interfaces are represented much more accurately than on Cartesian grids. However, adaptive Cartesian grids have also been advocated, with cell refinement at the boundaries [WPL02]. The greater geometric flexibility of these methods is achieved at the expense of

[13] For more information on photonic bandgaps, see Chapter 7.

simplicity of the algorithm. An additional difficulty arises in FEM for time-domain problems: the "mass" matrix (containing the inner products of the basis functions) appears in the time derivative term and makes the time-stepping procedure implicit, unless "mass-lumping" techniques are used.

4.5.6 Meshless Methods

The abundance of meshless methods, as well as many variations in the terminology adopted in the literature, make a thorough review unfeasible here – see [BLG94, BKO+96, CM96, DB01, KB98, LJZ95, BBO03] instead. Let me highlight only the main ideas and features.

The prevailing technique is the Moving Least Squares (MLS) approximation. Consider a "meshless" set of nodes (that is, nodes selected at arbitrary locations r_i, $i = 1, 2, \ldots n$) in the computational domain. For each node i, a smooth weighting function $W_i(r)$ with a compact support is introduced; this function would typically be normalized to one at node i (i.e. at $r = r_i$) and decay to zero away from that node. Intuitively, the support of the weighting function defines the "zone of influence" for each node.

Let u be a smooth function that we wish to approximate by MLS. For any given point r_0, one considers a linear combination of a given set of m basis functions $\psi_\alpha(r)$ (almost always polynomials in the MLS framework): $u_h^{(i)} = \sum_{\alpha=1}^{m} c_\alpha(r_0)\psi_\alpha(r)$. Note that the coefficients c depend on r_0. They are chosen to approximate the nodal values of u, i.e. the Euclidean vector $\{u(r_i)\}$, in the least-squares sense with respect to the weighted norm with the weights $W_i(r_0)$. This least-squares problem can be solved in a standard fashion; note that it involves only nodes containing r_0 within their respective "zones of influence" – in other words, only nodes i for which $W_i(r_0) \neq 0$.

C.A. Duarte & J.T. Oden [DO96] showed that this procedure can be recast as a partition of unity method, where the PU functions are defined by the weighting functions W as well as the (polynomial) basis set $\{\psi\}$. This leads to more general adaptive "*hp*-cloud" methods.

One version of meshless methods – "Meshless Local Petrov-Galerkin" (MLPG) method developed by S.N. Atluri *et al.* [AZ98, AS02, Liu02] – is particularly close to the variational version of FLAME described in [Tsu04b] and in Section 4.7 below. Our emphasis, however, is not on the "meshless" setup (even though it is conceivable for FLAME) but on the framework of multivalued approximation (that is not explicitly introduced in MLPG) and on the new *non*-variational version of FLAME (Section 4.3).

The trade-off for avoiding complex mesh generation in mesh-free methods is the increased computational and algorithmic complexity. The expressions for the approximating functions obtained by least squares are rather complicated [BKO+96, DB01, KB98, LJZ95, BBO03]. The *derivatives* of these functions are even more involved. These derivatives are part of the integrand in the Galerkin inner product, and the computation of numerical quadratures is a bottleneck in meshless methods. Other difficulties include the treatment

of Dirichlet conditions and interface conditions across material boundaries [CM96, DB01, KB98, LJZ95].

4.5.7 Special Finite Element Methods

There is also quite a number of special finite elements, and related methods, that incorporate specific features of the solution. In problems of solid mechanics, J. Jirousek and his coworkers in the 1970s [JL77, Jir78] proposed "Trefftz" elements, with basis functions satisfying the underlying differential equation exactly. This not only improves the numerical accuracy substantially, but also reduces the Galerkin volume integrals in the computation of stiffness matrices to surface integrals (via integration by parts). Since then, Trefftz elements have been developed quite extensively; see a detailed study by I. Herrera [Her00] and a review paper by J. Jirousek & A.P. Zielinski [JZ97].

Also in solid mechanics, A.K. Soh & Z.F. Long [SL00] proposed two 2D elements with circular holes, while S.A. Meguid & Z.H. Zhu [MZ95] developed special elements for the treatment of inclusions.

Enrichment of FE bases with special functions is well established in computational mechanics. The variational multiscale method by T.J.R. Hughes [Hug95] provides a general framework for adding fine-scale functions inside the elements to the usual coarse-scale FE basis. The additional amount of computational work is small if the fine scale bases are *local*, i.e. confined to the support of a single element. However, in this case the global effects of the fine scale are lost.

In the method of Residual-Free Bubbles by F. Brezzi *et al.* [BFR98], the standard element space is enriched with functions *satisfying the underlying differential equation* exactly. There is a similarity with the Trefftz–FLAME schemes described in Section 4.3. However, FLAME is a finite-difference technique rather than a Galerkin finite element method. The conformity in Residual-Free Bubbles is maintained by having the "bubbles" vanish at the interelement boundaries. Similar "bubbles" are common in hierarchical finite element algorithms (see e.g. Yserentant [Yse86]); still, traditional FE methods – hierarchical or not – are built exclusively on piecewise-polynomial bases.

C. Farhat *et al.* [FHF01] relax the conformity conditions and get a higher flexibility of approximation in return. As in the case of residual-free bubbles, functions satisfying the differential equation are added to the FE basis. However, the continuity at interelement boundaries is only weakly enforced via Lagrange multipliers.

The following observation by J.M. Melenk [Mel99] in reference to special finite elements is highly relevant to our discussion:

> "The theory of homogenization for problems with (periodic) microstructure, asymptotic expansions for boundary layers, and Kondrat'ev's corner expansions are a few examples of mathematical techniques yielding knowledge about the local properties of the solution.

> This knowledge may be used to construct local approximation spaces which can capture the behavior of the solution much more accurately than the standard polynomials for a given number of degrees of freedom. Exploiting such information may therefore be much more efficient than the standard methods ...”

In electromagnetic analysis, Treffz expansions were used by M. Gyimesi *et al.* in unbounded domains [GLOP96, GWO01].

4.5.8 Domain Decomposition

Although the setup of FLAME may suggest its interpretation as a Domain Decomposition technique, there are perhaps more differences than similarities between the two classes of methods. In FLAME, the domain cover consists of “micro” (stencil-size) subdomains. In contrast, Domain Decomposition methods usually operate with “macro” subdomains that are relatively large compared to the mesh size. Consequently, the notions and ideas of Domain Decomposition (e.g. Schwartz methods, mortar methods, Chimera grids, and so on) will not be directly used in our development. With regard to Domain Decomposition, the book by A. Toselli & O. Widlund [TW05] is recommended.

4.5.9 Pseudospectral Methods

In pseudospectral methods (PSM) [Boy01, DECB98, Ors80, PR04], numerical solution is sought as a series expansion in terms of Fourier harmonics, Chebyshev polynomials, etc. The expansion coefficients are found by collocating the differential equation on a chosen set of grid nodes.

Typically the series is treated as *global* – over the whole domain or large subdomains. There is, however, a great variety of versions of pseudospectral methods, some of which (“spectral elements”) deal with more localized approximations and in fact overlap with the *hp*-version of FEM (J.M. Melenk *et al.* [MGS01]).

The key advantage of PSM is their exponential convergence, provided that the solution is quite smooth over the whole domain.

One major difficulty is the treatment of complex geometries. In relatively simple cases this can be accomplished by a global mapping to a reference shape (square in 2D or cube in 3D) but in general may not be possible. Another alternative is to subdivide the domain and use spectral elements (with “spectral” approximation within the elements but lower order smoothness across their boundaries); however, convergence is then algebraic, not exponential, with respect to the parameter of that subdivision.

The presence of material interfaces is an even more serious problem, as the solution then is no longer smooth enough to yield the exponential convergence of the global series expansion.

An additional disadvantage of PSM is that the resultant systems of equations tend to have much higher condition numbers than the respective FD or FE systems (E.H. Mund [Mun00]). This is due to the very uneven spacing of the Chebyshev or Legendre collocation nodes typically used in PSM. Ill-conditioning may lead to accuracy loss in general and to stability problems in time-stepping procedures.

PSM have been very extensively studied over the last 30 years, and quite a number of approaches alleviating the above disadvantages have been proposed [DECB98, MGS01, Mun00], [Ors80]. Nevertheless it would be fair to say that these disadvantages are inherent in the method and impede its application to problems with complex geometries and material interfaces.

4.5.10 Special FD Schemes

Many difference schemes rely on special approximation techniques to improve the numerical accuracy. These special techniques are too numerous to list, and only the ones that are closely related to the ideas of this chapter are briefly reviewed below.

For some 1D equations, R.E. Mickens [Mic94] constructed "exact" FD schemes – that is, schemes with zero consistency error. He then developed a wider class of "nonstandard" schemes by modifying finite difference approximations of derivatives. These modified approximations are *asymptotically* (as the mesh size tends to zero) equivalent to the standard ones but for finite mesh sizes may yield higher accuracy. Similar ideas were used by I. Harari & E. Turkel [HT95] and by I. Singer & E. Turkel [ST98] to construct exact and high-order schemes for the Helmholtz equation. J.B. Cole [Col97, Col04] applied nonstandard methods to electromagnetic wave propagation problems (high-order schemes) in 2D and 3D.

The "immersed surface" methodology (A. Wiegmann and K.P. Bube [WB00]) generalizes the Taylor expansions to account for derivative jumps at material boundaries but leads to rather unwieldy expressions.

J.W. Nehrbass [Neh96] and L.A. Lambe *et al.* [LLN03] modified the central coefficient of the standard FD scheme for the Helmholtz equation to minimize, in some sense, the average consistency error over plane waves propagating in all possible directions. There is some similarity between the Nehrbass schemes and FLAME. However, the derivation of the Nehrbass schemes requires very elaborate symbolic algebra coding, as the averaging over all directions of propagation leads to integrals that are quite involved. In contrast, FLAME schemes are inexpensive and easy to construct.

Very closely related to the material of this chapter are the special difference schemes developed by G.R. Hadley [Had02a, Had02b, Web07] for electromagnetic wave propagation. In fact, these schemes are direct particular cases of FLAME, with Bessel functions forming a Trefftz–FLAME basis (although Hadley derives them from different considerations).

For unbounded domains, an artificial truncating boundary has to be introduced in FD and FE methods. The exact conditions at this boundary are nonlocal; however, local approximations are desirable to maintain the sparsity of the system matrix. One such approximation that gained popularity in the 1990s is the so called "Measured Equation of Invariance" (MEI) by K.K. Mei *et al.* [MPC+94, GRSP95, HR98a]. As it happens, MEI can be viewed as a particular case of Trefftz–FLAME, with the basis functions taken as potentials due to some test distributions of sources.

4.6 Discussion

The "Flexible Approximation" approach combines analytical and numerical tools: it integrates local analytical approximations of the solution into numerical schemes in a simple way. Existing applications and special cases of FLAME are listed in the following table (see Chapters 6 and 7 for applications of FLAME to electrostatics of colloidal systems and in nano-photonics). The cases in the table fall under two categories. The first one contains standard difference schemes revealed as direct particular cases of Trefftz–FLAME. The second category contains FLAME schemes that are substantially different from, and are more accurate than, their conventional counterparts, often with a higher rate of convergence for identical stencils. Practical implementation of Trefftz–FLAME schemes is substantially simpler than FEM matrix assembly and comparable with the implementation of conventional schemes (e.g. flux-balance schemes).

It is worth noting that FLAME schemes do not have any hidden parameters to contrive better performance. The schemes are completely defined by the choice of the basis set and stencil; it is the approximating properties of the basis that have the greatest bearing on the numerical accuracy.

The collection of examples in Table 4.4 inspires further analysis and applications of FLAME. The table is in no way exhaustive – for example, boundary layers in eddy current problems and in semiconductor simulation (the Scharfetter–Gummel approximation, S. Selberherr [Sel84, Fri05]), varying material parameters in some protein models, J.A. Grant *et al.* [GPN01], T. Washio [Was03], etc., could be added to this list.

Two broad application areas for FLAME – one at zero frequency (electrostatics of colloids and macromolecules in solvents) and another one at very high frequencies (photonics) – are considered in Chapters 6 and 7, respectively.

The method is most powerful when good local analytical approximations of the solution are available. For example, the advantage of the special field approximation in FLAME for a photonic crystal problem is crystal clear in [Tsu05a]; see Chapter 7. Similarly, problems with magnetizable or polarizable particles admit an accurate representation of the field around the particles in terms of spherical harmonics, and the resultant FLAME schemes are substantially more accurate than the standard control volume method.

Application	Basis functions used in FLAME	Stencil used in FLAME	Accuracy of FLAME schemes	Comparison with standard finite-difference schemes
Standard schemes for the 3D Laplace equation	Local harmonic polynomials	Depends on the order	2nd order for the 7-point stencil	Standard schemes are a simple particular case of FLAME
Mehrstellen scheme for the 3D Laplace equation	Harmonic polynomials in x, y, z up to order 4	19-point	4th order	The "Mehrstellen"-Collatz scheme revealed as a natural particular case of FLAME
1D Schrödinger equation	High-order Taylor approximations to the solution	3-point	Any desired order	The Numerov scheme is 4th order. 3-point schemes of order higher than 4 not available.
1D heat conduction with variable material parameter	Independent local solutions of the heat equation	3-point	Exact (machine precision in practice)	So-called "homogeneous" schemes [Sam01] are a particular case of FLAME.
Time-domain scalar wave equation (one spatial dimension)	Traveling waves (polynomials times sinusoids)	5-point	2nd order in the generic case	In the generic case, equivalent to central differences. Much higher accuracy if a dominant frequency is present.
Slanted material interface boundary	Local polynomials satisfying interface matching conditions	7-point in 3D, 5-point in 2D	2nd order	Standard schemes, unlike FLAME, suffer "staircase" effects
Unbounded problems	Multipole harmonics outside the computational domain	7-point in 3D	See [HFT04]	Standard finite-difference schemes not applicable to unbounded problems.
Charged colloidal particles, no salt	Spherical harmonics (up to quadrupole)	7-point	2nd order	Much higher accuracy than the standard flux-balance scheme.
Charged colloidal particles, monovalent salt (Poisson–Boltzmann)	Spherical Bessel harmonics (up to quadrupole)	7-point	2nd order	Much higher accuracy than the standard scheme.
Scattering from a dielectric cylinder (frequency domain)	Plane waves in air and cylindrical harmonics near scatterer	9-point	6th order	Much higher accuracy than the standard scheme.

Table 4.4. Examples and applications of FLAME.

Perfectly Matched Layer (frequency domain)	Outgoing plane waves	9-point	Under investigation	
Waves, eigenmodes and band structure in photonic crystals [PWT07, Tv07]	Cylindrical harmonics	9-point	6th order	Much higher accuracy than the standard scheme and FEM with 2nd order triangular elements.
Coupled plasmonic particles	Plane waves in air and cylindrical harmonics near particles	9-point	6th order	Much higher accuracy than the standard scheme.

Table 4.5. Examples and applications of FLAME (continued).

Trefftz–FLAME schemes are not variational, which makes their construction quite simple and sidesteps the notorious bottleneck of computing numerical quadratures. At the same time, given that this method is non-variational and especially non-Galerkin, one cannot rely on the well-established convergence theory so powerful, for example, in Finite Element analysis. For the time being, FLAME needs to be considered on a case-by-case basis, with the existing convergence results (Section 4.3.5) and experimental evidence (Section 4.4) in mind. Furthermore, again because the method is non-Galerkin, the system matrix is in general not symmetric, even if the underlying continuous operator is self-adjoint. In many – but not all – cases, this shortcoming is well compensated for by the superior accuracy and rate of convergence (Section 4.4).

FLAME schemes use nodal values as the primary degrees of freedom (d.o.f.). Other d.o.f. could certainly be used, for example edge circulations of the field. The matrix of edge circulations would then be introduced instead of the matrix of nodal values in the algorithm, and the notion of the stencil would be modified accordingly. In the FE context (edge elements), this choice of d.o.f. is known to have clear mathematical and physical advantages in various applications (A. Bossavit [Bos98], R. Hiptmair [Hip01], C. Mattiussi [Mat97], E. Tonti [Ton02]) and is therefore worth exploring in the FLAME framework as well.

It is hoped that the ideas presented in this chapter will prove useful for further development of difference schemes in various areas. Such schemes can be eventually incorporated into existing FD software packages for use by many researchers and practitioners.

In the foreseeable future, FEM, due to its unrivaled generality and robustness, will remain king of simulation. However, FLAME schemes may successfully occupy the niches where the geometric and physical layout is too complicated to be handled on conforming FE meshes, while the standard

finite-difference approximation is too crude. One example is the simulation of electrostatic multiparticle interactions in colloidal systems, where FEM is impractical and Fast Multipole methods may not be suitable due to nonlinearity and inhomogeneities (Chapter 6).[14]

Another example, albeit more complicated, is the simulation of macromolecules, including proteins and polyelectrolytes [DTRS07]. In such problems, electrostatic interactions of atoms in the presence of the solvent are extremely important but are still only part of an enormously complicated physical picture. Yet another example of a "niche application" where FLAME can work very well is wave analysis in photonic crystals (Chapter 7) [PWT07, Tv07].

The possible applications of FLAME could be significantly expanded if accurate local *numerical* approximations rather than analytical ones are used to generate a FLAME basis. This approach involves solution of local problems around grid stencils. Such "mini-problems" can be handled by finite element or integral equation techniques much more cheaply than the global problem. FLAME schemes in this case may continue to operate on simple and relatively coarse Cartesian grids that do not necessarily have to resolve all geometric features [DT06]. Applications of this methodology to problems of electromagnetic interference in high density VLSI modules are currently being explored.

Finally, any modern algorithm has to be *adaptive*. The possibility of adaption and a numerical example are considered in Section 6.2.3 on p. 300.

In addition to practical usage and to the potential of generating new difference schemes in various applications, there is also some intellectual merit in having a unified perspective on different families of FD techniques such as low- and high-order Taylor-based schemes, the Mehrstellen schemes, the "exact" schemes, some special schemes for electromagnetic wave propagation, the "measured equation of invariance," and more. This unified perspective is achieved through systematic use of local approximation spaces *in the finite difference context*.

4.7 Appendix: Variational FLAME

4.7.1 References

The variational version of FLAME was described in [Tsu04b, Tsu04c]; this section follows [Tsu04b]. Variational FLAME is very close to the "Meshless Local Petrov-Galerkin" (MLPG) method developed by S.N. Atluri and collaborators [AZ98, AS02][15] (see also G.R. Liu's book [Liu02]).

The variational version is now to a large extend superseded by a *non*-variational one – the "Trefftz–FLAME" schemes introduced in [Tsu05a,

[14] Software for large-scale Trefftz–FLAME simulations of electrostatic interactions in colloidal suspensions was developed by E. Ivanova and S. Voskoboynikov.

[15] I thank Jon Webb for bringing this to my attention [Web07].

Tsu06] and described in this chapter. The general setup – multivalued approximation over a domain cover by overlapping parches and a set of nodes – is common for all versions of FLAME.

4.7.2 The Model Problem

Although the potential application areas of FLAME are broad, for illustrative purposes we shall have in mind the model static Dirichlet boundary-value problem

$$Lu \equiv -\nabla \cdot \epsilon \nabla u = f \quad \text{in } \Omega \subset \mathbb{R}^n, \ (n = 2, 3); \quad u_{|\partial\Omega} = 0 \tag{4.56}$$

Here ϵ is a material parameter (conductivity, permittivity, permeability, etc.) that can be discontinuous across material boundaries and can depend on coordinates but not, in the linear case under consideration, on the potential u. The computational domain Ω is either two- or three-dimensional, with the usual mathematical assumption of a Lipschitz-continuous boundary. To simplify the exposition, precise mathematical definitions of the relevant functional spaces will not be given, and instead we shall assume that the solution has the degree of smoothness necessary to justify the analysis.

At any material interface boundary Γ, the usual conditions hold:

$$u_1 = u_2 \quad \text{on } \Gamma \tag{4.57}$$

$$\epsilon_1 \frac{\partial u_1}{\partial n} = \epsilon_2 \frac{\partial u_2}{\partial n} \quad \text{on } \Gamma \tag{4.58}$$

where the subscripts refer to the two subdomains Ω_1 and Ω_2 sharing the material boundary Γ, and n is the normal direction to Γ.

4.7.3 Construction of Variational FLAME

The basic setup for the variational version of FLAME is the same as for Trefftz–FLAME (Section 4.3.1, p. 198). The computational domain is covered by a set of overlapping patches: $\Omega = \cup\Omega^{(i)}$, $i = 1, 2, \ldots n$. There is a local approximation space $\Psi^{(i)}$ within each patch $\Omega^{(i)}$

$$\Psi^{(i)} = \text{span}\{\psi_\alpha^{(i)}, \quad \alpha = 1, 2, \ldots, m(i)\}$$

and a multivalued approximation – i.e. a collection of patch-wise approximations $\{\cup u_h^{(i)}\}$. Convergence in this framework (for $h \to 0$) is understood either in the nodal norm as $\|u_h - \mathcal{N}u\|_{E^n} \to 0$ or, alternatively, in the Sobolev norm as $(\sum_i \left\| u_h^{(i)} - u \right\|^2_{H^1(\Omega^{(i)})})^{1/2} \to 0$. As elsewhere in the chapter, the underscore signs denote column vectors.

The next ingredient in the variational formulation is a set of linear test functionals that will be denoted with primes:

$$\{\psi^{(i)'}\}, \quad \omega^{(i)} \equiv \mathrm{supp}(\psi^{(i)'}) \subset \Omega^{(i)}, \quad i = 1, 2, \ldots n \tag{4.59}$$

Simply put, this means that $\psi^{(i)'}(f)$ for any (sufficiently smooth) function f is completely unaffected by the values of f outside $\Omega^{(i)}$, *including the boundary of* $\Omega^{(i)}$. (The italicized portion of this statement is due to the fact that support $\mathrm{supp}(\psi^{(i)'})$ is, by its mathematical definition, a closed set, whereas domain $\Omega^{(i)}$ is open.) Thus a possible discontinuity of the local approximation $u_h^{(i)}$ at the patch boundary is unimportant. The local solution within the i-th patch is a linear combination of the chosen basis functions:

$$u_h^{(i)} \;=\; \sum_{\alpha=1}^{m(i)} c_\alpha^{(i)} \psi_\alpha^{(i)} \;=\; \underline{c}^{(i)T} \underline{\psi}^{(i)} \;\in\; \Psi^{(i)} \tag{4.60}$$

where $\underline{c}^{(i)}$, $\underline{\psi}^{(i)}$ are viewed as column vectors, with their individual entries marked with subscript α. In the variational formulation, the discrete system of equations is obtained by applying the chosen linear test functionals to the differential equation:

$$[u_h^{(i)},\ \psi^{(i)'}] \;=\; \langle f, \psi^{(i)'} \rangle \tag{4.61}$$

or equivalently

$$[\underline{c}^{(i)T} \underline{\psi}^{(i)},\ \psi^{(i)'}] \;=\; \langle f,\ \psi^{(i)'} \rangle \tag{4.62}$$

where $[u,\ \psi^{(i)'}]$ and $\langle f,\ \psi^{(i)'} \rangle$ are alternative notations for $\psi^{(i)'}(Lu)$ and $\psi^{(i)'}(f)$, respectively.[16]

This equation is in terms of the expansion coefficients c of (4.12), (4.60). To obtain the actual difference scheme in terms of *the nodal values*, one needs to relate the coefficient vector $\underline{c}^{(i)} \equiv \{c_\alpha^{(i)}\} \in \mathbb{R}^m$ of expansion (4.60) to the vector $\underline{u}^{(i)} \in r^M$ of the nodal values of $u_h^{(i)}$ on stencil $\#i$. (The superscript (i) for M and m has been dropped for simplicity of notation.) The transformation matrix $N^{(i)}$, with M rows and m columns, was defined above.

If $M = m$ and $N^{(i)}$ is nonsingular,

$$\underline{c}^{(i)} \;=\; (N^{(i)})^{-1} \underline{u}^{(i)} \tag{4.63}$$

and equation (4.62) becomes

$$[\underline{u}^{(i)T} (N^{(i)})^{-T} \underline{\psi},\ \psi^{(i)'}] \;=\; \langle f,\ \psi^{(i)'} \rangle \tag{4.64}$$

(It is implied that the functional $[\cdot\,,\,\cdot]$ in the left hand side is applied to the column vector $\{\underline{\psi}^{(i)}\}$ entry-wise.) Then (4.62) or (4.64) can equally well be written as

$$\underline{u}^{(i)T} (N^{(i)})^{-T} [\underline{\psi}^{(i)},\ \psi^{(i)'}] \;=\; \langle f, \psi^{(i)'} \rangle \tag{4.65}$$

[16] $[u,\ \psi^{(i)'}]$ should not be construed as an inner product of two functions because $\psi^{(i)'}$ is a linear functional rather than a function in the same space as u. I thank S. Prudhomme for taking a note of this.

Equivalently, one may note that matrix N governs the transformation from the original basis $\{\psi_\alpha^{(i)}\}$ in $\Psi^{(i)}$ to the nodal basis $\{\underline{\psi}_{\text{nodal}}^{(i)}\}$ such that $\psi_{\alpha\beta,\text{nodal}}^{(i)}(r_\beta) = \delta_{\alpha\beta}$. Indeed, two equivalent representations of $u_h^{(i)}$ in the original and nodal bases

$$u_h^{(i)} \;=\; \underline{u}^{(i)T}\underline{\psi}_{\text{nodal}}^{(i)} \;=\; \underline{c}^{(i)T}\underline{\psi}^{(i)} \tag{4.66}$$

yield, together with (4.63),

$$\underline{\psi}_{\text{nodal}}^{(i)} \;=\; (N^{(i)})^{-T}\underline{\psi}^{(i)} \tag{4.67}$$

which reveals that (4.62) is in fact

$$\underline{u}^{(i)T}[\underline{\psi}_{\text{nodal}}^{(i)},\, \psi^{(i)'}] \;=\; \langle f, \psi^{(i)'}\rangle \tag{4.68}$$

Expressions (4.64) and (4.68) are equivalent but suggest two different algorithmic implementations of the difference scheme. According to (4.64), one can first compute the Euclidean vector of inner products $\underline{\zeta}^{(i)} = [\underline{\psi}^{(i)},\, \psi^{(i)'}]$ and the difference scheme then is $(N^{(i)})^{-T}\underline{\zeta}^{(i)}$. Alternatively, according to (4.68), one first computes the nodal basis (4.67) and then the products $[\underline{\psi}_{\text{nodal}}^{(i)},\, \psi^{(i)'}]$.

The algorithm for generating variational-difference schemes for an equation $Lu = f$ can be summarized as follows (for $M = m$ and nonsingular $N^{(i)}$):

1. For a given node, choose a stencil, a set of local approximating functions $\{\psi\}$, and a test functional ψ'.
2. Calculate the values of the ψ's at the nodes and combine these values into the N matrix (4.14).
3. Solve the system with matrix N^T and the r.h.s. ψ to get the nodal basis.
4. Compute the coefficients of the difference scheme as

$$[\psi_{\text{nodal}},\, \psi'] \;\equiv\; (L\psi_{\text{nodal}},\, \psi')$$

Alternatively, stages 3) and 4) can be switched:

3′. Compute the values $[\psi,\, \psi'] \equiv (L\psi,\, \psi')$.

4′. Solve the system with matrix N^T and the r.h.s. $[\psi, \psi']$ to obtain the coefficients of the difference scheme.

Note that the r.h.s. of the system of equations involves *functions* $\{\underline{\psi}_{\text{nodal}}\}$ in the first version of the algorithm and *numbers* $[\underline{\psi},\, \psi']$ in the second version. While working with numbers is easier, the nodal functions can be useful and may be reused for different test functionals.

Variational-difference schemes (4.64) and (4.68) are consistent essentially by construction [Tsu04b] (see also Section 4.3.5 for related proofs).

Graphically, the procedure can be viewed as a "machine" for generating variational-difference FLAME schemes (Fig. 4.12).

Remark 11. With this generic setup, no blanket claim of convergence of the variational scheme can be made. The difference scheme is consistent by construction [Tsu04b] but its stability needs to be examined in each particular case.

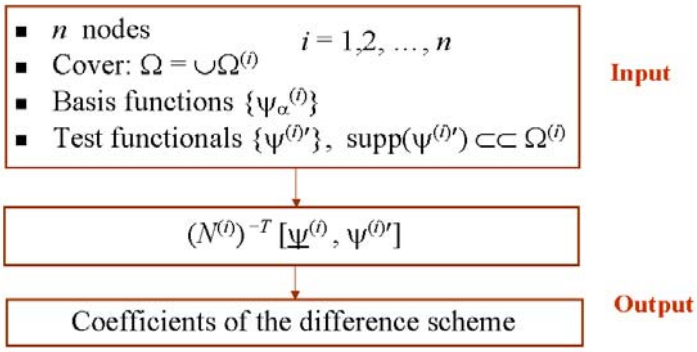


Fig. 4.12. A "machine" for variational-difference FLAME schemes. (Reprinted by permission from [Tsu04b] ©2004 IEEE.)

Remark 12. Implementation of (4.64) or (4.68) implies solving a small system of linear equations whose dimension is equal to the stencil size.

Volume integration in (4.64) is avoided if the test functional is taken to be either the Dirac delta or, alternatively, the characteristic ("window") function $\Pi(\omega^{(i)})$ of some domain $\omega^{(i)} \subset \Omega^{(i)}$: that is, $\Pi(\omega^{(i)}) = 1$ inside $\omega^{(i)}$ and zero outside. With the "window" function, one arrives at a control volume (flux balance) scheme with surface integration. (Typically, $\omega^{(i)}$ is the same size as a grid cell but centered at a node.) The computational cost is asymptotically proportional to the number of grid nodes but depends on the numerical quadratures used to compute the surface fluxes.

4.7.4 Summary of the Variational-Difference Setup

The setup of variational FLAME schemes can be summarized as follows:

- A system of overlapping patches is introduced.
- Desired approximating functions are used within each patch, independently of other patches.
- Simple regular grids can be used.
- When patches overlap, the approximation is generally multivalued (as is also the case in standard FD analysis).
- The nodal solution on the grid is single-valued and provides the necessary "information transfer" between the overlapping patches.
- Since a unique globally continuous interpolant is not defined, the Galerkin method in $H^1(\Omega)$ is generally not applicable. However, within each patch there is a sufficiently smooth local approximation (4.12), and a general moment (weighted residual) method can be applied, provided that the support of the test function is contained entirely within the patch.

In particular, by introducing the standard "control volume" box centered at a given node of the grid and setting the test function equal to one within that control volume and zero elsewhere, one arrives at a flux balance scheme. This is a generalization of the standard "control volume" technique to any set of suitably defined local approximating functions. Only surface integrals, rather than volume quadratures, are needed, which greatly reduces the computational overhead.

Application examples of the variational-difference version of FLAME are given in [Tsu04b]. We now turn to the non-variational version that in many respects is more appealing.

4.8 Appendix: Coefficients of the 9-Point Trefftz–FLAME Scheme for the Wave Equation in Free Space

The mesh size h is for simplicity assumed to be the same in both x- and y-directions. A 3×3 stencil is used. The eight Trefftz–FLAME basis functions are taken as plane waves in eight directions of propagation (toward the central node of the stencil from each of the other nodes).

$$\psi_\alpha \;=\; \exp(\mathrm{i}k\,\hat{\mathbf{r}}_\alpha \cdot \mathbf{r}), \qquad \alpha = 1, 2, \ldots, 8, \quad k^2 \;=\; \omega^2 \mu_0 \epsilon_0 \tag{4.69}$$

The origin of the coordinate system in this case is placed at the midpoint of the stencil and $\hat{\mathbf{r}}_\alpha$ is the unit vector in the direction toward the respective node of the stencil, i.e.

$$\hat{\mathbf{r}}_\alpha \;=\; \hat{x}\cos\frac{\alpha\pi}{4} \;+\; \hat{y}\sin\frac{\alpha\pi}{4}, \qquad \alpha = 1, 2, \ldots, 8 \tag{4.70}$$

The 9×8 nodal matrix (4.14) of FLAME comprises the values of the chosen basis functions at the stencil nodes, i.e.

$$N_{\beta\alpha} \;=\; \psi_\alpha(\mathbf{r}_\beta) \;=\; \exp(ik\,\hat{\mathbf{r}}_\alpha \cdot \mathbf{r}_\beta) \quad \alpha = 1, 2, \ldots, 8; \;\; \beta = 1, 2, \ldots, 9 \tag{4.71}$$

The coefficients of the Trefftz–FLAME scheme (4.20) are obtained by symbolic algebra as the null vector of N^T. As noted by F. Čajko [vT07], care should be exercised to avoid cancelation errors when the coefficients are computed numerically, as their accuracy should be commensurate with the high order of the scheme. The algebraic expressions for the coefficients are as follows.

For the central node:

$$s_1 \;=\; \frac{(e_{\frac{1}{2}} + 1)(e_{\frac{1}{2}} e_1 \;+\; 2e_{\frac{1}{2}} e_0 \;-\; 4e_{-\frac{1}{2}} e_1 \;+\; e_{\frac{1}{2}} \;-\; 4e_{-\frac{1}{2}} \;+\; e_1 \;+\; 2e_0 + 1)}{(e_0 \;-\; 1)^2 (e_{-\frac{1}{2}} \;-\; 1)^4}$$

For the four mid-edge nodes:

$$s_{2-5} \;=\; -\,\frac{e_{\frac{3}{2}}e_0 \;-\; 2e_{\frac{1}{2}}e_1 \;+\; 2e_{\frac{1}{2}}e_0 \;-\; 2e_{\frac{1}{2}} + e_0}{(e_0 \;-\; 1)^2(e_{-\frac{1}{2}} \;-\; 1)^4}$$

For the four corner nodes:

$$s_{6-9} \;=\; \frac{e_{-\frac{1}{2}}(2e_{\frac{1}{2}}e_0 \;-\; e_{-\frac{1}{2}}e_1 \;-\; 2e_{-\frac{1}{2}}e_0 \;-\; e_{-\frac{1}{2}} \;+\; 2e_0)}{(e_0 \;-\; 1)^2(e_{-\frac{1}{2}} \;-\; 1)^4}$$

where $e_\gamma \;=\; \exp(2^\gamma ihk)$, $\gamma = -\frac{1}{2}, 0, \frac{1}{2}, 1, \frac{3}{2}$.

4.9 Appendix: the Fréchet Derivative

In regular calculus, derivatives are used to linearize functions of real or complex variables locally: $f(x + \Delta x) - f(x) \approx f'(x)\Delta x$. More precisely,

$$f(x + \Delta x) \;-\; f(x) \;=\; f'(x)\Delta x \;+\; \delta(x, \Delta x) \tag{4.72}$$

where the residual term δ is small, in the sense that

$$\lim_{|\Delta x| \to 0} \frac{|\delta(x, \Delta x)|}{|\Delta x|} \;=\; 0 \tag{4.73}$$

In functional analysis, this definition is generalized substantially to give a local approximation of a nonlinear operator with a linear one. This leads to the notion of the *Fréchet derivative* in normed linear spaces; the absolute values in (4.73) are replaced with norms.

A formal account of this local linearization procedure in its general form, with rigorous definitions and proofs, can be found in any text on mathematical analysis. This Appendix gives a semi-formal illustration of the Fréchet derivative for the case that will be of most interest in Chapter 6 – the Poisson–Boltzmann operator. In a slightly simplified form, this operator is

$$Lu \;\equiv\; \epsilon\nabla^2 u \;-\; a\sinh(bu) \tag{4.74}$$

where u, by its physical meaning, is the electrostatic potential in an electrolyte with dielectric permittivity ϵ; a and b are known physical constants.

Let us give u a small increment Δu (for brevity of notation, dependence of the potential and its increment on coordinates is not explicitly indicated) and examine the respective increment of Lu:

$$\Delta(Lu) \;\equiv\; L(u + \Delta u) \;-\; Lu \;=\; \epsilon\nabla^2\Delta u \;-\; a[\sinh(b(u + \Delta u)) - \sinh(bu)]$$

Linearizing the hyperbolic sine, one obtains

$$\Delta(Lu) \;=\; \epsilon\nabla^2\Delta u \;-\; ab\cosh(bu)\,\Delta u \;+\; \delta(u, \Delta u)$$

Hence, up to first order terms in Δu,

$$\Delta(Lu) \;\approx\; L'(u)\Delta u$$

where the Fréchet derivative L' is the linear operator

$$L'(u) \;=\; \epsilon\nabla^2 \;-\; ab\cosh(bu)\cdot$$

5

Long-Range Interactions in Free Space

5.1 Long-Range Particle Interactions in a Homogeneous Medium

Computation of long-range forces between multiple charged, polarized and/or magnetized particles is critical in a variety of molecular and nanoscale applications: analysis of macromolecules and nanoparticles, ferrofluids, ionic crystals; in micromagnetics and magnetic recording, etc.

There is a substantial difference between problems with *known* and *unknown* values of charges or dipoles. For example, charges of ions in an ionic crystal and charges of colloidal particles can often be assumed known and fixed. On the other hand, the dipole moments of polarizable particles depend on the external field and therefore are in general unknown *a priori*.

Furthermore, the particles (charges or dipoles) may interact in a homogeneous or in an inhomogeneous medium. The inhomogeneous (and especially nonlinear) case is substantially more complicated and will be discussed in Chapter 6.

This chapter is concerned exclusively with problems where the charges or dipoles are known and the medium is linear homogeneous (free space being the obvious particular case). Even though this case is simpler than problems with unknown polarization of particles and with inhomogeneous media, the computational challenges are still formidable.

Any macroscopic volume contains an astronomical number of particles (Avogadro's number is $\sim 6.022 \times 10^{23}$ particles per mole of any substance). "Brute force" modeling of such enormous systems is obviously not feasible in any foreseeable future. Therefore one cannot help but restrict the simulation to a computational cell containing a relatively small number of particles (typically from hundreds to tens of thousands), with the assumption that the results are representative of the behavior of a larger volume of the material.

A new question immediately arises, however, once the simulation has been limited to a finite cell. To find the electrostatic (or in some cases magnetostatic) field within the cell, one needs to set *boundary conditions* on its surface.

Clearly, the actual boundary values of the field or potential are not known, as they depend on the distribution of *all* sources, including those outside the computational cell. The most common approximation is to impose *periodic conditions* by replicating the cell in all three directions. The whole space is then filled with identical cells, as schematically illustrated in Fig. 5.1.

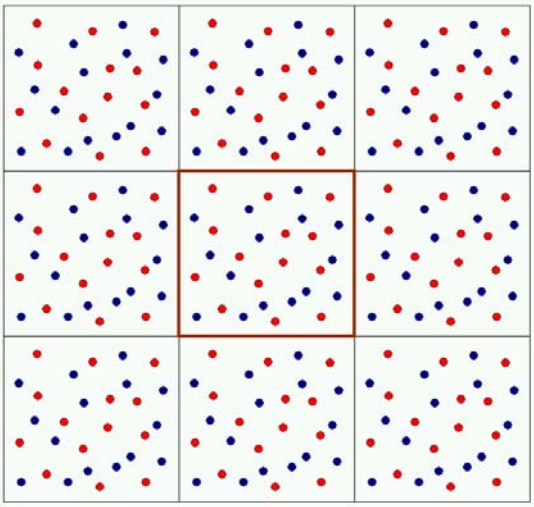


Fig. 5.1. A schematic illustration of the electrostatic problem with periodic conditions.

The obvious geometric restriction is that the cell has to have a space-filling shape such as a parallelepiped (rectangular, monoclinic or triclinic) or a truncated octahedron.[1] The latter is indeed used in some molecular dynamics simulations, as its shape is closer to spherical symmetry than that of a parallelepiped. For simplicity, however, we shall limit our discussion to the rectangular parallelepiped, keeping in mind that most computational methods considered in this chapter can be generalized to more complex shapes of the cell. Furthermore, we shall consider only *charges*, not dipoles; for dipole interactions, see e.g. S.W. de Leeuw *et al.* [dLPS86] and Z. Wang & C. Holm [WH01].

With infinitely many cells filling the whole space and infinitely many particles, it is clearly impossible to compute energy and forces by straightforward numerical summation. Even if the number of particles N were finite but large, direct summation of all pairwise energies, while theoretically possible, would

[1] See e.g. Eric W. Weisstein. "Truncated Octahedron." From MathWorld – A Wolfram Web Resource. http://mathworld.wolfram.com/TruncatedOctahedron.html

not be computationally efficient, as the number of operations θ is asymptotically proportional to N^2. Special techniques are therefore required.

The main features of the problem that will be considered in this chapter can be summarized as follows:

1. Charges q_i $(i = 1, 2, \ldots, N)$ are given. Their locations $\mathbf{r}_i = (x_i, y_i, z_i)$ within a rectangular parallelepiped with dimensions $L_x \times L_y \times L_z$ are also known.
2. The system of charges is electrically neutral, i.e. $\sum_{i=1}^{N} q_i = 0$.
3. The medium is homogeneous (free space is a particular case).
4. The boundary conditions are set as periodic: each charge at $\mathbf{r}_i = (x_i, y_i, z_i)$ has infinitely many identical images at $(x_i + n_x L_x,\ y_i + n_y L_y,\ z_i + n_z L_z)$, where $n_{x,y,z}$ are integers.

Violation of the neutrality condition – that is, a nonzero value of the total (or, equivalently, the average) charge in the computational cell – would lead, due to the periodic boundary conditions, to the nonzero average charge density throughout the infinite space, which does not give rise to mathematically or physically meaningful fields.

The goal is to compute energy and forces acting on the particles, at as low computational cost as possible. In the asymptotic sense at least, the number of operations θ growing as $\sim CN^2$, with some numerical factor C, is as a rule not acceptable, and one is looking for ways to reduce it as close as possible to the optimal $\theta \sim CN$ level. Of course, in the comparison of methods with the same asymptotic behavior, the magnitude of the C prefactor becomes important.

When the focus of the analysis is on the *asymptotic* behavior and not on the prefactor, the "big-oh" notation is very common and useful:

$$\theta \;=\; \mathcal{O}(N^\gamma) \quad \Longleftrightarrow \quad C_1 N^\gamma \;\leq\; \theta \;\leq\; C_2 N^\gamma \tag{5.1}$$

where $C_{1,2}$ are positive constants independent of N and γ is a parameter. (See also Introduction, p. 7.)

At least two classes of methods with close to optimal asymptotic number of arithmetic operations per particle are known. The first one – the summation method introduced in 1921 by P. Ewald [Ewa21] – is the main subject of this chapter.

The second alternative is Fast Multipole methods (FMM) by L. Greengard & V. Rokhlin [GR87b, CGR99]. The key idea to speed up multiparticle field computation by clustering the particles hierarchically can be traced back to the tree codes developed in the 1980s by J. Barnes & P. Hut [BH86] and to the algorithm by A.W. Appel [App85]. For 2D, FMM was developed by Greengard & Rokhlin [GR87b] and independently by L.L. van Dommelen & E.A. Rundensteiner [vDR89]. The 2D case is simplified by the availability of tools of complex analysis; 3D algorithms are much more involved and were perfected in the 1990s by Greengard & Rokhlin [GR97, CGR99, BG].

In FMM, the particles are clustered hierarchically; interactions between remote clusters can be computed with any desired level of accuracy via multipole expansions (truncated to a finite number of terms); this idea, when applied recursively, reduces the computational cost dramatically – from $\mathcal{O}(N^2)$ to the asymptotically optimal value $\mathcal{O}(N)$.

Many versions, modifications and implementations of the Greengard-Rokhlin FMM now exist. A very helpful and concise tutorial by R. Beatson & L. Greengard is available online [BG]. Notably, the operation count for the "classic" version of FMM in 3D is, according to [BG], approximately $150Np^2$, where p is the highest order of multipole moments retained in the expansion. (The numerical error decreases exponentially as p increases.) An improved version of FMM reduces the operation count to $\sim 270Np^{3/2} + 2Np^2$. Finally, a new algorithm combining multipole expansions with plane wave expansions[2] requires about $200Np + 3.5Np^2$ operations [BG]. The multipole/exponential expansion is described by T. Hrycak & V. Rokhlin [HR98b] for 2D and by Greengard & Rokhlin for 3D.

Implementation of FMM *for periodic boundary conditions* requires additional care. This case is discussed in Greengard & Rokhlin's 1987 paper [GR87b] (Section 4.1), in Greengard's dissertation [Gre87], and for 3D in more detail by K.E. Schmidt & M.A. Lee [SL91]. For more recent developments, see F. Figueirido *et al.* [FLZB97] and Z.H. Duan & R. Krasny [DK00, DK01].

The FMM works best, and has almost optimal operation count, if (i) the domain is unbounded; (ii) material characteristics are linear and homogeneous; (iii) the dipole moments (or charges) are known *a priori*; (iv) the number of particles is very large (on the order of 10^4 or higher). If these conditions are not fully satisfied, the FMM is less efficient. However, even when the situation is ideal for FMM, its algorithmic implementation is quite involved, and the large numerical prefactor in the operation count reduces the computational efficiency. In addition, in Molecular Dynamics (MD) simulations a fairly large number of terms (eight or more) have to be retained in the multipole expansion to avoid appreciable numerical violation of energy conservation laws [BSS97]. Due to this combination of circumstances, Ewald summation algorithms are still more popular in MD than FMM.

5.2 Real and Reciprocal Lattices

It is standard in solid state physics to characterize the computational cell geometrically by its three axis vectors $\mathbf{L}_1 = L_1\hat{l}_1$, $\mathbf{L}_2 = L_2\hat{l}_2$, $\mathbf{L}_3 = L_3\hat{l}_3$, where $\hat{l}_{1-3}$ are unit vectors. These vectors are not necessarily orthogonal, although

[2] The plane wave expansion of the (static!) Coulomb potential is counterintuitive but nevertheless efficient, due to the simple translation properties of plane waves. In 2D, the exponential representation comes from the obvious integration formula $(z - z_0)^{-1} = \int_0^\infty \exp(-x(z - z_0))\, dx$ for any complex z, z_0 with $\mathrm{Re}(z - z_0) > 0$ [HR98b]. The integral is then approximated by numerical quadratures.

in subsequent sections for simplicity we assume that they are. In the case of ionic crystals, the computational box may correspond to the Wigner–Seitz cell.

The real lattice $\mathcal{L}$ is defined as a set of vectors (or equivalently, points) $\mathbf{R} = n_1\mathbf{L}_1 + n_2\mathbf{L}_2 + n_3\mathbf{L}_3$ for all integers n_1, n_2, n_3.

It is also standard in solid state physics and crystallography to define the reciprocal lattice $\mathcal{K}$ of vectors $\mathbf{k}$ such that

$$\exp(\mathrm{i}\mathbf{R}\cdot\mathbf{k}) \;=\; 1 \tag{5.2}$$

The reciprocal lattice $\mathcal{K}$ is spanned by three vectors

$$\mathbf{k}_1 \;=\; 2\pi\,\frac{\mathbf{L}_2\times\mathbf{L}_3}{\mathbf{L}_1\cdot\mathbf{L}_2\times\mathbf{L}_3} \tag{5.3}$$

$$\mathbf{k}_2 \;=\; 2\pi\,\frac{\mathbf{L}_3\times\mathbf{L}_1}{\mathbf{L}_1\cdot\mathbf{L}_2\times\mathbf{L}_3} \tag{5.4}$$

$$\mathbf{k}_3 \;=\; 2\pi\,\frac{\mathbf{L}_1\times\mathbf{L}_2}{\mathbf{L}_1\cdot\mathbf{L}_2\times\mathbf{L}_3} \tag{5.5}$$

so that any reciprocal lattice vector $\mathbf{k} = m_1\mathbf{k}_1 + m_2\mathbf{k}_2 + m_3\mathbf{k}_3$ for some integers m_1, m_2, m_3.

Transformations between real and reciprocal (i.e. Fourier) spaces are key in the analysis. The Fourier series representation of a function $f(\mathbf{r})$ is

$$f(\mathbf{r}) \;=\; \sum_{\mathbf{k}\in\mathcal{K}} \hat{f}(\mathbf{k})\exp(\mathrm{i}\mathbf{k}\cdot\mathbf{r}) \tag{5.6}$$

with

$$\hat{f}(\mathbf{k}) \;=\; \frac{1}{V}\int_{\mathrm{cell}} f(\mathbf{r})\exp(-\mathrm{i}\mathbf{k}\cdot\mathbf{r})\,dV \tag{5.7}$$

In the remainder, the lattices vectors will be assumed orthogonal and directed along the Cartesian axes; therefore subscripts x, y, z will be used instead of 1, 2, 3 to denote these lattices vectors.

Further details can be found in textbooks on solid state physics, for example N.W. Ashcroft & N.D. Mermin [AM76].

5.3 Introduction to Ewald Summation

Developed early in the 20th century [Ewa21] as an analytical method for computing electrostatic energy and forces in ionic crystals, the Ewald method became, after the introduction of "Particle–Mesh" methods by R.W. Hockney & J.W. Eastwood in [HE88], a computational algorithm of choice for periodic charge and dipole distributions. Nowadays many versions of Ewald summation exist (see e.g. excellent reviews by C. Sagui & T.A. Darden [SD99] and by M. Deserno & C. Holm [DH98a]).

The main features of the problem were already summarized in the previous section; we now turn to a more rigorous formulation.

An electrically neutral collection of charges $\{q_i\}_{i=1}^N$ is considered in a rectangular box $L_x \times L_y \times L_z$.[3] The charges and their locations $\mathbf{r}_i = (x_i, y_i, z_i)$ are known.[4] Due to the periodic conditions assumed, each charge has infinitely many images at $\mathbf{r}_i + \mathbf{n}.*\mathbf{L}$, where vector $\mathbf{L} = (n_x L_x, n_y L_y, n_z L_z)$, $\mathbf{n} \in \mathbb{Z}^3$ is a 3D index, and n_x, n_y, n_z are arbitrary integers. ($\mathbf{n} = 0$ corresponds to the charge itself.) Here, and occasionally elsewhere in this chapter, I adopt Matlab-style notation for entry-wise multiplication of vectors: $\mathbf{n}.*\mathbf{L} \equiv (n_x L_x, n_y L_y, n_z L_z)$.[5]

One would think that the electrostatic potential can easily be written out as a superposition of Coulomb potentials of all charges (including images):

$$u(\mathbf{r}) \;=\; \frac{1}{4\pi\epsilon} \sum_{\mathbf{n}\in\mathbb{Z}^3} \sum_{i=1}^{N} \frac{q_i}{|\mathbf{r} - \mathbf{r}_i + \mathbf{n}.*\mathbf{L}|} \qquad \text{(to be clarified)} \tag{5.8}$$

where the SI system of units has been adopted. Similarly, at first glance the expression for electrostatic energy $\mathcal{E}$ is

$$\mathcal{E} \;=\; \frac{1}{4\pi\epsilon}\,\frac{1}{2} \sum_{\mathbf{n}\in\mathbb{Z}^3}^{*} \sum_{1\le i,j\le N} \frac{q_i q_j}{|\mathbf{r}_i - \mathbf{r}_j + \mathbf{n}.*\mathbf{L}|} \qquad \text{(to be clarified)} \tag{5.9}$$

Subscripts i, j refer to two charges in the simulation box. By convention, the asterisk (or in some publications the prime) on top of the summation sign indicates that the singular term with $i = j$ and $n = 0$ is omitted.

Since both potential and energy are expressed via infinite series, the question of convergence (or lack thereof) is critical. The net charge of the computational cell is, by definition of the problem, zero; let the total dipole moment of the cell be $\mathbf{p} = \sum_{q_i \in \text{cell}} q_i \mathbf{r}_i$. Convergence of the series is governed by the asymptotic behavior of its terms as index $n \to \infty$. The contribution of the $\mathbf{n}$-th image of the cell to the potential in the cell is, asymptotically,

$$u(\mathbf{r}, \mathbf{n}) \;\sim\; \frac{p}{|\mathbf{n}.*\mathbf{L}|^2} \;\sim\; \frac{p}{n^2} \tag{5.10}$$

At the same time, the number of periodic images corresponding to the same $\mathbf{n}$ is asymptotically proportional to n^2 (to see that, assume for simplicity that all three dimensions of the cell are scaled to unity and picture the n-th layer of images of the cell as approximately a spherical shell of volume $4\pi n^2$). This

[3] For simplicity, we shall not consider more complex box shapes such as triclinic or truncated octahedral, even though their treatment in Ewald algorithms is similar.

[4] In Molecular Dynamics, at each time step particles assume different positions and the Ewald method is then applied to update the energy and forces at that step.

[5] In mathematics, such entry-wise multiplication is known as Hadamard product; see R.A. Horn & C.R. Johnson [HJ94].

means that the n-th layer of images contributes on the order of n^2 terms, each of which *by the absolute value* is on the order of n^{-2}. Consequently, the series for the electrostatic potential does *not converge absolutely.*[6]

If the dipole moment of the cell happens to be zero (which in practice can only be assured under special symmetry conditions), the rate of decay of the terms in the series will be dictated by the next surviving multipole moment (e.g. quadrupole, if it is nonzero). Then the series *will* converge absolutely due to the faster rate of decay of its terms. Absolute convergence substantially simplifies the analysis and, among other things, makes it legal to change the order of summation in the infinite series.

In the general, and most interesting, case of a nonzero dipole moment, the sum of the series for both potential and energy depends on the order of summation of its terms. That is, expressions (5.8), (5.9) are not even rigorously defined until the order of summation is specified. The value of the potential thus depends on which charge contributes to the total field "first," which one contributes "second," etc. This is unacceptable on physical grounds and, in addition, quite bizarre mathematically due to the Riemann rearrangement theorem. This theorem states that the terms of any conditionally convergent series can be rearranged to obtain any preassigned sum from $-\infty$ to ∞ (inclusive) – that is, any value of potential and energy could be obtained by summing up the contributions in a certain order.

The cause of this nonphysical result is the artificial infinite and perfectly periodic structure that has been assumed. In contrast, the potential of a *finite* system of charges is well defined. An infinite system is nonphysical at least in some respects, so it is not a complete surprise that paradoxes do arise. A mathematically rigorous way to define and analyze the infinite periodic system is to start with a finite one and then let its size tend to infinity. The conditional convergence then manifests itself in a clear way: the total potential, and thus the field, depend on the overall geometric shape of the body [Smi81, dLPS80a, dLPS80b, SD99, DH98a] and on the conditions on its boundary. This shape dependence does not disappear even if the boundary is moved far away.

An accurate mathematical analysis along these lines was carried out by E.R. Smith [Smi81] (see also de Leeuw *et al.* [dLPS80a, dLPS80b, dLPS86]). Smith considered a finite-size collection of particles (e.g. a finite ionic crystal) as built of a number of layers of cells around a "master" cell and computed the electrostatic energy (per unit cell) for the progressively increasing number of such layers, with the shape of the body remaining fixed. This problem is mathematically valid and well-posed, and Smith's final result does contain a term depending on the shape of the body and also on the dielectric constant of the surrounding medium.

[6] Recall that a series is called absolutely convergent if the series of absolute values converges. Otherwise a convergent series is called *conditionally convergent.*

A physical explanation of this shape dependence is not complicated. Indeed, a body containing a large number of cells carrying a dipole moment $\mathbf{p}$ can be considered as having average polarization (i.e. dipole moment per unit volume) of approximately $\mathbf{P} = \mathbf{p}/V$, with volume $V = L_x L_y L_z$. It is well known from electrostatics that the corresponding equivalent charge density on the surface of the body is $\rho_S = \mathbf{P} \cdot \hat{\mathbf{n}}$, where $\hat{n}$ is the outward unit normal to the surface. This surface charge creates an additional field and contributes to the energy of the system; this contribution does not diminish even if the size of the surface tends to infinity.[7]

In the following section, we view the computation of the electrostatic potential as a boundary value problem. This treatment is instructive and, generally speaking, standard in electrostatics; yet it is uncommon in the studies of Ewald methods.

5.3.1 A Boundary Value Problem for Charge Interactions

Let the computational cell be a rectangular parallelepiped $\Omega = [0, L_x] \times [0, L_y] \times [0, L_z]$. The governing electrostatic equation for the electrostatic potential is

$$Lu \equiv -\nabla^2 u = \frac{\rho_\delta}{\epsilon} \quad \text{in } \Omega = [0, L_x] \times [0, L_y] \times [0, L_z] \tag{5.11}$$

where the density of point charges ρ_δ can be written via Dirac δ-functions:

$$\rho_\delta = \sum_{i=1}^{N} q_i \delta(\mathbf{r} - \mathbf{r_i}) \tag{5.12}$$

Periodic boundary conditions are assumed:

$$u(0, y, z) = u(L_x, y, z); \qquad \frac{\partial u(0, y, z)}{\partial x} = \frac{\partial u(L_x, y, z)}{\partial x} \tag{5.13}$$

and similar conditions on the other two pairs of faces. In addition, to eliminate an additive constant in the potential, the zero mean is imposed:

$$\int_\Omega u \, d\Omega = 0 \tag{5.14}$$

It is not difficult to prove that the solution of this boundary value problem is unique. Indeed, if there are two solutions of (5.11)–(5.14), u_1 and u_2, then their difference $v \equiv u_1 - u_2$ satisfies the Laplace equation in Ω as well as the periodic boundary conditions. The Fourier series expansion of v is

[7] In addition, if the surrounding medium outside the body is a dielectric, it will also be polarized and will in general affect the equivalent surface charge density and the overall field and energy.

$$v(\mathbf{r}) \;=\; \sum_{\mathbf{k}\in\mathcal{K}} \tilde{v}(\mathbf{k}) \exp(\mathrm{i}\,\mathbf{k}\cdot\mathbf{r}) \tag{5.15}$$

and the periodic boundary conditions ensure that the second derivative of v exists as a regular periodic function, not just a distribution, in the whole space. (See Appendix 6.15, p. 343, for information on distributions.) Then the Laplace operator in the Fourier space amounts just to multiplication with $-k^2$, so the fact that v satisfies the Laplace equation implies $k^2\tilde{v}(\mathbf{k}) = 0$. Hence all Fourier coefficients $\tilde{v}(\mathbf{k})$ for $k \neq 0$ are zero; but $\tilde{v}(\mathbf{0}) = 0$ as well due to the zero mean condition for v. Since all Fourier coefficients of v are zero, $v = 0$ and $u_1 = u_2$.

Further, not only is the solution unique but also problem (5.11) – (5.14) is well-posed. This can be stated more precisely in several different ways. Let us, for example, examine the minimum eigenvalue of the problem

$$Lu \;=\; \lambda u \tag{5.16}$$

(with periodic boundary conditions and the zero-mean constraint in place). If u is an eigenfunction of this problem, then

$$(Lu, u) \;=\; \lambda(u, u) \tag{5.17}$$

where $(\cdot\,,\,\cdot)$ is the standard complex L_2 inner product, i.e.

$$(u, v) \;\equiv\; \int_\Omega uv^*\, d\Omega \tag{5.18}$$

where v^* is the complex conjugate of v. Using Parseval's identity, one can equally well compute the inner products in Fourier space and rewrite (5.17) as

$$\sum\nolimits_{\mathbf{k}\in\mathcal{K};k\neq 0} k^2|\tilde{u}(\mathbf{k})|^2 \;=\; \lambda \sum\nolimits_{\mathbf{k}\in\mathcal{K};k\neq 0} |\tilde{u}(\mathbf{k})|^2 \tag{5.19}$$

where $\tilde{u}(\mathbf{k})$ are the Fourier coefficients of u.

Since $k \geq 1$, it is clear from the expression above that

$$\lambda \;\geq\; k_{\min}^2 \;=\; \left(\frac{2\pi}{L_x}\right)^2 + \left(\frac{2\pi}{L_y}\right)^2 + \left(\frac{2\pi}{L_z}\right)^2 \tag{5.20}$$

This boundedness of the minimum eigenvalue shows that the problem is indeed well-posed.

The Fourier Transform of the point charge density will be needed very frequently:

$$\mathcal{F}\{\rho_\delta\}(\mathbf{k}) \;=\; \frac{1}{V}\,\tilde{\rho}(\mathbf{k}) \;=\; \frac{1}{V}\sum_{i=1}^{N} q_i \exp(-\mathrm{i}\,\mathbf{k}\cdot\mathbf{r}_i) \tag{5.21}$$

In solid state physics and crystallography, coefficients $\tilde{\rho}(\mathbf{k})$ are known as *structure factors* and are often denoted with $S(\mathbf{k})$. I shall, however, continue to use

the $\tilde{\rho}$ notation because it underscores the connection with charge density ρ in real space.

With the dot product written out explicitly, the structure factor is

$$\tilde{\rho}(\mathbf{k}) \;=\; \sum_{i=1}^{N} q_i \exp\left(-\mathrm{i}\left(k_x x_i \,+\, k_y y_i \,+\, k_z z_i\right)\right) \tag{5.22}$$

The treatment of Coulomb interactions as a boundary value problem accomplishes several goals:

Well-posedness. The ambiguity related to the shape-dependent term has been removed; the problem is well-posed.

Finite domain. The problem is limited to one finite computational cell – no need to consider infinite sets of images and infinite sums.

Wider selection of methods. Not only FT-based techniques but other methods well established for boundary value problems (e.g. finite differences) become available.

The reader may note an apparent contradiction between the well-posedness of the boundary value problem and the inherent ambiguity of summation of the conditionally convergent infinite series. The following section considers this question in more detail and presents the solution of the Poisson equation via the Ewald series.

5.3.2 A Re-formulation with "Clouds" of Charge

As discussed in Section 5.3, the infinite series (5.8) for the electrostatic potential of the periodic array of cells does not converge absolutely and therefore cannot directly be used for theoretical analysis or practical computation.

As already noted, the rigorous analysis by E.R. Smith [Smi81] involves a *finite* series for the potential and energy of a finite-size body, and passing to the limit as the size of the body increases but its shape is kept the same. The end result can be written as a sum of two absolutely convergent Ewald series (considered in detail below), plus a shape-dependent term. Fixing the shape of the body can be interpreted as specifying the order of summation in the infinite series: the summation is carried out layer-by-layer. Changing the shape leads to a rearrangement of terms in the original conditionally convergent series (5.8) and in general to a different result.

The shape-dependent term is attributable to the field of charges on the surface of the body (e.g. a crystal) due to the polarization of that body. From this physical perspective, it is clear that the periodic conditions (5.13) in the boundary value problem correspond to the case where Smith's shape-dependent term is absent. Thus the periodic conditions represent only a particular case of a more general physical situation; however, the general case can always be recovered by adding the shape-dependent term. It can also be argued [DTP97] that in a real physical system of finite size the surface

charges will tend to rearrange themselves to minimize their contribution to free energy.

In the remainder, we shall therefore disregard the shape-dependent term and focus on the boundary value problem described by the Poisson equation (5.11) with the periodic boundary conditions (5.13) and the zero-mean constraint (5.14). This problem, as noted in the previous section, is well-posed.

The following idea allows one to write the solution via rapidly convergent sums. Intuitively, this idea can be interpreted as splitting up the potential of each point charge into two parts, by adding and subtracting an auxiliary "cloud" of charge (Fig. 5.2), usually with a Gaussian distribution of charge density. In the first subproblem (point charges with clouds), the interactions are short-range due to the screening effect of the clouds; these interactions can therefore be computed directly. The second subproblem (clouds only) does not contain singularities and can be solved, especially for periodic boundary conditions, using Fourier Transforms (FT). A radical improvement is achieved

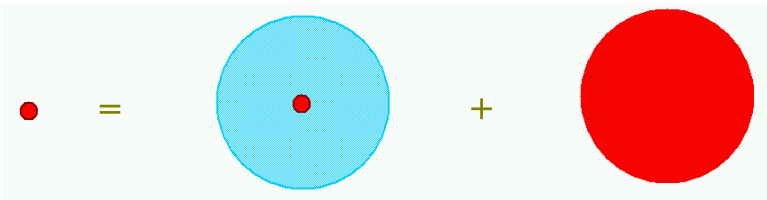


Fig. 5.2. The point charge problem split into two parts. (Reprinted by permission from [Tsu04a] ©2004 IEEE.)

by employing Fast FT on a finite grid, with appropriate charge-to-grid assignment schemes. Fourier Transform ("reciprocal space") methods are so standard that even the conventional terminology reflects it. For example, the notion of reciprocal energy and forces refers to the way these quantities are *computed* (by FT) rather than to what they physically *are* (interactions due to Gaussian clouds).

As a preliminary step in the derivation of Ewald summation methods, let us work out expressions for the field distribution due to a Gaussian cloud of charge.

5.3.3 The Potential of a Gaussian Cloud of Charge

Let the charge density be defined (in the spherical system) as a Gaussian distribution centered (for convenience) at the origin:

$$\rho_{\text{cloud}} = \rho_0 \exp(-\beta^2 r^2) \tag{5.23}$$

where ρ_0 and β are parameters (in Ewald methods, β is called the *Ewald* parameter). Note that this form of charge density is taken for computational convenience, as it is relatively easy to deal analytically with Gaussians.

We first consider a "stand-alone" cloud, with no periodicity, and then turn to the problem with periodic images. The total charge of a single cloud is

$$q \;=\; \int_{\mathbb{R}^3} \rho\, dV \;=\; \rho_0 \int_0^\infty \exp(-\beta^2 r^2)\, 4\pi r^2 dr \;=\; \rho_0\, \frac{\pi^{3/2}}{\beta^3} \tag{5.24}$$

Hence

$$\rho_0 \;=\; q\, \frac{\beta^3}{\pi^{3/2}} \tag{5.25}$$

The field of the cloud can then be found using Gauss's Law of electrostatics: the flux of the $\mathbf{D}$ vector through any closed surface is equal to the total charge inside that surface.

For the Gaussian cloud in a homogeneous dielectric with permittivity ϵ, this yields, in the metric system of units,[8]

$$4\pi r^2 \epsilon E \;=\; \rho_0 \int_0^r \exp(-\beta^2 r^2)\, 4\pi r^2\, dr \;=\; \mathrm{erf}(\beta r) \;-\; 2\frac{\beta}{\sqrt{\pi}}\, r \exp(-\beta^2 r^2)$$

This immediately gives the $\mathbf{E}$ field and then the potential of the Gaussian cloud of charge:

$$u_{\mathrm{cloud}}(r) \;=\; \int_r^\infty E(r)\, dr \;=\; \frac{\mathrm{erf}(\beta r)}{4\pi\epsilon r} \tag{5.26}$$

As a reminder, the error function is defined as

$$\mathrm{erf}(r) \;\equiv\; \frac{2}{\sqrt{\pi}} \int_0^r \exp(-r^2)\, dr \tag{5.27}$$

and the complementary error function

$$\mathrm{erfc}(r) \;\equiv\; 1 \;-\; \mathrm{erf}(r) \tag{5.28}$$

The Taylor expansion of erf around zero is known to be

$$\mathrm{erf}(r) \;=\; \frac{2}{\sqrt{\pi}} \left(r \;-\; \frac{1}{3} r^3 \;+\; \mathrm{h.o.t.} \right), \quad r \ll 1 \tag{5.29}$$

where "h.o.t." stands for "higher order terms" in r. The cloud potential (5.26) at $r = 0$ then is

$$u_{\mathrm{cloud}}(0) \;=\; \frac{\beta}{2\pi^{3/2}\epsilon} \tag{5.30}$$

[8] The usage of the same symbol (in this case, r) as both the dummy integration variable and the integration limit helps to avoid superfluous notation and should not cause any confusion.

Note that the error function tends very rapidly to one when its argument goes to infinity (and simultaneously erfc tends to zero). For example, $\mathrm{erfc}(4) \approx 1.54 \cdot 10^{-8}$, $\mathrm{erfc}(6) \approx 2.16 \cdot 10^{-17}$. Consequently, potential (5.26) of a Gaussian cloud decays as $\sim 1/r$, but the potential of a point charge with a screening cloud of the opposite sign decays extremely quickly with increasing r – as $\mathrm{erfc}(r)/r$).

Next, we shall need the Fourier Transform of the Gaussian charge density and potential. (The main rationale for using FT is that differentiation turns into multiplication by $\mathrm{i}k_{x,y,z}$ in the Fourier domain.) We have to be prepared to deal with multiple charges and the corresponding clouds centered at different locations, so the (slight) simplification that the charge is located at the origin must now be dropped.

The FT of a Gaussian is known to be also a Gaussian. Let us start with 1D for simplicity and consider a Gaussian function centered at x_i:

$$\rho^{(i)}(x) \;=\; \rho_{0x}^{(i)} \exp(-\beta^2 (x - x_i)^2) \tag{5.31}$$

For the time being, ρ_{0x} is just an arbitrary factor; however, we anticipate that in 3D, when combined with similar factors $\rho_{0y}^{(i)}$, $\rho_{0z}^{(i)}$, it will yield the proper normalization constant ρ_0 of (5.25).

The Fourier transform of this Gaussian is

$$\begin{aligned} \mathcal{F}\{\rho^{(i)}\}(k_x) \;&\equiv\; \rho_{0x}^{(i)} \int_{-\infty}^{\infty} \exp(-\beta^2 (x - x_i)^2)\, \exp(-\mathrm{i}k_x x)\, dx \\ &=\; \rho_{0x}^{(i)} \frac{\sqrt{\pi}}{\beta} \exp\left(-\frac{k_x^2}{4\beta^2}\right) \exp(-\mathrm{i}k_x x_i) \end{aligned} \tag{5.32}$$

where k is a Fourier (= reciprocal space) variable and subscript "x" is used in anticipation of y- and z-components of k to be needed later.

5.3.4 The Field of a Periodic System of Clouds

The FT above is for a *stand-alone* Gaussian in the whole space. However, we need to deal with a *periodic system* of Gaussians; what is the FT in this case? More precisely, we define the "periodized" charge density as

$$\mathrm{PER}\{\rho^{(i)}\}(\mathbf{r}) \;\equiv\; \sum_{\mathbf{n} \in \mathbb{Z}^3} \rho^{(i)}(\mathbf{r} - \mathbf{n}. * \mathbf{L}) \tag{5.33}$$

where again $\mathbf{n}. * \mathbf{L}$ is Matlab-style notation for Hadamard-product, i.e. entry-wise multiplication of Euclidean vectors or matrices (see footnote 5 on p. 244).

This charge density is periodic by construction, and hence its Fourier transform is actually a Fourier series, so that

$$\mathrm{PER}\{\rho^{(i)}\}(\mathbf{r}) \;=\; \sum_{\mathbf{k} \in \mathcal{K}} \tilde{\rho}_{\mathrm{PER}}^{(i)}(\mathbf{k}) \exp(\mathrm{i}\mathbf{k} \cdot \mathbf{r}) \tag{5.34}$$

where $\tilde{\rho}^{(i)}_{\mathrm{PER}}$ are the coefficients of the Fourier series.

We can now take advantage of a simple relationship between the discrete FT of a periodic array of clouds and the continuous FT of a single cloud:

$$\mathcal{F}\{\mathrm{PER}\{\rho^{(i)}\}\} \;\equiv\; \tilde{\rho}^{(i)}_{\mathrm{PER}}(\mathbf{k}) \;=\; \frac{1}{V}\,\tilde{\rho}^{(i)}(\mathbf{k}), \quad \mathbf{k} \in \mathcal{K} \tag{5.35}$$

where $V = L_x L_y L_z$ is the volume of the computational cell. This relationship (for 1D, well known in Signal Analysis) is derived in Appendix 5.6.

An explicit expression for this spectrum is obtained by substituting the FT of the stand-alone Gaussian (5.32) into (5.35), for each of the three coordinates x, y, z:

$$\begin{aligned}\mathcal{F}\left\{\mathrm{PER}\{\rho^{(i)}_{\mathrm{cloud}}\}\right\}(\mathbf{k}) \;&=\; \rho^{(i)}_0 \frac{1}{V}\left(\frac{\sqrt{\pi}}{\beta}\right)^3 \exp\left(-\frac{k^2}{4\beta^2}\right)\exp(-\mathrm{i}\mathbf{k}\cdot\mathbf{r}_i)\\ &=\; \frac{q_i}{V}\,\exp\left(-\frac{k^2}{4\beta^2}\right)\exp(-\mathrm{i}\mathbf{k}\cdot\mathbf{r}_i)\end{aligned} \tag{5.36}$$

where vector $\mathbf{k} = (k_{0x}m_x,\, k_{0y}m_y,\, k_{0z}m_z)$, $k_{0x} = 2\pi/L_x$ and similarly for the other two coordinates; $k^2 = k_x^2 + k_y^2 + k_z^2$; $\rho^{(i)}_0$ was defined in (5.25).

Hence the FT of the charge density of *all* clouds is

$$\mathcal{F}\{\rho_{\mathrm{clouds}}\}(\mathbf{k}) \;\equiv\; \frac{1}{V}\,\tilde{\rho}_{\mathrm{clouds}}(\mathbf{k}) \;=\; \frac{1}{V}\sum_{i=1}^{N} q_i \exp(-\mathrm{i}\mathbf{k}\cdot\mathbf{r}_i)\exp\left(-\frac{k^2}{4\beta^2}\right) \tag{5.37}$$

where subscript "clouds" (in plural) implies the collective contribution of all clouds of charge (including their periodic images).

5.3.5 The Ewald Formulas

With the Fourier Transform of the sources now at hand, we can solve the Poisson equation and derive the Ewald summation formulas. The Poisson equation (5.11) in the Fourier domain is extremely simple:

$$k^2\tilde{u}_{\mathrm{clouds}}(\mathbf{k}) \;=\; \frac{\tilde{\rho}_{\mathrm{clouds}}}{\epsilon} \tag{5.38}$$

Hence the electrostatic potential in the Fourier domain is

$$\tilde{u}_{\mathrm{clouds}}(\mathbf{k}) \;=\; \frac{\tilde{\rho}_{\mathrm{clouds}}}{k^2\epsilon V} \;=\; \frac{1}{\epsilon V}\sum_{i=1}^{N} q_i \exp(-\mathrm{i}\mathbf{k}\cdot\mathbf{r}_i)\,\frac{1}{k^2}\,\exp\left(-\frac{k^2}{4\beta^2}\right), \quad k \neq 0 \tag{5.39}$$

for $k = 0$, due to the zero-mean constraint for charges and potentials

$$\tilde{u}_{\mathrm{cloud}}(\mathbf{0}) \;=\; 0 \tag{5.40}$$

The inverse FT of $\tilde{u}_{\text{clouds}}$ will now yield the cloud potential in real space. We are thus in a position to derive Ewald formulas for the electrostatic energy. The starting point is the usual expression for the energy in terms of charge and potential:

$$\mathcal{E} \;=\; \frac{1}{2}\sum\nolimits_{i=1}^{N} q_i \overset{i}{u}(\mathbf{r}_i) \tag{5.41}$$

where the "top-i" in $\overset{i}{u}$ indicates that the self-potential is eliminated, i.e. the potential $\overset{i}{u}(\mathbf{r}_i)$ is due to all charges but q_i.[9]

As intended, we now add and subtract the potential of all clouds:

$$\mathcal{E} \;=\; \frac{1}{2}\sum\nolimits_{i=1}^{N} q_i \left(\overset{i}{u}(\mathbf{r}_i) \;+\; \overset{i}{u}_{\text{clouds}}(\mathbf{r}_i)\right) \;-\; \frac{1}{2}\sum\nolimits_{i=1}^{N} q_i \overset{i}{u}_{\text{clouds}}(\mathbf{r}_i) \tag{5.42}$$

The first summation term in the expression above is the energy of pairwise interactions of charges with the neighboring "charge+cloud" systems; since the field of such a system is short-range, these pairwise interactions can be computed directly at the computational cost proportional to the number of charges.[10] This "direct" energy is then

$$\begin{aligned}\mathcal{E}_{\text{dir}} \;&=\; \frac{1}{2}\sum\nolimits_{i=1}^{N} q_i \left(\overset{i}{u}(\mathbf{r}_i) \;+\; \overset{i}{u}_{\text{clouds}}(\mathbf{r}_i)\right)\\ &=\; \frac{1}{4\pi\epsilon}\,\frac{1}{2}\sum_{\mathbf{n}\in\mathbb{Z}^3}^{*}\;\sum_{i,j=1}^{N} \frac{q_i q_j \operatorname{erfc}(\beta\,|\mathbf{r}_i-\mathbf{r}_j+\mathbf{n}.*\mathbf{L}|)}{|\mathbf{r}_i-\mathbf{r}_j+\mathbf{n}.*\mathbf{L}|}\end{aligned} \tag{5.43}$$

In practice, summation over all $\mathbf{n}$ is hardly ever necessary because the error function becomes negligible at separation distances much smaller than the size of the computational box.

The very fast decay of the complementary error function with distance makes the direct-sum interactions effectively short-range. In practice, a cutoff radius r_{cutoff} is chosen in such a way that $\operatorname{erfc}(\beta r_{\text{cutoff}})$ is negligible (more about that in the following section), and the respective terms in the sum are ignored:

---

[9] This rather inelegant adjustment of the potential is needed to eliminate the non-physical infinite self-energy of point charges. A rigorous definition of $\overset{i}{u}$, however, is not completely trivial. If charge q_i is excluded, the remaining system of charges is not electrically neutral, and the boundary value problem with periodic conditions is not well-posed. One can simply define $\overset{i}{u}$ as $u - u^{(i)}_{\text{self}}$, where $u^{(i)}_{\text{self}}$ is just the Coulomb potential of charge q_i in empty space. The fact that $u^{(i)}_{\text{self}}$ and $\overset{i}{u}$ do not satisfy periodic boundary conditions is unimportant because the only role of these quantities is to regularize expressions for energy by removing the singularity in an arbitrarily small neighborhood of the point charge.

[10] It is convenient to assume that the volume density of particles is fixed and the number of particles grows as the volume of the computational box grows.

$$\mathcal{E}_{\text{dir}} \approx \frac{1}{4\pi\epsilon}\frac{1}{2}\sum_{i,j=1,|\mathbf{r}_i-\mathbf{r}_j|<r_{\text{cutoff}}}^{N} \frac{q_i q_j \operatorname{erfc}(\beta\,|\mathbf{r}_i-\mathbf{r}_j|)}{|\mathbf{r}_i-\mathbf{r}_j|} \tag{5.44}$$

We now turn to the second sum in (5.42). Each term of this sum contains its own *modified* potential $\overset{i}{u}$ (with the contribution of its respective cloud eliminated); this is inconvenient, as it is much more straightforward to compute the potential of all clouds without exception. We therefore rewrite this second sum and the expression for the energy as

$$\mathcal{E} = \mathcal{E}_{\text{dir}} - \frac{1}{2}\sum\nolimits_{i=1}^{N} q_i u_{\text{clouds}}(\mathbf{r}_i) + \frac{1}{2}\sum\nolimits_{i=1}^{N} q_i u_{\text{cloud}}^{(i)}(\mathbf{r}_i) \tag{5.45}$$

where $\mathcal{E}_{\text{dir}}$ is given by (5.43). The first sum in the right hand side of this equation is easily interpreted as the energy of point charges in the field created by the clouds. It has been our intention from the beginning to compute this term in the Fourier domain by the Plancherel–Parseval theorem; this is indeed sensible, as the charge distribution of the clouds is smooth enough for the high-order Fourier harmonics to be sufficiently small.

$$\begin{aligned}\mathcal{E}_{\text{rec}} &= \frac{1}{2}\int_{\Omega} \rho_\delta\, u_{\text{clouds}}\, d\Omega = \frac{1}{2}\sum_{\mathbf{k}\in\mathcal{K}} \tilde{\rho}_\delta(\mathbf{k})\, u^*_{\text{clouds}}(\mathbf{k}) \\ &= \frac{1}{2}\sum_{\mathbf{k}\in\mathcal{K}}\sum_{i=1}^{N} q_i \exp(-\mathrm{i}\mathbf{k}\cdot\mathbf{r}_i)\, \tilde{u}^*_{\text{clouds}}(\mathbf{k})\end{aligned} \tag{5.46}$$

The potential of clouds in the Fourier domain has already been found in (5.39). The following expression for the reciprocal energy ensues:

$$\mathcal{E}_{\text{rec}} = \frac{1}{4\pi\epsilon}\frac{1}{2V}\sum_{k\neq 0} \frac{\exp(-\pi^2 k^2/\beta^2)}{k^2}\,|\tilde{\rho}_\delta(\mathbf{k})|^2 \tag{5.47}$$

Finally, the i-th term of the last summation in the energy decomposition (5.45) has an immediate interpretation as the energy of the i-th point charge in the field of its respective cloud (loosely speaking, "self-energy"). The potential at the center of the cloud is given by (5.30), and thus the self-energy term is

$$\mathcal{E}_{\text{self}} = -\frac{1}{4\pi\epsilon}\frac{\beta}{\sqrt{\pi}}\sum_{j=1}^{N} q_j^2 \tag{5.48}$$

The Ewald formulas for the electrostatic energy are summarized in the overview Section 5.5.

5.3.6 The Role of Parameters

There are two main adjustable parameters in Ewald methods: β and the cutoff radius r_{cutoff}. The latter limits the direct computation of pairwise interactions

only to charges within the cutoff distance from one another. The potential of a charge surrounded by the screening Gaussian cloud decays as $\mathrm{erfc}(\beta r)$. Since $\mathrm{erfc}(4.5) \approx 2 \times 10^{-10}$, one may want to choose, say,

$$r_{\mathrm{cutoff}} \;\geq\; 4.5/\beta \tag{5.49}$$

We shall assume that the cutoff radius is taken to be sufficiently large, so that the error due to cutoff is substantially smaller than all other numerical errors and can therefore be neglected.

Remark 13. Early on in the development of molecular dynamics, "cutoff" had a different meaning: electrostatic interactions were simply ignored beyond the cutoff. This approach results in an abrupt change of the potential and (theoretically) infinite fields and forces at the cutoff radius, violation of energy conservation, etc. More accurate and sophisticated methods for electrostatic interactions were developed to eliminate such computational artifacts. In Ewald methods, the "cutoff" is applied to the erfc terms, which is orders of magnitude more accurate than a cutoff in Coulomb terms.

In addition to the cutoff radius and β, grid-based Ewald algorithms (discussed in the subsequent sections) have other adjustable parameters, in particular, the grid size and the order of the charge interpolation scheme. Moreover, not just the parameters, but the *approaches* for grid-based computation vary.

The trade-offs for β are not difficult to see. If this parameter increases, the effective size of the cloud decreases, and the cutoff radius can be taken smaller in accordance with (5.49). This reduces the number of pairwise interactions that are computed directly in the Ewald sum. However, the charge density in the cloud decays more rapidly for higher β, and therefore more spatial harmonics have to be retained in the spatial FT to achieve the same level of accuracy.

For illustration, consider two extreme choices of β. Suppose that the volume density of charges remains constant but the number of charges and consequently the volume of the computational box grow.

Let us first keep β, and therefore the cutoff radius (5.49), constant. Then for each charge the number of its neighbors within the cutoff remains the same as well. In the direct sum (for $\mathcal{E}_{\mathrm{dir}}$) the computational cost per charge is then independent of the number of charges.

For a given level of accuracy, the infinite series for $\mathcal{E}_{\mathrm{rec}}$ can be truncated at some maximum k proportional to β and hence constant in the case under consideration. But this implies that the computational cost for the reciprocal sum grows with the number of charges N as $\mathcal{O}(N^2)$. Indeed, the number of spatial harmonics that need to be retained is proportional to the volume of the box and hence to N, because $m_x = k_x L_x/(2\pi)$, etc. The growing number of charges is accompanied by the same growth in the number of spatial harmonics, leading to the very poor $\mathcal{O}(N^2)$ scaling of the cost.

The opposite effect, but with the same unfavorable outcome, occurs if the cutoff radius is chosen to be proportional to the growing size of the box. Then

β can be reduced accordingly, making the reciprocal sum easier to compute. However, as the cutoff radius expands, a greater number of direct pairwise interactions have to be computed. The end result is the same asymptotic computational cost of $\mathcal{O}(N^2)$.

It is clear from these considerations that the β parameter controls the trade-off between the complexity of direct and reciprocal sums. One might guess that there must be the best choice of β that minimizes the overall cost. Indeed, this cost is known to be $\mathcal{O}(N^{3/2})$ [SD99, TB96], which is still suboptimal. A drastic improvement can be achieved by taking advantage of the Fast Fourier Transform (FFT) on an auxiliary grid.

5.4 Grid-based Ewald Methods with FFT

5.4.1 The Computational Work

In the Ewald expression for total energy, the cost of computing the individual terms is unequal. The number of operations required to compute the self-energy term is obviously optimal, i.e. proportional to the number of charges. For direct energy, the computational cost becomes optimal if a cutoff radius (beyond which the potential and field of the charge+cloud system becomes negligible) is introduced. In this case, each charge interacts only with its neighbors within the cutoff distance. For a fixed volume density of particles and a fixed cutoff distance the computational cost for the direct sum is again optimal.

However, reciprocal energy, if calculated in a straightforward way, becomes a bottleneck due to the computation of structure factors

$$\tilde{\rho}(\mathbf{m}) \;=\; \frac{1}{V}\sum\nolimits_{i=1}^{N} q_i \exp\left(-\mathrm{i}2\pi\left(m_x\frac{x_i}{L_x}+m_y\frac{y_i}{L_y}+m_z\frac{z_i}{L_z}\right)\right) \tag{5.50}$$

This is expression (5.21) with a slight change of notation: $\mathbf{m}$ is used instead of $\mathbf{k}$ for the reciprocal vector as a mnemonic reminder that *m*esh-based methods are under consideration. The total number of these factors is equal to the size of the reciprocal grid $M = M_x \times M_y \times M_z$, and the computation of each of these factors involves summation over all N particles; hence the total number of operations for the reciprocal sum is too high – asymptotically proportional to NM.

The structure factors are the Fourier Transform of the point charge density, and it is natural to consider *Fast* FT as a way to achieve a substantial efficiency improvement. But how exactly can this be done?

FFT operates with expressions of the form (5.50) *but* over a *discrete* set of values of the coordinates – that is, on a grid. More precisely, the 1D *discrete* FT of a sequence $\{w(n)\}_{n=1}^{N}$ is

$$\tilde{w}(m) \equiv \mathcal{F}\{w\}(m) = \sum_{n=1}^{N_x} w(n) \exp\left(-\mathrm{i}\,\frac{2\pi mn}{N_x}\right), \quad m = 1, 2, \ldots, N_x \tag{5.51}$$

where N_x is the number of grid points along the x-coordinate (not to be confused with the number of charges N). The *inverse* transform is

$$w(n) \equiv \mathcal{F}^{-1}\{\tilde{w}\}(n) = \frac{1}{N_x} \sum_{m=1}^{N} \tilde{w}(m) \exp\left(\mathrm{i}\,\frac{2\pi nm}{N_x}\right) \tag{5.52}$$

In addition to the factor $1/N_x$, the inverse transform differs from the forward one in the sign of the exponential, implying that

$$\mathcal{F}^{-1} = \frac{1}{N_x}\,\mathcal{F}^* \tag{5.53}$$

or equivalently

$$\mathcal{F}^*\mathcal{F} = \mathcal{F}\mathcal{F}^* = N_x \tag{5.54}$$

Indeed, if the FT is written in matrix-vector form, the mn-th matrix entry for the forward transform is $\exp(-\mathrm{i}2\pi mn/N_x)$, while for the inverse it is

$$\frac{1}{N_x} \exp\left(\mathrm{i}\,\frac{2\pi mn}{N_x}\right) = \frac{1}{N_x}\left[\exp\left(-\mathrm{i}\,\frac{2\pi mn}{N_x}\right)\right]^*$$

Note that if in its definition the forward FT is rescaled to include the factor of $1/\sqrt{N_x}$, then the same square-root factor will replace $1/N_x$ in the inverse transform and the FT will become unitary – i.e. its inverse will be equal to its complex conjugate. Despite some mathematical advantages of such rescaling, we shall adhere to the more common definition (5.51) with no scaling factors for the forward transform and no square roots.

An immediate consequence of the above connection between the inverse and conjugate Fourier operators is the Plancherel and Parseval relationship between the inner products and energies in the real and Fourier spaces:

$$(\tilde{w}, \tilde{v}) \equiv (\mathcal{F}w, \mathcal{F}v) = (w, \mathcal{F}^*\mathcal{F}v) = N_x\,(w, v) \tag{5.55}$$

(Plancherel) and for $w = v$

$$(\tilde{w}, \tilde{w}) = N_x\,(w, w) \tag{5.56}$$

(Parseval).

Discrete FT on *three-dimensional* grids consists in three consecutive applications of 1D transforms:

$$\tilde{w}(\mathbf{m}) = \sum_{n_x=1}^{N_x} \sum_{n_y=1}^{N_y} \sum_{n_z=1}^{N_z} w(n_x, n_y, n_z) \exp\left(-\mathrm{i}2\pi\left(\frac{m_x n_x}{N_x} + \frac{m_y n_y}{N_y} + \frac{m_z n_z}{N_z}\right)\right) \tag{5.57}$$

where w is now a function defined on a real-space grid and its transform $\tilde{w}$ is defined on a 3D reciprocal grid. The mesh in real space has $N_m = N_x \times N_y \times N_z$ nodes[11] and the reciprocal one has $M = M_x \times M_y \times M_z$ nodes.

With the basic definitions established, we can now return to the computation of the FT of the point charge density. This computation reduces to the discrete FT on the grid if, as comparison of expressions (5.50) and (5.22) shows, coordinates x_i, y_i, z_i take on a *discrete* set of values: $x_i = n_x L_x/N_x$, $y_i = n_y L_y/N_y$, $z_i = n_z L_z/N_z$ for some integer numbers n_x, n_y, n_z.

As coordinates of the particles do in fact vary continuously, in order to apply the (discrete) FFT one needs to find coefficients $w(n_x, n_y, n_z)$ that would approximáte the *continuous*-parameter exponentials as a linear combination of the *discrete*-parameter ones:

$$\exp\left(-\mathrm{i}2\pi(m_x x_i/L_x + m_y y_i/L_y + m_z z_i/L_z)\right)$$
$$\approx \sum_{n_x,n_y,n_z} b_i(n_x, n_y, n_z) \exp\left(-\mathrm{i}2\pi(m_x n_x/N_x + m_y n_y/N_y + m_z n_z/N_z)\right) \tag{5.58}$$

where summation is, in principle, over the whole grid, but in practice is over a small subset of nodes around the location (x_i, y_i, z_i) of the i-th charge; b's are coefficients to be specified.

Obviously, if the particles were located at grid nodes (n_x, n_y, n_z), the values of w would simply be equal to the values of the charges at the respective nodes of the grid (and $w = 0$ at grid nodes where no charges are present).

Remark 14. If the charges are located *between* grid nodes, the assignment of the w values can be intuitively understood as "charge allocation" to grid nodes. Despite this very common and intuitively natural interpretation, mathematically this assignment has to do with the representation of the exponential factors by linear combinations of discrete-parameter exponentials in (5.58).

In general we need a suitable mapping of the set of charge values $\{q_i\}_{i=1}^N$ to a set of grid-based coefficients $w(n_x, n_y, n_z)$. Naturally, this mapping is sought as a *linear* one and can be written in matrix form as

$$\underline{w} \;=\; \mathcal{I}_{q\to m}\, \underline{q} \tag{5.59}$$

Here $\underline{w} \in \mathbb{R}^{N_m}$ is the Euclidean vector of values of w at the mesh nodes; $\underline{q} \in \mathbb{R}^N$ is the Euclidean vector comprising the charges. $\mathcal{I}_{q\to m}$ is an $N_m \times N$ matrix that maps charges ("q") to mesh ("m") coefficients.

Fig. 5.3 gives an illustrative example of this mapping – for simplicity, in 2D. A charge q_i is shown in a grid cell with node numbers[12] 7, 8, 40, 41,

[11] Subscript "m" (for "mesh") is used instead of "g" (for grid) to avoid possible confusion between subscripts "g" and "q" (especially in handwriting) and with the usage of g and G for Green's functions.

[12] Here the nodes are referred to by their global numbers from 1 to N rather than the triple-index (n_x, n_y, n_z).

with their respective weights of 0.3, 0.4, 0.1, 0.2 (as an example). This means that e.g. $w(7) = 0.3q_i$. It is important to point out from the outset that in general the nonzero coefficients of mapping $\mathcal{I}_{q\to m}$ are not limited to the nodes adjacent to the charge. These coefficients can (and in practice do, as will be discussed later) involve several layers of nodes and, at least in principle, even the whole grid, although the latter would not be efficient computationally.

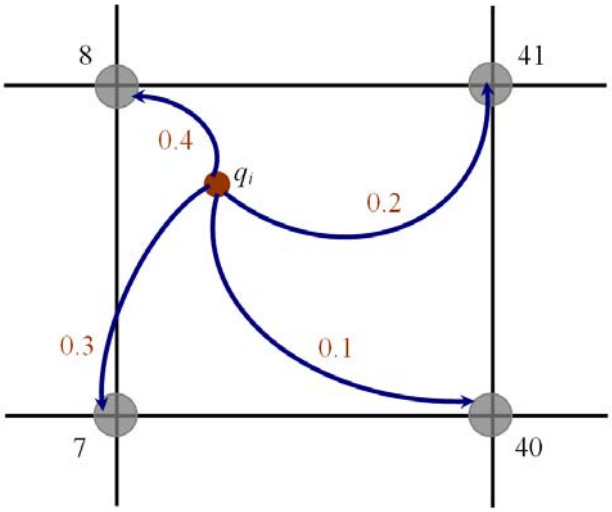


Fig. 5.3. An example of charge mapping $\mathcal{I}_{q\to m}$ onto a grid.

In this example, the i-th *column* of matrix $\mathcal{I}_{q\to m}$ contains the coefficients 0.3, 0.4, 0.1, 0.2 in their respective rows 7, 8, 40, 41; the other entries of this column are zero. The other columns of this matrix correspond to other charges and have a similar form.

Approximate values of Fourier coefficients for the point charge density (i.e. approximate structure factors) are then obtained by the discrete FT (5.57); we shall now write this transformation in matrix form as

$$\underline{\tilde{\rho}}_\delta \;=\; \frac{1}{V}\,\mathcal{F}\underline{w} \;=\; \frac{1}{V}\,\mathcal{F}\mathcal{I}_{q\to m}\,\underline{q} \tag{5.60}$$

where $\underline{\tilde{\rho}}_\delta \in \mathbb{R}^M$ is the Euclidean vector of structure factors on the reciprocal grid and $\mathcal{F}$ is the matrix of the discrete FT.

The potential in Fourier space is found, according to (5.39), by multiplying the structure factors with the Gaussian exponentials $\exp\left(-k^2\,/\,(4\beta^2)\right)$ and dividing by k^2 (i.e. solving the Poisson equation in Fourier space). Since these operations apply to each component of $\underline{\tilde{\rho}}_\delta$ separately, in matrix notation they are represented by a diagonal matrix:

$$\underline{\tilde{u}} \;=\; D\underline{\tilde{\rho}}_\delta \;=\; D\mathcal{F}\underline{\rho}_\delta \;=\; \frac{1}{V}\,D\mathcal{F}\mathcal{I}_{q\to m}\,\underline{q} \tag{5.61}$$

where the entries of the diagonal matrix D are $\exp\left(-k^2/(4\beta^2)\right)/k^2$ for $k \neq 0$.[13] For $k = 0$, the respective entry of D is irrelevant, as the mean value of the charge and potential is zero; this entry can be conveniently set to zero. Note that D is purely real:

$$D^* \;=\; D \tag{5.62}$$

By Parseval's theorem, reciprocal energy can be computed in the Fourier space:

$$\mathcal{E}_{\rm rec} \;=\; \frac{1}{2}\,(\rho_\delta, u) \;=\; \frac{1}{2V}\,(\tilde{\rho}_\delta, \tilde{u}) \;\approx\; \frac{1}{2V}\,(\underline{\tilde{\rho}}_\delta, \underline{\tilde{u}}) \;=\; \frac{1}{2V}\,(D\mathcal{F}\underline{\rho}_\delta, \mathcal{F}\underline{\rho}_\delta)\,\mathcal{I}_{q\to m}\,\underline{q} \tag{5.63}$$

To compute the field $\mathbf{E} = -\nabla u$ (and consequently the forces acting on the point charges) one needs to differentiate the electrostatic potential. This can be done either analytically in the Fourier domain or numerically, by finite differences on the grid.

We shall start with the analytical differentiation in Fourier space, which corresponds simply to multiplication with $\mathrm{i}\mathbf{k}$. Therefore components of the field $E_\alpha = -(\nabla u)_\alpha$ ($\alpha = x$, y or z) in Fourier space can be expressed in Euclidean vector form as

$$\underline{\tilde{E}}_\alpha \;=\; \frac{1}{V}\,G_\alpha\mathcal{F}\mathcal{I}_{q\to m}\,\underline{q}, \qquad \alpha = x, y, z, \qquad \underline{\tilde{E}}_\alpha \in \mathbb{R}^M \tag{5.64}$$

where each entry of the diagonal matrix G_α is obtained by multiplying the respective entry of D with $-\mathrm{i}k_\alpha$. As D is purely real, G is a purely imaginary diagonal matrix:

$$G^* \;=\; -\,G \tag{5.65}$$

The actual (real-space) field at the grid nodes is obtained by the inverse FT:

$$\underline{E}_\alpha \;=\; \frac{1}{V}\,\mathcal{F}^{-1}G_\alpha\mathcal{F}\mathcal{I}_{q\to m}\,\underline{q}, \qquad \alpha = x, y, z \tag{5.66}$$

Finally, to find the field values *at the actual locations of the particles*, one needs to *interpolate* the field from grid nodes to these locations. The interpolation procedure is defined by a suitably chosen $N \times N_m$ matrix $\mathcal{I}_{m\to q}$ that is conceptually analogous to the $N_m \times N$ matrix $\mathcal{I}_{q\to m}$ described earlier. The α-component of the field ($\alpha = x$, y, z) at the particles can then be written as a Euclidean vector $\underline{E}_{q,\alpha} \in \mathbb{R}^N$:

$$\underline{E}_{q,\alpha} \;=\; \frac{1}{V}\,\mathcal{I}_{m\to q}\mathcal{F}^{-1}G_\alpha\mathcal{F}\mathcal{I}_{q\to m}\,\underline{q}, \qquad \alpha = x, y, z \tag{5.67}$$

[13] The order in which these values appear on the diagonal of D depends on the global numbering of nodes of the reciprocal grid.

Finally, the Euclidean vector $\underline{F}_\alpha \in \mathbb{R}^N$ of the force components acting on the particles is

$$\underline{F}_\alpha \;=\; \underline{q}.*\underline{E}_{q,\alpha} \;=\; \frac{1}{V}\,\underline{q}.*\mathcal{I}_{m\to q}\mathcal{F}^{-1}G_\alpha\mathcal{F}\mathcal{I}_{q\to m}\,\underline{q} \tag{5.68}$$

where the Matlab-style notation ".*" is again used for entry-wise multiplication of vectors; i.e. $a.*b \equiv [a_1b_1, a_2b_2, \ldots, a_Nb_N]$ for two arbitrary vectors a, b in $\mathbb{R}^N$ (see footnote 5 on p. 244).

The force values computed this way are obviously only *approximations* of the true values, numerical errors coming from grid interpolation procedures and from the truncation of the Fourier transform to a finite number of terms. These approximate values may not in general obey Newton's Third Law and, consequently, the physically important conservation of momentum; however, it is prudent to require that they do and to find suitable restrictions on the interpolation procedures guaranteeing that Newton's Third Law holds.

Equivalently, one wants the following reciprocity condition to be true. Let a *unit* charge at point $\mathbf{r}_i$ create a field $\mathbf{E}_{i\to j}$ at point $\mathbf{r}_j$. Reciprocally, let a unit charge at point $\mathbf{r}_j$ create a field $\mathbf{E}_{j\to i}$ at point $\mathbf{r}_i$. Newton's Third Law condition is $\mathbf{E}_{i\to j} = -\mathbf{E}_{j\to i}$.

Let us examine what this requirement translates to in matrix form. According to the expression for the field values at the particles, any field component at location $\mathbf{r}_j$ due to the *unit* charge at $\mathbf{r}_i$ is

$$\underline{E}_{q,\alpha}(\mathbf{r}_j) \;=\; \frac{1}{V}\,\mathcal{I}_{m\to q}(\mathbf{r}_j)\,\mathcal{F}^{-1}G_\alpha\mathcal{F}\,\mathcal{I}_{q\to m}(\mathbf{r}_i) \tag{5.69}$$

It is important in this expression to show the dependence of the interpolation matrices on the location of the charge and the observation point – dependence that for the sake of brevity was not explicitly indicated previously. Despite the abundance of symbols, this expression has a clear and direct interpretation: first, assign charge to grid[14] ($\mathcal{I}_{q\to m}$), then Fourier-transform it ($\mathcal{F}$), solve the Poisson equation and analytically differentiate the potential in Fourier space (G_α), inverse-transform the result back to real space ($\mathcal{F}^{-1}$), and finally interpolate the field from mesh to the location of the charge ($\mathcal{I}_{m\to q}$).

In the reciprocal case – a unit charge at $\mathbf{r}_j$ creating a field at $\mathbf{r}_i$ – the field value is

$$E_{q,\alpha}(\mathbf{r}_i) \;=\; \frac{1}{V}\,\mathcal{I}_{m\to q}(\mathbf{r}_i)\,\mathcal{F}^{-1}G_\alpha\mathcal{F}\,\mathcal{I}_{q\to m}(\mathbf{r}_j) \tag{5.70}$$

To check the reciprocity, field $E_{q,\alpha}(\mathbf{r}_i)$ can be linked to the complex *conjugate* of $E_{q,\alpha}(\mathbf{r}_j)$:

$$E^*_{q,\alpha}(\mathbf{r}_j) \;=\; \frac{1}{V}\,\mathcal{I}^T_{q\to m}(\mathbf{r}_i)\mathcal{F}^*G^*_\alpha\mathcal{F}^{-*}\mathcal{I}^T_{m\to q}(\mathbf{r}_j)$$

Recalling that $G^*_\alpha = -G_\alpha$ (5.65) and that $\mathcal{F}^* = N_m\mathcal{F}^{-1}$ (5.53), we have

[14] See Remark 14 on p. 258.

$$E^*_{q,\alpha}(\mathbf{r}_j) \;=\; -\frac{1}{V}\mathcal{I}^T_{q\to m}(\mathbf{r}_i)\mathcal{F}^{-1}G_\alpha\mathcal{F}\mathcal{I}^T_{m\to q}(\mathbf{r}_j) \tag{5.71}$$

Since the electric field of the unit charge is real, the asterisk in the left hand side can be dropped. Then, by comparing the fields $E_{q,\alpha}(\mathbf{r}_j)$ and $E_{q,\alpha}(\mathbf{r}_i)$, we observe that the reciprocity principle (and hence Newton's Third Law and the conservation of momentum) will hold numerically, i.e.

$$E_{q,\alpha}(\mathbf{r}_j) \;=\; -\,E_{q,\alpha}(\mathbf{r}_i) \tag{5.72}$$

provided that the charge-to-grid and grid-to-charge interpolation operators are adjoint:

$$\mathcal{I}^T_{q\to m} \;=\; \mathcal{I}_{m\to q} \tag{5.73}$$

This condition was obtained by R.W. Hockney & J.W. Eastwood [HE88] in a different manner. Note that by setting $\mathbf{r}_j = -\mathbf{r}_i$ in the field reciprocity condition (5.72) one also verifies the absence of self-force.[15]

The general field approximation procedure can now be specialized – in particular, by choosing different grid interpolation operators. Two distinct possibilities are Lagrangian interpolation (Section 5.4.3) and spline interpolation (Section 5.4.4). A somewhat different approach, the "Particle–Particle Particle–Mesh Ewald" (P3M) method by Hockney & Eastwood is reviewed in Section 5.4.5.

5.4.2 On Numerical Differentiation

We previously computed fields and forces by differentiating the potential *analytically*, i.e. by multiplying it with i**k** in Fourier space. However, this procedure requires three inverse Fourier transforms (one for each component of the field/force). Another possibility is to compute the potential using one inverse transform and then differentiate the potential *numerically*.

Let Δ_α be a difference operator approximating the partial derivative in the α-direction ($\alpha = x,\, y,\, z$). This operator maps a function defined on the grid to its "discrete derivative" defined on the same grid. Well known examples of such difference operators in 1D are backward difference

$$(\Delta_{\text{b.d.}}u)_i \;\equiv\; \frac{u_i - u_{i-1}}{h_x} \tag{5.74}$$

(forward difference is completely analogous) and central difference

$$(\Delta_{\text{c.d.}}u)_i \;\equiv\; \frac{u_{i+1} - u_{i-1}}{2h_x} \tag{5.75}$$

where u is a function defined on the grid and h_x is the grid size in the x-direction. Due to periodic conditions in Ewald methods, index shifts such as

[15] There is no singularity in the self-field, as the solution has been implicitly regularized by removing the $k = 0$ term in Fourier space.

$i+1$ should be understood modulo N_x. Difference operators are discussed in more detail in Chapter 2.

The Fourier transform $\tilde{\Delta} \equiv \mathcal{F}\{\Delta\}$ of a difference operator Δ is defined to make the action of $\tilde{\Delta}$ in Fourier space correspond to the action of Δ in real space; formally,

$$\tilde{\Delta}(\mathcal{F}u) \;\equiv\; \mathcal{F}\{\Delta\}(u) \;=\; \mathcal{F}\{\Delta u\} \tag{5.76}$$

or in a more symbolic and concise way,

$$\tilde{\Delta}\mathcal{F} \;=\; \mathcal{F}\Delta \tag{5.77}$$

The Fourier transforms of backward and central difference operators can easily be found:

$$\tilde{\Delta}_{\text{b.d.}} \;\equiv\; \mathcal{F}\{(\Delta_{\text{b.d.}}u)\} \;=\; \frac{1-\exp(-\mathrm{i}k_x h_x)}{h_x} \tag{5.78}$$

$$\tilde{\Delta}_{\text{c.d.}} \;\equiv\; \mathcal{F}\{(\Delta_{\text{c.d.}}u)\} \;=\; \frac{\exp(\mathrm{i}k_x h_x)-\exp(-\mathrm{i}k_x h_x)}{2h_x} \tag{5.79}$$

In the limit $h_x \to 0$, both difference operators tend to the analytical derivative $\mathrm{i}k_x$.

With analytical differentiation of the potential, we previously had expression (5.66) (reproduced below for easy reference) for fields at the nodes:

$$\underline{E}_\alpha \;=\; \frac{1}{V}\,\mathcal{F}^{-1}G_\alpha\mathcal{F}\mathcal{I}_{q\to m}\,\underline{q}, \quad \alpha = x,y,z \tag{5.80}$$

Analytical differentiation (that is, the factor $\mathrm{i}\mathbf{k}$) was incorporated into the G_α matrix. If one uses numerical rather than analytical differentiation, the following expression ensues:

$$\underline{E}_\alpha \;=\; \frac{1}{V}\,\Delta_\alpha\mathcal{F}^{-1}D\mathcal{F}\mathcal{I}_{q\to m}\,\underline{q}, \quad \alpha = x,y,z \tag{5.81}$$

Note that matrix D, rather than $G_\alpha = \mathrm{i}k_\alpha D$, appears in this last expression. Numerical differentiation is performed in real space (Δ_α), but for theoretical analysis it is convenient to convert this operation to reciprocal space using the FT of Δ_α and the identity $\Delta_\alpha\mathcal{F}^{-1} = \mathcal{F}^{-1}\tilde{\Delta}_\alpha$ (5.77):

$$\underline{E}_\alpha \;=\; \frac{1}{V}\,\mathcal{F}^{-1}\tilde{\Delta}_\alpha(\mathbf{k})\,D(\mathbf{k})\,\mathcal{F}\mathcal{I}_{q\to m}\,\underline{q}, \quad \alpha = x,y,z \tag{5.82}$$

Thus numerical and analytical differentiation yield algebraically quite similar expressions; the only difference is that matrix G_α is replaced with $\tilde{\Delta}_\alpha D$ in the numerical formula. (Both $\tilde{\Delta}_\alpha$ and D are diagonal matrices.) In the previous section, the reciprocity principle for the field was shown to be a direct consequence of matrix G_α being skew-Hermitian. If the same condition holds for $\tilde{\Delta}_\alpha D$, reciprocity (and hence Newton's Third Law) will hold for numerical

differentiation as well. This is equivalent to $\tilde{\Delta}_\alpha$ being purely imaginary, as D is purely real.

The FT examples for backward and central difference operators can be easily generalized to show that for the transform $\tilde{\Delta}_\alpha$ to be purely imaginary, operator Δ_α must be central-difference-like (i.e. defined on a symmetric stencil, with antisymmetric coefficients). This condition was established, in a somewhat different way, by R.W. Hockney & J.W. Eastwood [HE88].

5.4.3 Particle–Mesh Ewald

As noted in Section 5.4.1, different versions of Ewald summation can be obtained by choosing different charge-to-grid interpolation operators. Lagrange interpolation leads to the so-called *Particle–Mesh Ewald* (PME) method (T. Darden *et al.* [DYP93], H.G. Petersen [Pet95]).

Let us first recall how Lagrange interpolation is defined on a given set of knots (points) $\{x_i\}_{i=1}^{N_x}$ in 1D. In general, the spacing between the neighboring knots does not have to be uniform; however, we shall deal with uniform grids only, as Fast Fourier Transforms in grid-based Ewald methods require that.

For any given knot x_α ($\alpha = 1,2, \ldots, N_x$) the Lagrange interpolation polynomial

$$\psi_\alpha(x) \;=\; \prod_{1\le j\le N_{\text{knots}};\, j\ne\alpha} \frac{x - x_j}{x_\alpha - x_j} \tag{5.83}$$

has the Kronecker-delta property of being equal to one at point x_α and zero at all other knots. Note that N_{knots} is equal to the order p_L of the Lagrange polynomial plus one.

The sum of the Lagrange polynomials corresponding to all nodes is obviously itself a polynomial of order $\le p_L$; due to the Kronecker-delta property, this sum is equal to one at all $N_{\text{knots}} = p_L + 1$ knots. Hence the sum must be equal to one for all x:

$$\sum_{\alpha=1}^{N_{\text{knots}}} \psi_\alpha(x) \;\equiv\; 1 \tag{5.84}$$

We are now in a position to define the charge-to-mesh interpolation operator $\mathcal{I}_{q\to m}$. A "portion" of charge i at a point x_i allocated to each grid node x_α is $\psi_\alpha(x_i)$; formally then,

$$\mathcal{I}_{q\to m,\alpha i} \;=\; \psi_\alpha(x_i), \quad \alpha = 1, 2, \ldots N_x, \quad , i = 1, 2, \ldots, N_{\text{knots}} \tag{5.85}$$

Let us illustrate this in the simplest possible case: interpolation by Lagrange polynomials of order $p_L = 1$, based on $N_{\text{knots}} = 2$ points $x_{1,2}$. From (5.83)

$$\psi_1(x) \;=\; \frac{x - x_2}{x_1 - x_2}; \quad \psi_2(x) \;=\; \frac{x - x_1}{x_2 - x_1} \tag{5.86}$$

Clearly, $\psi_1 + \psi_2 \equiv 1$ as it should be. For a charge located at point $x_q \in [x_1, x_2]$, its fraction $\psi_1(x_q) \;=\; (x_q - x_2)/(x_1 - x_2)$ gets assigned to node 1 and fraction $\psi_2(x_q) \;=\; (x_q - x_1)/(x_2 - x_1)$ gets assigned to node 2.

Let us now consider a numerical illustration of the accuracy of Lagrange interpolation for some realistic cases. It will be convenient to normalize the grid to unit spacing – in particular, for easy comparison with Smooth PME [EPB+95] where this normalization is also natural. We shall examine the accuracy of representing complex exponentials $\exp(\mathrm{i}2\pi m_x x_q/N_x)$ as a linear combination of grid-based exponentials $\exp(\mathrm{i}2\pi m_x x_\alpha/N_x)$ with Lagrangian weights w_α; that is,

$$\exp(\mathrm{i}2\pi m_x x_q/N_x) \approx \sum_{\alpha=1}^{N_{\mathrm{knots}}} w_\alpha \exp(\mathrm{i}2\pi m_x x_\alpha/N_x); \quad \text{with } w_\alpha = \psi_\alpha(x_q) \tag{5.87}$$

Let the order of the Lagrange interpolation p_L and consequently a number of Lagrange interpolation knots N_{knots} vary. However, the number of knots will always be assumed even, to maintain a symmetric arrangement of nodes around the charge and for consistency with Smooth PME where the same assumption is made. The leftmost and rightmost knots are then at $x_{\min} = \mathrm{floor}(x_q N_x) - N_L/2 + 1$, and $x_{\max} = \mathrm{floor}(x_q N_x) + N_L/2$, respectively, where "floor" denotes the nearest integer not greater than the given number. Multiplication of the x_q coordinate by N_x reflects the scaling to the unit spacing between the knots.

As a practical example, consider a 1D grid with $N_x = 32$ nodes along the x-axis. Suppose we wish to approximate, using Lagrange interpolation, the fourth Fourier harmonic $\exp(\mathrm{i}2\pi m_x x_q/N_x)$, $m_x = 4$ by the grid-based exponentials $\exp(\mathrm{i}2\pi m_x x_\alpha/N_x)$.

The real part of this fourth harmonic, along with its first-order Lagrange approximation, is plotted in Fig. 5.4 as a function of the (unscaled) charge location x_q, $0 \le x_q \le 1$. We can see that even first-order approximation provides a reasonable level of accuracy. For higher orders of interpolation, the approximation would be visually indistinguishable from the exact exponential. Let us then turn to *error* plots in Fig. 5.5. Three distinct error "bands" happen to correspond to three different orders of Lagrange interpolation: errors in the range of $\sim 10^{-2} - 10^{-1}$ for $p_L = 1$, in the range of $\sim 10^{-3} - 10^{-2}$ for $p_L = 3$, and in the range of $\sim 10^{-4} - 10^{-3}$ for $p_L = 5$. As we shall see in the following section, the approximation accuracy can be significantly increased by using *spline* (rather than Lagrange) interpolation.

In 3D, the interpolation (="charge assignment") operator can be defined in a natural way as a product of the respective 1D operators. That is, grid node $(x_\alpha, y_\alpha, z_\alpha)$ is assigned the fraction $\psi_\alpha(x_\alpha - x_q)\psi_\alpha(y_\alpha - y_q)\psi_\alpha(z_\alpha - z_q)$ of a charge located at (x_q, y_q, z_q).

For further details on the Particle–Mesh Ewald method that employs Lagrange interpolation see T. Darden *et al.* [DYP93], H.G. Petersen [Pet95], and M. Deserno & C. Holm [DH98a], [DH98b].

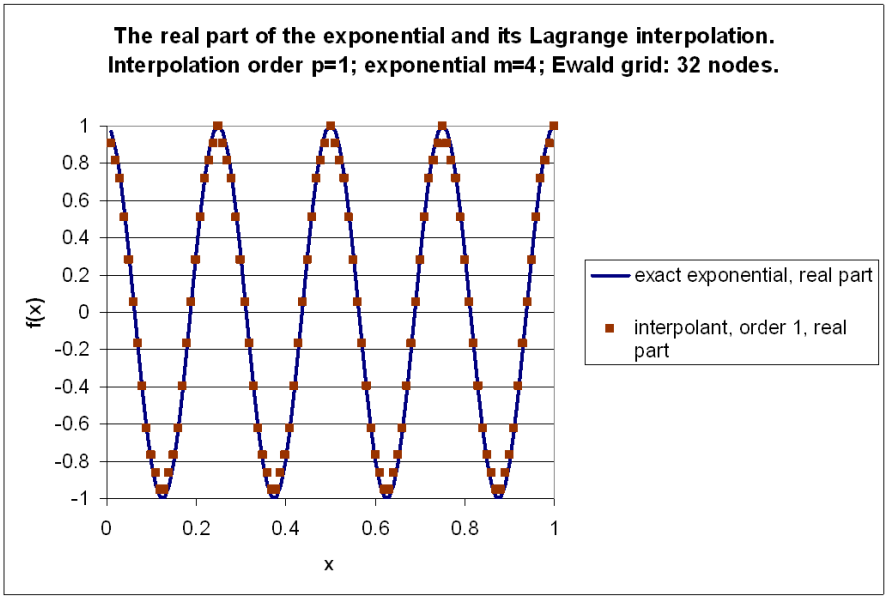


Fig. 5.4. The real part of the fourth Fourier harmonic (solid line) and its first-order Lagrange interpolation (symbols).

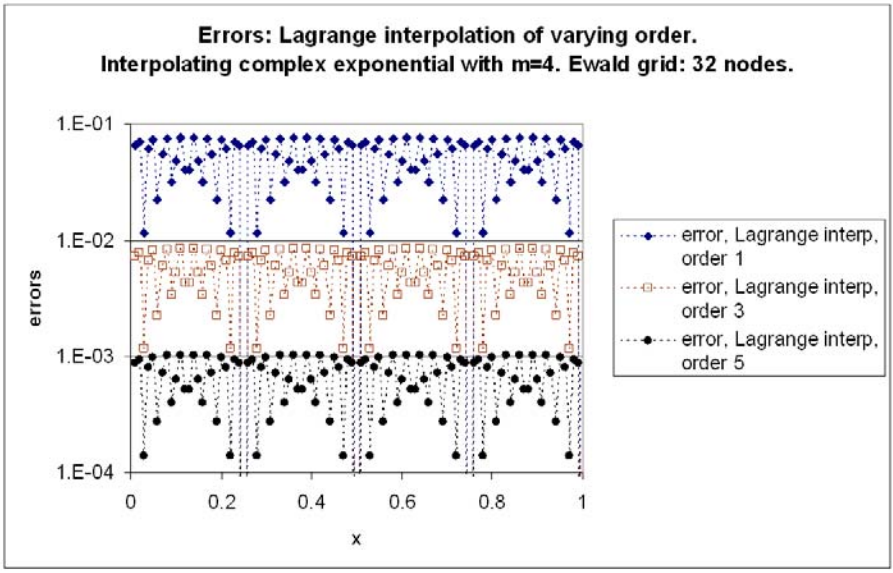


Fig. 5.5. Lagrange interpolation errors for the fourth Fourier harmonic; 32 grid nodes. Varying order of interpolation.

5.4.4 Smooth Particle–Mesh Ewald Methods

An alternative to the Lagrange approximation of exponentials is *Euler spline* interpolation employed in the "Smooth PME" method by U. Essmann *et al.* [EPB+95].

Let us first recall the basic definitions related to spline interpolation. Consider a set of nodes (*knots*) $x_0 < x_1 < \ldots < x_n$ on the x-axis, with the corresponding values y_i ($i = 1,2, \ldots, n$) of some function $y = f(x)$. A spline is a piecewise-polynomial curve that (a) passes through all given points (x_i, y_i); (b) has at least $p-1$ continuous derivatives on $[x_0, x_n]$; and (c) is a polynomial of order $\leq p$ within each subinterval $[x_i, x_{i+1}]$.

B-splines defined and analyzed in detail by C. De Boor [Boo01] and I.J. Schoenberg [Sch73] form a basis in the space of all splines of a given order over a given set of knots. For our purposes, *cardinal* B-splines – for which the knots are a set of consecutive integers – are needed.

Several different but equivalent definitions of cardinal B-splines $\hat{M}_n(x)$ are available. The hat sign is introduced here to distinguish this spline from its slightly different version used later in this section. The "hat" notation should not be confused with the Fourier Transform that I normally denote with a tilde.

Perhaps the most natural definition of these splines is via Fourier Transforms:

$$\mathcal{F}\{\hat{M}_n\}(k) \;=\; \left(\frac{\sin(k/2)}{k/2}\right)^{n+1} \tag{5.88}$$

where k is, as usual, the Fourier variable. For $n = 0$, Fourier transform (5.88) is the usual sinc function that corresponds to a rectangular pulse in real space:

$$\hat{M}_0(x) \;=\; \begin{cases} 1, & -\frac{1}{2} \leq x \leq \frac{1}{2} \\ 0, & \text{otherwise} \end{cases} \tag{5.89}$$

Since multiplication in Fourier space corresponds to convolution in real space, it follows that

$$\hat{M}_n \;=\; \hat{M}_0 * \hat{M}_0 * \ldots * \hat{M}_0 \tag{5.90}$$

where the convolution operations involve $(n+1)$ instances of $\hat{M}_0$. As a side note, in probability theory convolution of probability density functions (pdf) of *independent* random variables is the pdf of the sum of these variables; hence cardinal B-spline $\hat{M}_n$ (5.90) is the pdf of the sum of n independent random variables uniformly distributed over the interval $[-\frac{1}{2}, \frac{1}{2}]$.

This definition via convolution is not convenient or effective computationally. An alternative definition by Schoenberg [Sch73] via recursion relations in real space lends itself easily to computation. The following brief summary of Schoenberg's definition is due primarily to Essmann *et al.* [EPB+95]).

First, the backward difference of any function f is

$$\Delta f(u) \;\equiv\; f(u) \;-\; f(u-1) \tag{5.91}$$

and higher-order backward differences for $n \geq 2$

$$\Delta^n f(u) \;\equiv\; \Delta(\Delta^{n-1} f(u)) \tag{5.92}$$

It can be shown by induction that

$$\Delta^n f(u) \;=\; \sum\nolimits_{m=0}^{n} (-1)^m \frac{n!}{m!\,(n-m)!}\, f(u-m) \tag{5.93}$$

The cardinal B-spline $M_n u$ of order n is defined via the n-th backward difference of $(u_+)^{n-1}$, where $u_+ \equiv \max(u, 0)$:

$$M_n u \;\equiv\; \frac{1}{(n-1)!}\,\Delta^n\,(u_+)^{n-1} \;=\; \frac{1}{(n-1)!}\sum_{m=0}^{n} (-1)^m \frac{n!}{m!\,(n-m)!}\,(u-m)_+^{n-1} \tag{5.94}$$

Note that the difference between $\hat{M}_n$ (with the "hat") and M_n are in the index and argument shift: $M_n(x) = \hat{M}_{n-1}(x - n/2)$. Shown in Fig. 5.6 are plots of the first few splines $M_n(x)$.

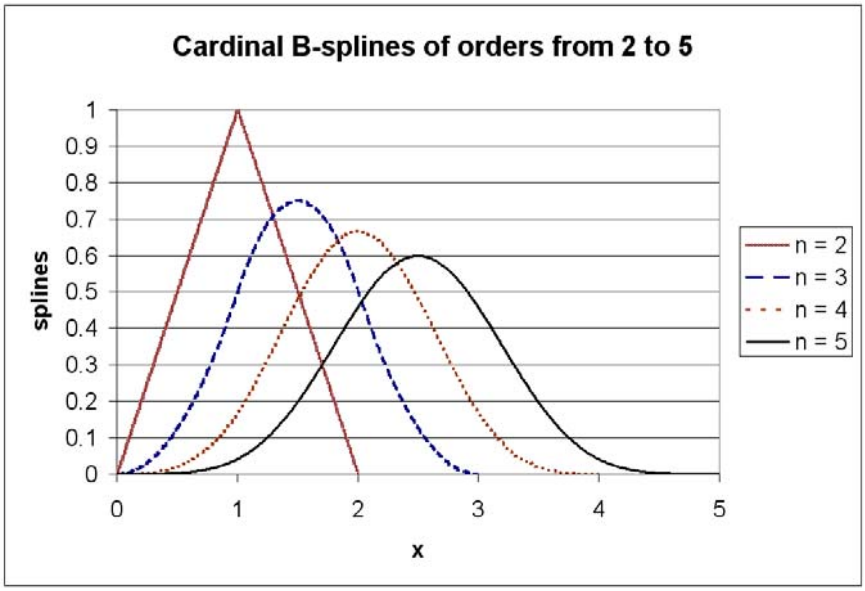


Fig. 5.6. Cardinal B-splines M_n of orders from 2 to 5.

With the B-splines now introduced, we return to the interpolation problem for complex exponentials on the grid.

Since the above definition of B-splines involves a set of integer knots $0, 1, \ldots, n$, let us rescale the coordinates to turn the grid into a lattice of

integers: $u = N_x x/L_x$. Smooth PME takes advantage of the Euler spline interpolation formula

$$\exp(2\pi \mathrm{i} m_x u_q) \;\approx\; b_x(m_x) \sum_{l=-\infty}^{\infty} M_n(u_q N_x - l) \exp\left(2\pi \mathrm{i}\, \frac{m_x}{N_x}\, l\right) \tag{5.95}$$

where the b coefficients are

$$b_x(m_x) \;=\; \exp\left(2\pi \mathrm{i}(n-1)\, \frac{m_x}{N_x}\right) \Bigg/ \sum_{l=0}^{n-2} M_n(l+1)\, \exp\left(2\pi \mathrm{i}\, \frac{m_x}{N_x}\, l\right) \tag{5.96}$$

The fact that these coefficients do not depend on x is crucial. Indeed, compare the Euler approximation with the trivial *exact* representation of the complex exponential:

$$\exp(2\pi \mathrm{i} m_x x_q) \;=\; \hat{b} \exp(2\pi \mathrm{i} m_x x_\alpha),$$

$$\text{with } \hat{b} \;\equiv\; \hat{b}(m_x, x_q) \;=\; \exp(2\pi \mathrm{i} m_x (x_q - x_\alpha))$$

where x_α is a mesh node. Despite its exactness, this representation is not practically useful, as the number of arithmetic operations needed to compute all of the $\hat{b}(m, x_q)$ factors is proportional to the number of charges *times* the number of Fourier harmonics, which is quite unattractive. In contrast, the Euler b coefficients are computed as functions of m only.

Since these coefficients depend only on the Fourier variable but not on the spatial variable, they can be incorporated into the $G(k)$ term in expressions like (5.67), which can be interpreted as a modification of Green's function on the grid. This perspective is chosen e.g. by Deserno & Holm [DH98a], although their terminology and overall approach are somewhat different from mine. Alternatively, one can continue to view the b factors as part of interpolation operators $\mathcal{I}$ rather than part of the mesh Green's function.

Fig. 5.7 may serve as a gauge of Euler spline interpolation errors. Parameters are the same as for the Lagrange interpolation in the previous section (see Fig. 5.5 on p. 266): the Ewald grid has $N_x = 32$ nodes and the fourth Fourier harmonic ($m_x = 4$) is being approximated. For a fair comparison with Lagrange interpolation, one needs to keep in mind that cardinal spline M_n is composed of polynomials of order $n-1$; for example, M_2 is piecewise-linear (see Fig. 5.6). Comparing Fig. 5.7 and Fig. 5.5, one observes that the cardinal spline algorithm provides higher accuracy of approximating the complex exponential than Lagrange interpolation. The relative advantage of splines increases with the growing order of interpolation.

5.4.5 Particle–Particle Particle–Mesh Ewald Methods

In the previous sections, we considered two most common alternatives for the charge-to-grid assignment operator (and consequently for its adjoint – grid-to-charge interpolation): namely, Lagrange and spline interpolation. The

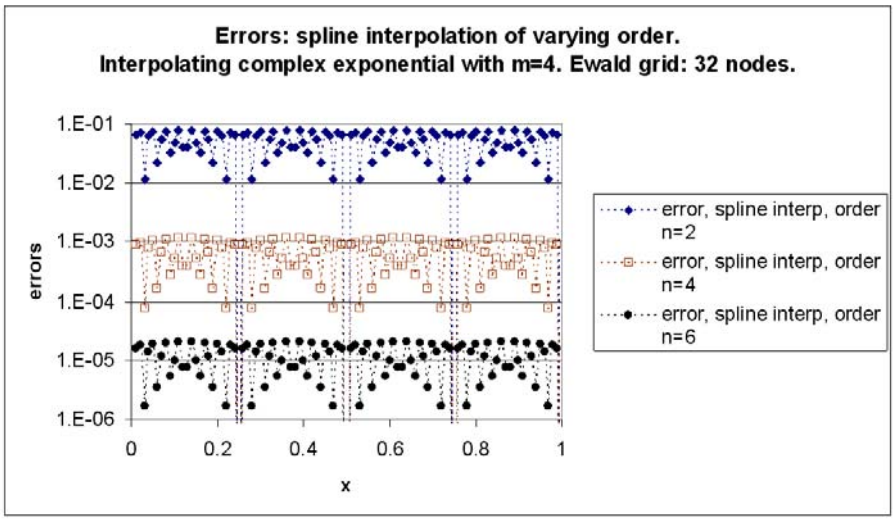


Fig. 5.7. Spline interpolation errors for the fourth Fourier harmonic; 32 grid nodes. Varying spline order.

G term in expressions for the forces (5.67) and in other similar expressions corresponded to the solution of the Poisson equation in Fourier space. (I ignore the possible adjustment of G mentioned for Smooth PME in the previous section, as it produces an algebraically equivalent result.)

There is, however, a substantially different approach. For a given interpolation procedure, one may relinquish the *direct* connection of G with the solution of the Poisson equation, allow G to float and then try to minimize the numerical error in the forces. By definition, this approach – if successful – is the most accurate one, at least with respect to the minimization criterion chosen.

R.W. Hockney & J.W. Eastwood [HE88] did in fact develop such an optimized algorithm and called it the "Particle–Particle Particle–Mesh" (or P3M) Ewald method. Although the P3M and Smooth PME interpolation procedures appear to have been developed independently, both employ B-splines and are essentially the same (apart from unimportant node index shifts). Since the interpolation (charge assignment) operators are the same but the G matrix in P3M minimizes (in a certain sense) the numerical error in force values, P3M is at least in principle more accurate than Smooth PME. A detailed theoretical and numerical investigation by M. Deserno & C. Holm [DH98a, DH98b] confirms that. However, the two algorithms are very close, and by borrowing the optimization idea from P3M, T. Darden *et al.* [DTP97] modified Smooth PME to make the accuracy of the two methods almost identical.

The technical details of P3M optimization and fine-tuning of its parameters are quite involved and will not be reported here. Interested readers are referred to the monograph and papers already cited [HE88, DH98a, DH98b].

5.4.6 The York–Yang Method

While P3M and PME (including Smooth PME) algorithms are now well established and widely used in both public domain and commercial software for molecular dynamics, other ideas have also been put forward and are worth at least a brief review.

In 1994, D.M. York & W. Yang [YY94] rewrote the Ewald sum in a form that has some advantages. In standard Ewald methods, energy is calculated as (see p. 254, (5.45))

$$\mathcal{E} \;=\; \frac{1}{2}\sum\nolimits_{i=1}^{N} q_i \left(\overset{i}{u}(\mathbf{r}_i) \;+\; \overset{i}{u}_{\text{clouds}}(\mathbf{r}_i)\right) \;-\; \frac{1}{2}\sum\nolimits_{i=1}^{N} q_i u_{\text{clouds}}(\mathbf{r}_i) \;+\; \frac{1}{2}\sum\nolimits_{i=1}^{N} q_i u^{(i)}_{\text{cloud}}(\mathbf{r}_i) \tag{5.97}$$

The immediate goal is to rewrite the reciprocal energy formula, to the extent possible, in terms of *cloud-cloud* (rather than *charge*-cloud) interactions. To that end, the cloud-cloud interaction term is added and subtracted, yielding

$$\mathcal{E} \;=\; \frac{1}{2}\sum\nolimits_{i=1}^{N} q_i \left(\overset{i}{u}(\mathbf{r}_i) \;+\; \overset{i}{u}_{\text{clouds}}(\mathbf{r}_i)\right) \;-\; \frac{1}{2}\sum\nolimits_{i=1}^{N} q_i u_{\text{clouds}}(\mathbf{r}_i) \;+\; \frac{1}{2}\sum\nolimits_{i=1}^{N} q_i u^{(i)}_{\text{cloud}}(\mathbf{r}_i) \;-\; \frac{1}{2}\int_{\Omega} \rho_{\text{clouds}} u_{\text{clouds}}\, d\Omega \;+\; \frac{1}{2}\int_{\Omega} \rho_{\text{clouds}} u_{\text{clouds}}\, d\Omega \tag{5.98}$$

Combining now terms with common factors, we get

$$\mathcal{E} \;=\; \frac{1}{2}\sum\nolimits_{i=1}^{N} q_i \left(\overset{i}{u}(\mathbf{r}_i) \;+\; u_{\text{clouds}}(\mathbf{r}_i)\right) \;-\; \frac{1}{2}\int_{\Omega} (\rho_\delta + \rho_{\text{clouds}})\, u_{\text{clouds}}\, d\Omega \;+\; \frac{1}{2}\int_{\Omega} \rho_{\text{clouds}} u_{\text{clouds}}\, d\Omega \tag{5.99}$$

The key observation now is that the first two terms (i.e. the first line in the expression above) represent short-range interactions and can therefore be computed directly. The last term (the second line) – the cloud-cloud interaction – is long-range but can be efficiently computed via FT, as in Ewald methods, or, alternatively, by numerical volume integration based on the values of charge density and potential at grid nodes. The final result is

$$4\pi\epsilon\,\mathcal{E} \;=\; \frac{1}{2}\sum_{i\neq j;\, r_{ij}<r_{\text{cutoff}}} q_i q_j\, \frac{\operatorname{erfc}(\beta r_{ij}/\sqrt{2})}{r_{ij}} \;-\; \frac{\beta}{\sqrt{2\pi}}\sum\nolimits_{i=1}^{N} q_i^2$$

$$+ \frac{1}{2} \int_{\Omega} \rho_{\text{clouds}} u_{\text{clouds}} \, d\Omega \tag{5.100}$$

See also the original paper [YY94] but note that the final expression there has a typographical error ($\sqrt{2}$ omitted in the self-energy term).

The "reciprocal" energy is now represented by the volume integral in (5.100) Since the cloud charge density is sufficiently smooth, the computation of this integral is numerically a relatively simple matter. It can be done not only in the Fourier space (as the word "reciprocal" would suggest and as done in the original paper by York & Yang) but also in real space.

Let us consider the latter alternative in some more detail.

5.4.7 Methods Without Fourier Transforms

As an alternative to reciprocal space methods, C. Sagui & T. Darden [SD01] apply finite-difference methods, with mutligrid solvers, to find the cloud potential efficiently in the context of the York–Yang algorithm. Once the potential is found, the "cloud" integral in (5.100) can be evaluated by a quadrature formula on the real-space grid.

FD schemes are discussed in detail in Chapters 2 and 4. To make this section self-consistent, I include a quick summary of the facts and features that are essential in the context of the York–Yang–Sagui–Darden method.

The Poisson equation for cloud potential is

$$\nabla^2 u_{\text{clouds}} = - \frac{\rho_{\text{clouds}}}{\epsilon} \tag{5.101}$$

subject to periodic boundary conditions. The charge density is itself spatially periodic (as it includes all clouds and their spatial images), even though for simplicity the explicit "PER" notation used previously has now been dropped.

Suitable difference schemes for this equation include classical Taylor-based methods of different order on different stencils and "Mehrstellen" schemes (see Chapters 2 and 4). The latter are advocated by C. Sagui, T. Darden and others [BSB96] due to the relatively compact stencil that reduces interprocessor communication in parallel computing. Since the charge density is smooth, the right hand side of the difference scheme is typically obtained simply by sampling the charge density at stencil nodes and taking a weighted average of these sampled values.

However, as noted in [Tsu04a] and explained in Chapters 2 and 4, the numerical accuracy of the FD solution can be substantially improved by splitting the solution up into homogeneous and inhomogeneous parts

$$u^{(i)}_{\text{clouds}} = u^{(i)}_0 + u^{(i)}_\rho, \quad \nabla^2 u^{(i)}_0 = 0; \quad \nabla^2 u^{(i)}_\rho = - \rho_{\text{clouds}} \tag{5.102}$$

Here superscript (i) emphasizes the local nature of this splitting: it is valid over a small domain containing a given grid stencil around node i (see Chapter 4 for a more complete and rigorous description of this framework). Note that no

global inhomogeneous solution u_ρ is needed to construct the difference scheme, as the scheme itself is purely local and depends only on the local properties of the potential.

Let now $L_h^{(i)}$ be any suitable difference approximation of the Laplace operator, and let $\mathcal{N}^{(i)}u$ denote the set of nodal values of potential u on grid stencil i. Since the homogeneous component $u_0^{(i)}$ of the solution by construction satisfies the Laplace equation, the difference operator can be applied to it to yield

$$L_h^{(i)}\mathcal{N}^{(i)}u_0^{(i)} \;=\; \epsilon_c \tag{5.103}$$

Since $L_h^{(i)}$ approximates the Laplace operator and since $u_0^{(i)}$ satisfies the Laplace equation, the consistency error ϵ_c can be expected, under reasonable mathematical assumptions, to tend to zero as the mesh is refined. (See Chapter 2 for a detailed discussion of this matter.) Substituting now the "difference potential" $u_0^{(i)} = u_{\text{clouds}} - u_\rho^{(i)}$, we have

$$L_h^{(i)}\mathcal{N}^{(i)}u_{\text{clouds}} \;=\; L_h^{(i)}\mathcal{N}^{(i)}u_\rho^{(i)} \;+\; \epsilon_c \tag{5.104}$$

It follows immediately that the difference scheme for the (approximate) grid-based potential u_h

$$L_h^{(i)}u_h \;=\; L_h^{(i)}\mathcal{N}^{(i)}u_\rho^{(i)} \tag{5.105}$$

has the consistency error of ϵ_c, i.e. precisely the same as for the *Laplace* equation. This implies that the solution accuracy does *not* depend on the sources of the field at all – in particular, the accuracy will not deteriorate even if the charge clouds are very sharp (large values of the Ewald β parameter). In fact, the accuracy will be the same even if scheme (5.105) is applied to point sources (= "infinitely sharp" clouds), as long as these sources do not coincide with grid points, so that the right hand side of scheme (5.105) remains mathematically valid.

The independence of consistency error from the Ewald β (however large) is a definite advantage of this approach. In contrast, the accuracy of classical schemes deteriorates for sharper clouds (i.e. larger values of β).

There is nothing paradoxical about the superior performance of the scheme with potential splitting: this scheme in essence operates on the (locally defined) *difference* potential satisfying the *Laplace* equation; the influence of the sources is confined to the inhomogeneous part $u_\rho^{(i)}$ of the total potential.

Since the cloud potential (for Gaussian charge density) is known – see (5.26) – $u_\rho^{(i)}$ can be computed analytically as a sum of contributions from clouds located in the vicinity of grid stencil i. "Vicinity" is defined by an adjustable radius r_0 (clouds centered at a distance $\leq r_0$ from stencil i contribute to $u_\rho^{(i)}$; the others do not).[16] For a fixed r_0 and a fixed volume density

[16] This setup must not be confused with the cutoff that is sometimes introduced to artificially truncate the range of particle interactions. Here, r_0 is not a "cutoff"

of particles, the operation count for the computation of $u_\rho^{(i)}$ and hence of the right hand side is optimal (proportional to the number of clouds).

The use of difference schemes with potential splitting as an alternative to Fourier-based methods is still largely unexplored. A test example with 99 charges (33 TIP3P water molecules) was considered in [Tsu04a] as a first step in this direction. The fourth order Mehrstellen scheme with the potential splitting was applied. For reference, the quasi-exact energy was computed by an "overkill" Ewald summation with terms retained up to round-off. As Table 5.1 shows, the accuracy gain in the proposed approach is appreciable, especially for finer meshes.

Table 5.1. Relative energy errors for different $n \times n \times n$ meshes. Unit cube; $\beta = 32$; $r_0 = r_{cutoff\,FFP} = 0.225$. [Tsu04a]

n	FD with potential splitting	FFP York–Yang
30	6.70×10^{-4}	1.11×10^{-3}
40	1.34×10^{-5}	1.99×10^{-5}
50	8.74×10^{-8}	5.74×10^{-7}
60	2.39×10^{-8}	4.96×10^{-7}

5.5 Summary and Further Reading

The problem of computing electrostatic energy and Coulomb forces on a periodic 3D lattice of charged particles in free space is not as simple as it might seem at first glance. Energy and forces can be formally expressed via infinite series of Coulomb terms, but these series are only *conditionally* convergent. This implies that the result depends on the order of summation and – even worse – by Riemann's series rearrangement theorem could be made to converge to *any* given value or diverge to $\pm\infty$.

P. Ewald [Ewa21] worked out alternative, unconditionally (and quickly) convergent series expressions for energy and forces of crystal lattices, and E.R. Smith [Smi81] gave a rigorous mathematical justification for Ewald summation. In Smith's approach, the shape of the crystal is fixed and expressions for energy and forces are examined as the dimensions of the body grow. In addition to the Ewald series, Smith's expressions contain a shape-dependent term that can physically be attributed to the presence of equivalent charges on the surface of the body. This term does not vanish as the size of the crystal increases and can be expected to contribute to the energy per unit cell in the crystal. It can be argued, however [DTP97], that in real crystals the

radius in this sense. The contributions of particles located beyond r_0 are not neglected; they simply contribute to the "homogeneous" part $u_0^{(i)}$ of the solution. More details and examples in Chapter 4 should clarify this point further.

actual arrangement of surface charges will tend to minimize total free energy, thereby diminishing or eliminating the additional shape-dependent term. In this chapter, this term for simplicity was not included in the expressions; it does not present any computational difficulty and can be restored at any time if necessary.

There are several ways to interpret the Ewald transformation of the original Coulomb series. Arguably, the most physically transparent interpretation is the addition and subsequent subtraction of auxiliary "clouds" of charge, usually with a Gaussian distribution of charge density. A charge with its surrounding screening cloud creates, by Gauss's Law, only a short-range field, which is easy to handle computationally. The subproblem with the clouds alone features a relatively smooth charge density distribution and its potential and field can therefore be found semi-analytically via Fourier transforms.

As a result, Ewald expressions include three main series. The first one is a "direct" term summing all pairwise interactions of particle–cloud systems that are sufficiently close to one another. The second term accounts for the interaction of point charges with clouds. The usual reference to this term as "reciprocal" does not have *physical* significance but rather reflects the most common way of *computing* this term in Fourier (i.e. "reciprocal") space. Finally, the third term is the necessary correction for the interaction energy of each charge with its own cloud and is easily computable.

Efficient Ewald methods can be obtained by applying Fast Fourier Transforms. Since these transforms are discrete, a grid has to be introduced and complex exponentials that in Ewald sums are evaluated *at particle locations* have to be approximated by similar exponentials evaluated *at grid nodes*. This procedure is commonly referred to as "charge assignment" to grid nodes, which, while not perfectly accurate mathematically, has intuitive appeal.

The general structure of grid-based Ewald methods is as follows:

1. "Charge assignment" to grid.
2. FFT of the grid-based charge density.
3. Solution of the Poisson equation in Fourier space (which amounts to simple division by $k^2\epsilon$, for $k \neq 0$).
4. Energy computation in Fourier space.
5. Inverse FFTs yielding potential and field in real space.
6. Grid-to-charge interpolation of the field yielding electrostatic forces that act on the charges.

Mathematically, this procedure can be written in the following general form:

$$\underline{E}_{q,\alpha} \;=\; \frac{1}{V}\,\mathcal{I}_{m\to q}\,\mathcal{F}^{-1}G_\alpha\mathcal{F}\,\mathcal{I}_{q\to m}\,\underline{q}, \quad \alpha = x, y, z \tag{5.106}$$

where interpolation operators $\mathcal{I}$, operator G and discrete Fourier transforms $\mathcal{F}$ were formally defined in the main text of this section; $\underline{q} \in \mathbb{R}^N$ is the Euclidean vector of charge values; V is the volume of the computational domain. It

was also shown that if "particle-to-grid" and "grid-to-particle" interpolation operators are adjoint, Newton's Third Law holds numerically.

The field, and therefore the force, can be computed either by analytical differentiation of the potential in Fourier space (i.e. by multiplication with i$\mathbf{k}$) or, alternatively, by *numerical* differentiation, as done for example in the original work by R.W. Hockney and J.W. Eastwood [HE88]. Unfortunately, differentiation does reduce the accuracy of force calculation (as a rule of thumb, by about an order of magnitude in practice), as compared to the accuracy of energy calculation.

The Ewald sums for energy can be expressed as

$$4\pi\epsilon\,\mathcal{E}_{\text{dir}} = \frac{1}{2}\sum_{\mathbf{n}\in\mathbb{Z}^3}^{*}\sum_{i,j=1}^{N}\frac{q_i q_j\,\text{erfc}(\beta\,|\mathbf{r}_i-\mathbf{r}_j+\mathbf{n}.*\mathbf{L}|)}{|\mathbf{r}_i-\mathbf{r}_j+\mathbf{n}.*\mathbf{L}|} \tag{5.107}$$

$$4\pi\epsilon\,\mathcal{E}_{\text{rec}} = \frac{1}{2V}\sum_{k\neq 0}\frac{\exp\left(-\pi^2k^2/\beta^2\right)}{k^2}\,|\tilde{\rho}(\mathbf{k})|^2 \tag{5.108}$$

$$4\pi\epsilon\,\mathcal{E}_{\text{self}} = -\frac{\beta}{\sqrt{\pi}}\sum_{j=1}^{N}q_j^2 \tag{5.109}$$

$$\mathcal{E} = \mathcal{E}_{\text{dir}} + \mathcal{E}_{\text{rec}} + \mathcal{E}_{\text{self}} \tag{5.110}$$

Here Matlab-style notation ".*" is again used for entry-wise multiplication of vectors (see footnote 5 on p. 244). $\tilde{\rho}(\mathbf{k})$ is the FT of point charges:

$$\tilde{\rho}(\mathbf{k}) = \sum\nolimits_{i=1}^{N} q_i\exp\left(-\text{i}(k_x x_i+k_y y_i+k_z z_i)\right) \tag{5.111}$$

The $\mathbf{k}$ vectors form a discrete set: $k_x = 2\pi m_x/L_x$, etc.; $m_x\in\mathbb{Z}$. L_x, L_y, L_z are the dimensions of the computational box.

The formulas for electrostatic forces are (see e.g. A.Y. Toukmaji & J. Board [TB96]):

$$4\pi\epsilon\,\mathcal{F}_{\text{dir},\alpha}(i) = q_i\sum_{j=1,j\neq i}^{N}\sum_{\mathbf{n}\in\mathbb{Z}^3}q_j\frac{r_{ij,\mathbf{n},\alpha}}{r_{ij,\mathbf{n}}^3}\left\{\text{erfc}(\beta r_{ij,\mathbf{n}})\right.$$
$$\left.+\frac{2\beta}{\sqrt{\pi}}\,r_{ij,\mathbf{n}}\exp(-(\beta r_{ij,\mathbf{n}})^2)\right\} \tag{5.112}$$

$$4\pi\epsilon\,\mathcal{F}_{\text{rec},\alpha}(i) = \frac{2q_i}{L}\sum_{j=1,j\neq i}^{N}q_j\sum_{k\neq 0}\frac{k_\alpha}{k^2}\exp\left(-\left(\frac{\pi k}{\beta L}\right)^2\right)\sin\left(\frac{2\pi}{L}\mathbf{k}\cdot\mathbf{r}_{ij}\right) \tag{5.113}$$

$$\mathcal{F}_{\text{total},\alpha}(i) = \mathcal{F}_{\text{dir},\alpha}(i) + \mathcal{F}_{\text{rec},\alpha}(i) \tag{5.114}$$

where $\alpha = x, y, z$ and $\mathbf{r}_{ij,\mathbf{n}} = \mathbf{r}-\mathbf{r}_i+\mathbf{n}.*\mathbf{L}$. (There is no self-force.)

Pressure can be computed by differentiating the energy with respect to the dimension of the computational box [TB96, DH98a, DH98b, DYP93, DTP97, SD99].

In some problems involving particle distributions near surfaces, periodic boundary conditions apply only in two directions along the surface, with particles distributed in a slab of finite thickness. The absence of periodicity in the direction perpendicular to the surface makes this problem *more* difficult computationally than its 3D-periodic counterpart considered in this chapter. A good starting point for the reader interested in this problem is A. Arnold's PhD thesis [Arn04] and a paper by M. Mazars [Maz05], with references therein.

5.6 Appendix: The Fourier Transform of "Periodized" Functions

Using the FT of a single Gaussian cloud as a starting point, we would like to find the FT (in fact, the Fourier series) of a periodic system of such Gaussians shifted by integer multiples of L_x, L_y, L_z in the three directions.

The 1D version of this task is well known in Signal Analysis and, in particular, is closely related to the Sampling Theorem. Adopting the language of Signal Analysis for convenience, let $f(t)$ be a signal with the continuous-time FT

$$\tilde{f}(\omega) \;\equiv\; \mathcal{F}\{f\} \;=\; \int_{t=-\infty}^{\infty} f(t)\,\exp(-\mathrm{i}\omega t)\,dt \tag{5.115}$$

The inverse transform is

$$f(t) \;\equiv\; \mathcal{F}^{-1}\{F\} \;=\; \frac{1}{2\pi}\int_{\omega=-\infty}^{\infty} \tilde{f}(\omega)\,\exp(\mathrm{i}\omega t)\,d\omega \tag{5.116}$$

Consider now a "periodized" version of f, i.e. a superposition of f with all its time-shifted images

$$\mathrm{PER}\{f\} \;\equiv\; \sum_{n=-\infty}^{\infty} f(t-nT) \tag{5.117}$$

where T is the basic shift.[17] It is clear that $\mathrm{PER}\{f\}$ is a periodic function with period T and can be expanded into a Fourier series:

$$\mathrm{PER}\{f\} \;=\; \sum_{n=-\infty}^{\infty} \hat{f}(n)\exp(\mathrm{i}\omega_0 nT), \quad \text{with } \omega_0 = \frac{2\pi}{T} \tag{5.118}$$

Our goal can now be stated precisely: relate the Fourier series coefficients $\hat{f}(n)$ of $\mathrm{PER}\{f\}$ to the continuous-time transform of f. To do so, we formally apply

[17] We treat T as a fixed parameter and therefore write simply $\mathrm{PER}\{f\}$ rather than $\mathrm{PER}_T\{f\}$ or $\mathrm{PER}\{f,T\}$.

continuous-time FT to PER$\{f\}$ and manipulate (at the "engineering" level of rigor) the infinite sums and Dirac delta-functions that emerge as a result:

$$\begin{aligned}\mathcal{F}\{\mathrm{PER}\{f\}\} &= \int_{t=-\infty}^{\infty} \mathrm{PER}\{f\} \exp(-\mathrm{i}\omega t) dt \\ &= \int_{t=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} f(t-nT) \exp(-\mathrm{i}\omega t) dt\end{aligned}$$

Using the fact that the FT of time-shifted signals differ only by an exponential phase factor, we obtain

$$\mathcal{F}\{\mathrm{PER}\{f\}\} = \tilde{f}(\omega) \sum_{n=-\infty}^{\infty} \exp(-\mathrm{i}\omega nT) = \tilde{f}(\omega) \sum_{n=-\infty}^{\infty} \exp\left(-\mathrm{i}2\pi n \frac{\omega}{\omega_0}\right) \tag{5.119}$$

As shown in the following Appendix, the infinite sum of exponentials above is

$$\sum_{n=-\infty}^{\infty} \exp\left(-\mathrm{i}2\pi n \frac{\omega}{\omega_0}\right) = \omega_0 \sum_{n=-\infty}^{\infty} \delta(\omega - n\omega_0)$$

Hence

$$\mathcal{F}\{\mathrm{PER}\{f\}\} = \omega_0 \sum_{n=-\infty}^{\infty} \delta(\omega - n\omega_0) \tag{5.120}$$

We can now write PER$\{f\}$ via the inverse transform

$$\mathrm{PER}\{f\} = \frac{1}{2\pi} \int_{\omega=-\infty}^{\infty} \omega_0 \sum_{n=-\infty}^{\infty} \delta(\omega - n\omega_0) \exp(\mathrm{i}\omega t) \tag{5.121}$$

Comparing this with the generic Fourier series (5.118), we obtain the desired relationship between the Fourier coefficients of the "periodized" signal and the FT of the original one:

$$\hat{f}(n) = \frac{\omega_0}{2\pi} \tilde{f}(n\omega_0) = \frac{1}{T} \tilde{f}(n\omega_0) \tag{5.122}$$

5.7 Appendix: An Infinite Sum of Complex Exponentials

The result in this Appendix is well known in Signal Analysis and is closely related to the Poisson summation formula (see e.g. S. Mallat [Mal99]). For the expressions to look more familiar, it is convenient to switch to the language of signals in the time domain. Infinite series and delta-functions are handled at the "engineering" level of rigor.

Consider a pulse train of Dirac delta functions:

$$f(t) = \sum\nolimits_{n=-\infty}^{\infty} \delta(t - t_0 - nT) \tag{5.123}$$

where t_0 is a given time shift and T is the period. Its formal expansion into the Fourier series reads

$$f(t) \;=\; \sum\nolimits_{n=-\infty}^{\infty} c_n \exp(\mathrm{i}n\omega_0 t), \qquad \omega_0 = \frac{2\pi}{T} \tag{5.124}$$

with the Fourier coefficients

$$c_n \;=\; \frac{1}{T}\int_{[t,t+T]\ni t_0} f(t)\exp(-\mathrm{i}n\omega_0 t)\,dt \tag{5.125}$$

As $f(t)$ in the case under consideration is comprised of δ-functions, the integration above is bogus and reduces to

$$c_n \;=\; \frac{1}{T}\exp(-\mathrm{i}n\omega_0 t_0) \tag{5.126}$$

Substituting coefficients c_n into the Fourier series, we obtain

$$\sum\nolimits_{n=-\infty}^{\infty} \delta(t-t_0-nT) \;=\; \frac{1}{T}\sum\nolimits_{n=-\infty}^{\infty} \exp(\mathrm{i}n\omega_0(t-t_0)) \tag{5.127}$$

Thus the infinite sum of complex exponentials can be expressed via the delta functions as

$$\sum\nolimits_{n=-\infty}^{\infty} \exp(\mathrm{i}n\omega_0(t-t_0)) \;=\; T\sum\nolimits_{n=-\infty}^{\infty} \delta(t-t_0-nT) \tag{5.128}$$

6

Long-Range Interactions in Heterogeneous Systems

6.1 Introduction

This book is motivated by problems where nanoscale phenomena and applications meet significant computational challenges and interesting numerical techniques. One case in point is long-range electro- or magnetostatic multiparticle interactions in *homogeneous* media, with applications in molecular dynamics, polymer and biomolecular simulation. Ewald summation and related algorithms (Chapter 5) are very effective for this type of problem and exemplify the blending of numerical techniques with applications.

This chapter considers a substantial generalization of this problem: long-range interactions in *inhomogeneous* media. The inhomogeneity implies spatial variation and in some cases nonlinearity of material characteristics. One class of nano- and molecular-scale problems where the inhomogeneity is crucial involves particles or macromolecules in solvents, as shown very schematically in Fig. 6.1. The precise interpretation of this figure may depend on a particular application: for example, colloidal particles with the dielectric constant ϵ_{p} in a solvent with the dielectric constant ϵ_{s}; mesoscale "beads" (connected by "springs," not shown in the picture) in a solvent for polymer models; polymer globules; macromolecules (composed of individual atoms), etc. There may of course be additional heterogeneities due to the presence of a substrate or other dielectrics.

The effective dielectric constant of protein molecules is relatively low, 2–4 (T. Simonson [Sim03]), and the same is true for colloidal particles and other bio- and macromolecules. In contrast, aqueous solutions around these molecules or particles have a much higher value of the permittivity, $\sim$80. From the physical perspective, the dielectric contrast of the media, with the commensurate changes in polarization $\mathbf{P}$, produce polarization charges equal to $-\nabla \cdot \mathbf{P}$.

Remark 15. If divergence is understood in the generalized (distributional) sense, $-\nabla \cdot \mathbf{P}$ incorporates both volume charges (for smooth variations of

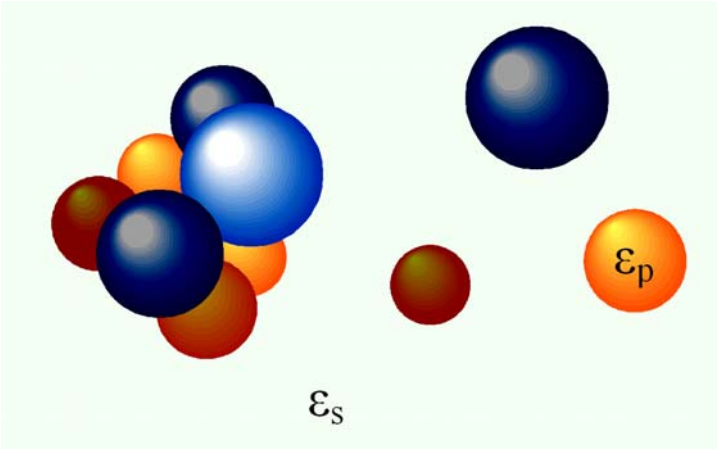


Fig. 6.1. A schematic view of heterogeneous problems involving particles (of any nature) in a solvent.

polarization) and surface charges (for abrupt changes). See Appendix 6.15 for information on distributions.

In addition, electrostatic fields in solvents are screened by electrolytes (due to the redistribution of microions, as discussed in detail later in this chapter). The polarization charge and the re-distributed microions obviously affect the electrostatic potential and field.

Ewald methods are directly applicable to *explicit* models of heterogeneous systems: the ionic and polarization effects are taken into account by explicitly including all microcharges in the model. This leads to a very large total number of charges or particles that need to be kept track of in the model. Moreover, the polarization charges are *a priori* unknown, as they depend on the field, and the computation of their values by necessity involves an iterative process wrapped around Ewald algorithms. All of that results not only in a high computational cost, but also in substantial complexity of the overall procedure.

An alternative to Ewald techniques, the Fast Multipole Method (FMM) due to L. Greengard & V. Rokhlin [GR87b, CGR99, BG] (see also Section 5.1 on p. 239), has similar limitations in heterogeneous problems. This method is designed for multiparticle interactions in free space and therefore also requires explicit treatment of all microcharges, with an iterative process for the values of these charges if they are not given. A simpler alternative is available near flat dielectric surfaces (e.g. substrates), where equivalent "image" charges representing the influence of the dielectric can be introduced. This approach is quite common and useful for theoretical analysis and intuitive insights. However, in the computational model the images further increase the number of

degrees of freedom (variables).[1] Even more importantly, the computation of image charges is quite involved even for spherical dielectric boundaries; for more complex shapes this approach becomes completely impractical. To get a flavor of this, see R. Messina's paper [Mes02].

A practical proposition is to treat field computation in heterogeneous media as a boundary value problem. While this approach is widely accepted and preferred in many areas of applied science and engineering, its use in macromolecular and nanoscale simulation so far have been limited (more about this below). Two very general techniques for boundary value problems are the Finite Element Method (FEM) and Finite Difference (FD) schemes. Another general methodology, integral equations, is well suited for linear piecewise-homogeneous media with geometrically compact boundaries;[2] it is not a good option for the multiparticle problems considered in this chapter.

FEM, described in Chapter 3, is arguably *the* most powerful simulation methodology for boundary value problems. In FEM, the computational domain is partitioned into small subdomains (elements) – in 3D, most frequently tetrahedra.[3] In many engineering applications, this partitioning is a great strength, as it results in a geometrically very accurate representation of the physical structure. However, for multiparticle simulations, with a large number of particles at arbitrary locations, mesh generation may be impractical, and the resultant system of equations may be too computationally expensive to solve. FEM can still be used effectively for a small number of particles; an example is given in Section 6.12.

FD schemes are attractive because of their relative simplicity: regular Cartesian grids can be used, and the discretization procedure is not difficult. The obvious downside, compared to FEM, is that curved or slanted boundaries cannot be rendered accurately on a regular grid. As a simple 2D illustration,

[1] In many instances, I use the term "degrees of freedom" in its physical sense of "free variables" or "free parameters" of a physical system. However, this term has a more distinctive mathematical meaning in the Finite Element context, where "degrees of freedom" are linear functionals on the finite element space (Chapter 3).

[2] For piecewise-homogeneous media, the unknown functions in the integral equation method are some equivalent sources on material interfaces. These equivalent sources create the same field in the host medium as the actual sources do. The 3D problem thus consists in finding a *2D* distribution of sources on surfaces. Although the dimensionality of the problem is reduced from 3D to 2D, the computational cost in general may well be *higher* than for FEM or FD. This is because the system matrices for integral equations are, unlike FEM or FD matrices, dense. FMM and wavelet transforms improve the efficiency of integral equation methods and make them competitive with, and in some cases preferable to, methods based on differential equations; see W.C. Chew *et al.* [CJMS01] and J.S. Zhao & W.C. Chew [ZC00].

[3] Hexahedral elements are also common, and many other types of elements are used as well, especially in commercial codes. For example, the ANSYS™ (www.ansys.com) element library contains over 150 different elements.

a circular boundary on a Cartesian grid is represented by the shaded area in Fig. 6.2. The material parameter (e.g. the dielectric constant) is usually evaluated, in classical FD schemes, at the midpoints of grid edges (marked by the asterisks in the figure).

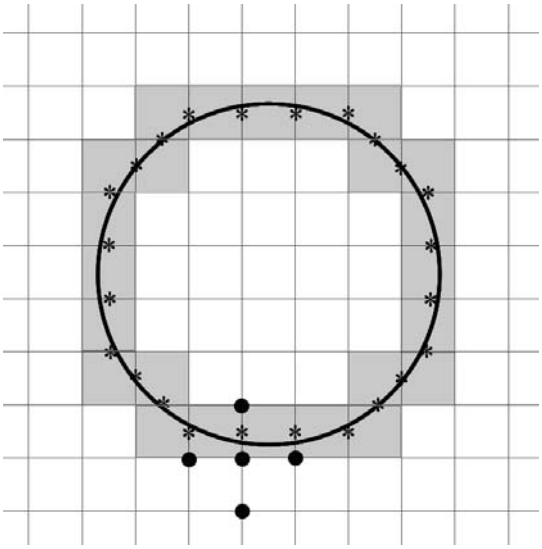


Fig. 6.2. The shaded area approximates a circular boundary on a Cartesian grid. The material parameter is typically evaluated at edge midpoints (asterisks). Small circles indicate an example of a grid stencil.

The obvious *geometric* nature of the "staircase" approximation of the boundary would be central in image processing and similar applications. For the solution of boundary value problems, this geometric effect is relevant only insofar as it affects the numerical accuracy. It is the *algebraic* approximation of the solution, rather than the geometric layout itself, that is critical. In classic FD, the algebraic approximation of the potential near material interfaces is poor, due to the discontinuity of the field, and that is a major source of numerical error. Standard FD relies on smooth Taylor polynomials that cannot represent the jump conditions on the boundary very well. The remedy is to switch from the Taylor polynomials to other approximating functions that can more closely mimic the behavior of the actual solution.

In Trefftz–FLAME schemes (Chapter 4), such approximating functions are taken as local analytical solutions of the underlying continuous problem. To give an example, a way of constructing these functions for spherical colloidal particles can be outlined as follows (see Section 6.7 and [Tsu05a, Tsu06] for details).

Inside any particle, the potential is governed by the Laplace equation and can therefore be expanded into spherical harmonics. Outside the particle, the potential satisfies, to a certain level of approximation, the Poisson–Boltzmann Equation (PBE). Once linearized, the PBE becomes the Helmholtz equation, and its solution can be expanded into harmonics involving spherical Bessel functions. Each basis function of FLAME is obtained by matching, via the boundary conditions, spherical harmonics inside and outside the particle. This produces Trefftz–FLAME basis functions satisfying the underlying equation (in this case, Laplace/linearized PBE) and the boundary conditions. This is sensible from both the mathematical and physical viewpoint.

To illustrate the usage of FLAME for particle problems, I start with two-dimensional examples of circular and elliptic particles in a dielectric host medium with no solvent (Sections 6.2 and 4.4.9) and then consider a similar problem for a spherical dielectric particle (Section 6.3). An introduction to the Poisson–Boltzmann equation (the classical Gouy–Chapman problem) is given in Section 6.4. Physical limitations of the Poisson–Boltzmann model are briefly described in Section 6.5. I explain the construction of FLAME schemes for colloidal particles in a solvent in Sections 6.6–6.7 and consider the treatment of nonlinearity of the PBE in Section 6.8. Illustrative numerical examples are presented in Section 6.12. Sections 6.9–6.11 deal with related and important topics: the DLVO theory, dispersion forces and thermodynamic potentials.

6.2 FLAME Schemes for Static Fields of Polarized Particles in 2D

A simple but compelling illustration of the efficiency of Trefftz–FLAME for particle problems was already given in Section 4.1. Fig. 4.2 (p. 191) compares two meshes that provide about the same level of accuracy for the field of a cylindrical particle in a uniform external field. The FLAME grid has more than two orders of magnitude fewer degrees of freedom (d.o.f.) than the FE mesh: 900 ($= 30 \times 30$) vs. 125,665.

In this section, FLAME schemes for electrostatic multiparticle problems are considered in more detail. The medium outside the particles is either a simple dielectric or an electrolyte. In the first case, the problem is analogous to the magnetostatic one, with magnetized particles in a medium with constant permeability.

For a simple dielectric with no electrolyte present, the electrostatic potential is governed by the Laplace equation both inside and outside the particles, with the standard conditions on the boundary of each particle

$$u_{\mathrm{in}}(r_{\mathrm{p}}) \;=\; u_{\mathrm{out}}(r_{\mathrm{p}}) \tag{6.1}$$

$$\epsilon_{\mathrm{in}}\,\frac{\partial u_{\mathrm{in}}}{\partial r} \;=\; \epsilon_{\mathrm{out}}\,\frac{\partial u_{\mathrm{out}}}{\partial r}, \qquad r = r_{\mathrm{p}} \tag{6.2}$$

where $\epsilon_{\rm p}$ and $\epsilon_{\rm out}$ are the relative permittivities of the particles and the outside medium, respectively. The assumption of equal dielectric permittivities of all particles is not essential and is taken only to avoid additional indexes in the notation.

Let us construct a Trefftz–FLAME basis in the vicinity of a cylindrical particle. In 3D, FLAME bases for spherical particles are generated in a very similar way (Sections 6.3, 6.7). Local approximating functions are chosen to satisfy the underlying differential equations and the interface boundary conditions. Since the potential is governed by the Laplace equation both inside and outside the particle, the basis functions are sought as cylindrical harmonics (Fig. 6.3)

$$\psi_{2n}(r,\phi) \;=\; \begin{cases} r^n \cos n\phi, & r \le r_{\rm p} \\ (a_n r^n \,+\, b_n r^{-n}) \cos n\phi, & r \ge r_{\rm p} \end{cases}, \qquad n = 0, 1, \ldots \tag{6.3}$$

$$\psi_{2n-1}(r,\phi) \;=\; \begin{cases} r^n \sin n\phi, & r \le r_{\rm p} \\ (a_n r^n \,+\, b_n r^{-n}) \sin n\phi, & r \ge r_{\rm p} \end{cases}, \qquad n = 1, 2, \ldots \tag{6.4}$$

In these expressions,[4] (r,ϕ) is the polar coordinate system with its origin at the center of the particle; $r_{\rm p}$ is the radius of the particle. Coefficients a_n and b_n are to be determined via the boundary conditions and are easily shown to be the same for both "sine" and "cosine" subsets of the basis; this is already reflected in the expressions.[5]

At first glance, one would expect only *one* term, with the negative power of r, to appear in the formula for ψ *outside* the particle. This would certainly be the case if the basis function were considered in the whole plane: only the negative power of r decays at infinity. However, FLAME approximations are always purely local; conceptually, the basis is introduced in a small "patch" (subdomain) containing the grid stencil (see Chapter 4). For illustration, Fig. 6.3 shows a 5-point stencil in a patch $\Omega^{(i)}$. Superscript (i) is, for simplicity of notation, dropped for the ψ functions and other variables.

The only axisymmetric basis function is $\psi_0 \equiv 1$. The Coulombic potential of a charged particle – increasing as $\log r$ outside the particle and constant inside – does not appear in the basis, as it does not satisfy the *homogeneous* conditions on the particle boundary. If the particle is charged, the corresponding inhomogeneous equation is treated in FLAME, as explained in Section 4.3.4, by local potential splitting.

To finalize the definition of the basis, one finds, in a straightforward fashion, the two unknown coefficients a_n, b_n from the two boundary conditions (6.1), (6.2). The result is

$$a_n \;=\; \frac{\epsilon_{\rm in} + \epsilon_{\rm out}}{2\epsilon_{\rm out}}, \qquad b_n = \frac{\epsilon_{\rm out} - \epsilon_{\rm in}}{2\epsilon_{\rm out}}\, r_{\rm p}^{2n} \tag{6.5}$$

[4] The "greater or equal" signs are used in *both* subcases of (6.3) and (6.4) intentionally, to emphasize the continuity of the basis functions at $r = r_{\rm p}$.

[5] It is often more convenient to use complex exponentials, rather than trigonometric functions of ϕ, but here all functions are kept real.

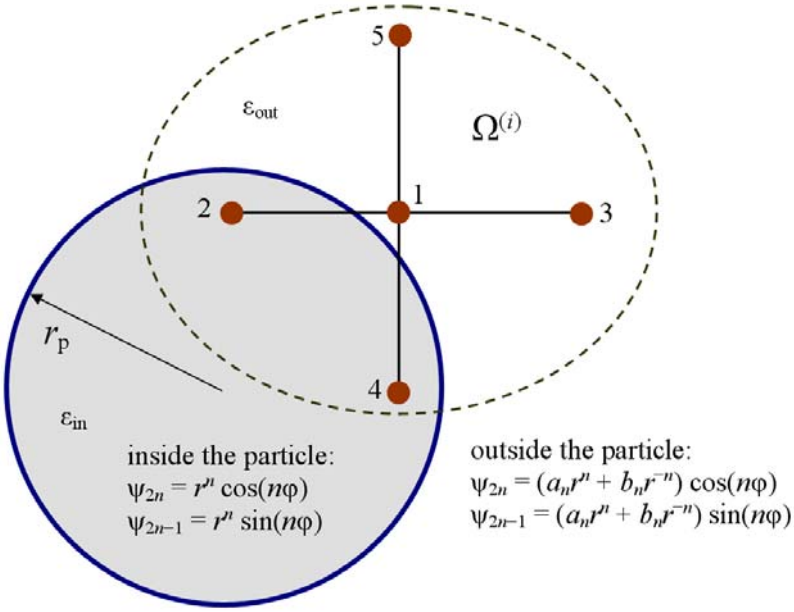


Fig. 6.3. FLAME basis functions near a cylindrical particle are defined as cylindrical harmonics inside and outside the particle, matched via the boundary conditions. Some stencil nodes in FLAME may lie inside the particle and some outside. The consistency error of the FLAME scheme is low in all cases.

These values of the coefficients complete the definition of the approximating functions in FLAME (6.3), (6.4). The number of functions to be included in the basis depends on the chosen stencil. For the 5-point stencil, four basis functions are needed. The selection of three of them is clear: $\psi_0 \equiv 1$ and $\psi_{1,2}$ (the dipole terms). The forth basis function could be taken as any linear combination of the two quadrupole ψ functions, for $n = 2$; for example, it can simply be chosen as ψ_3 of (6.4). The following numerical example clarifies this construction of the FLAME basis and the computation of the FLAME scheme.

Example 14. Suppose, in reference to Fig. 6.3, that the radius of the particle is $r_p = 1$, the particle is centered at the origin, the midpoint of the stencil is located at $x_1 = 0.9$, $y_1 = 0.8$, the dielectric constants are $\epsilon_{in} = 10$, $\epsilon_{out} = 1$, the mesh size is $h = 0.75$ in both directions, and the five stencil nodes are numbered as shown in the figure. The (transposed) nodal matrix in FLAME is

$$N^T = \begin{pmatrix} \psi_0(r_1,\phi_1) & \psi_0(r_2,\phi_2) & \dots & \psi_0(r_5,\phi_5) \\ \psi_1(r_1,\phi_1) & \psi_1(r_2,\phi_2) & \dots & \psi_1(r_5,\phi_5) \\ \psi_2(r_1,\phi_1) & \psi_2(r_2,\phi_2) & \dots & \psi_2(r_5,\phi_5) \\ \psi_3(r_1,\phi_1) & \psi_3(r_2,\phi_2) & \dots & \psi_3(r_5,\phi_5) \end{pmatrix} \tag{6.6}$$

Since $\psi_0 \equiv 1$, all entries in the first row of the matrix are simply equal to one. The remaining entries depend on the cylindrical coordinates of all stencil nodes:

Stencil node	x	y	r	ϕ
1	0.9	0.8	1.20416	0.72664
2	0.15	0.8	0.81394	1.38545
3	1.65	0.8	1.83371	0.45145
4	0.9	0.05	0.90139	0.055498
5	0.9	1.55	1.79234	1.04473

Let us compute, say, the third row of the (transposed) nodal matrix. As (6.6) shows, this row contains the values of the basis function ψ_2 at the five stencil nodes. Expression (6.3) is, in the case of ψ_2 and with coefficients a_n, b_n shown explicitly,

$$\psi_2(r,\phi) = \begin{cases} r\cos\phi = x, & r \le r_\mathrm{p} \\ (2\epsilon_\mathrm{out})^{-1}\left((\epsilon_\mathrm{in}+\epsilon_\mathrm{out})r + (\epsilon_\mathrm{out}-\epsilon_\mathrm{in})\, r_\mathrm{p}^2 r^{-1}\right)\cos\phi, & r \ge r_\mathrm{p} \end{cases} \tag{6.7}$$

Substituting the coordinates of all nodes, we find that the third row of the matrix is approximately (2.15689, 0.15, 6.86682, 0.9, 3.6893). Repeating such a straightforward calculation for the remaining rows, we obtain the complete nodal matrix:

$$N^T_\mathrm{example} \approx \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1.91724 & 0.8 & 3.32937 & 0.05 & 6.35379 \\ 2.15689 & 0.15 & 6.86682 & 0.9 & 3.6893 \\ 4.83795 & 0.24 & 13.4693 & 0.09 & 14.1284 \end{pmatrix} \tag{6.8}$$

The null space of this matrix is one-dimensional, and the FLAME difference scheme is

$$\underline{s} \in \mathrm{Null}\, N^T_\mathrm{example} \approx (1,\, 0.15706,\, -0.05774,\, -0.81446,\, -0.28485)^T \tag{6.9}$$

(up to an arbitrary factor). The first coefficient, corresponding to the central node of the stencil, has been normalized to one for convenience.

Example 15. As an extension of the previous example, let us now assume that the cylindrical surface of the particle is uniformly charged, with surface charge density ρ_S or, equivalently, with charge density per unit axial length $\rho_l = 2\pi r_\mathrm{p}\rho_S$. How does this affect the FLAME scheme?

As explained in Section 4.3.4 on p. 203, the FLAME matrix remains the same as for the homogeneous equation – i.e. is specified by (6.8) in this example. The right hand side of the system has a nonzero entry defined in FLAME

via potential splitting. The general procedure is described in Section 4.3.4; in the example under consideration, the splitting is

$$u \;=\; u_0 + u_f, \qquad u_f \;=\; \begin{cases} 0, & r \le r_{\rm p} \\ -\rho_S r_{\rm p} \epsilon_{\rm out}^{-1} \log \frac{r}{r_{\rm p}} & r \ge r_{\rm p} \end{cases} \tag{6.10}$$

Indeed, it is straightforward to verify that u_f satisfies the Laplace equation both inside and outside the particle, as well as the inhomogeneous boundary condition – the jump of the radial component of the $\mathbf{D}$ vector does correspond to the surface charge.

Once the coordinates of each stencil node are substituted into this expression for u_f, the vector of nodal values of u_f is found to be

$$\mathcal{N}^{(i)} u_f \;=\; (-0.18578, 0, -0.60634, 0, -0.58352)^T \tag{6.11}$$

where operator $\mathcal{N}^{(i)}$ indicates the nodal values on a given stencil i; see Section 4.3.4. The entry corresponding to this stencil in the right hand side of the FLAME system is, from (6.9) and (6.11),

$$\underline{s}^T \mathcal{N}^{(i)} u_f \;\approx\; 0.01545$$

The FLAME scheme on this stencil can now be explicitly written as (with about five digits of accuracy)

$$u_1 \,+\, 0.15706 u_2 \,-\, 0.05774 u_3 \,-\, 0.81446 u_4 \,-\, 0.28485 u_5 \;=\; 0.01545$$

This completes the numerical example.

6.2.1 Computation of Fields and Forces for Cylindrical Particles

Solution of the FLAME system yields potential values at the grid nodes. A typical goal of the simulation, however, is to compute forces. The electrostatic force[6] acting on a given particle can be found, as known from electromagnetic theory, by integrating the Maxwell Stress Tensor (MST) over a closed surface containing this particle and no other particles.[7]

The electrostatic part $\overleftrightarrow{T}^{\rm el}$ of the MST is defined as (see e.g. J.D. Jackson [Jac99], J.A. Stratton [Str41] or W.K.H. Panofsky & M. Phillips [PP62]):

$$\overleftrightarrow{T}^{\rm el} \;=\; \epsilon \begin{pmatrix} E_x^2 - \frac{1}{2}E^2 & E_x E_y & E_x E_z \\ E_y E_x & E_y^2 - \frac{1}{2}E^2 & E_y E_z \\ E_z E_x & E_z E_y & E_z^2 - \frac{1}{2}E^2 \end{pmatrix} \tag{6.12}$$

[6] Similar considerations apply to magnetostatic forces in magnetic fields.

[7] For particles in electrolytes, there is also an osmotic pressure force due to uneven concentration of microions around the particle. This type of force will be considered in Section 6.11.

where ϵ is the dielectric constant of the medium in which the particles are immersed, E is the amplitude of the electric field and $E_{x,y,z}$ are its Cartesian components.

The electrostatic force is

$$\mathbf{F} \;=\; \int_S \overleftrightarrow{T}^{\mathrm{el}} \cdot \mathbf{dS} \;=\; \epsilon \int_S \left[(\mathbf{E} \cdot \hat{n})\mathbf{E} \;-\; \frac{1}{2} E^2 \hat{n} \right] dS \tag{6.13}$$

Here S is an arbitrary closed surface containing one, and only one, particle; $\hat{n}$ is the exterior unit normal vector to S. For cylindrical particles, forces are computed per unit axial length and the surface integral reduces to a line integral; however, with the 3D case in mind, I shall still call it "surface" integration.

For numerical integration, the field $\mathbf{E} = -\nabla u$ needs to be computed at arbitrary points on surface S, and hence an interpolation procedure is called for. Although various forms of interpolation could be considered, the most natural one employs the local approximating functions used to construct the FLAME scheme. The local FLAME approximation over stencil number i is

$$u_h^{(i)} \;=\; \sum\nolimits_\alpha c_\alpha^{(i)} \psi_\alpha^{(i)} \;+\; u_f^{(i)} \tag{6.14}$$

The expansion coefficients c_α and the values of the numerical potential u_h at the stencil nodes are linearly related:

$$\underline{u}^{(i)} \;=\; N^{(i)} \underline{c}^{(i)} \;+\; \mathcal{N}^{(i)} u_f^{(i)} \tag{6.15}$$

where $N^{(i)}$ is the matrix of nodal values of basis functions $\psi_\alpha^{(i)}$ on the stencil.

Once the FLAME system of equations has been solved and the numerical solution – the nodal values on the grid – has been found, one may view (6.15) as a system of equations with respect to the expansion coefficients $\underline{c}^{(i)}$:

$$N^{(i)} \underline{c}^{(i)} \;=\; \underline{u}^{(i)} \;-\; \mathcal{N}^{(i)} u_f^{(i)} \tag{6.16}$$

This system is typically overdetermined: the number of rows in $N^{(i)}$ (equal to the number of stencil nodes – e.g. five) is usually greater than the number of columns (= the number of FLAME basis functions – e.g. four). However, if the null space of $N^{(i)T}$ is one-dimensional,[8] the system is consistent. That is, the right hand side of the system belongs to the image of $N^{(i)}$ – or, equivalently, is orthogonal to the null space of $N^{(i)T}$. This follows from the very definition of the FLAME scheme for inhomogeneous equations:

$$\underline{s}^{(i)T} \underline{u}^{(i)} \;=\; \underline{s}^{(i)T} \mathcal{N}^{(i)} u_f^{(i)} \tag{6.17}$$

Since the coefficient vector $\underline{s}^{(i)}$, according to the FLAME procedure, is in the null space of $N^{(i)T}$, and since this null space is by assumption one-dimensional,

[8] Recall that this null space defines the FLAME difference scheme.

equation (6.17) states that the right hand side of (6.16) is indeed orthogonal to the null space of $N^{(i)T}$.

Hence the vector of expansion coefficients for the FLAME solution $u_h^{(i)}$ over stencil i can be found from the consistent system[9] (6.16). Coefficients $\underline{c}^{(i)}$ then define, via (6.14), the FLAME interpolation $u_h^{(i)}$ in the vicinity of stencil i (technically, in the "patch" $\Omega^{(i)}$ containing the stencil). The electric field is

$$\mathbf{E}_h^{(i)} \;=\; -\sum\nolimits_\alpha c_\alpha^{(i)} \nabla \psi_\alpha^{(i)} \;-\; \nabla u_f^{(i)} \tag{6.18}$$

The electrostatic force is found by numerical integration of the MST (6.13).

Remark 16. Theoretically, the value of the force does not depend on the choice of the integration surface, but numerically it does. Numerical results for rectangular and circular integration paths are compared in the example below.

Remark 17. As argued in Chapter 4, in FD methods (FLAME included) approximation between the nodes is inherently multivalued. The solution is defined locally, over subdomains ("patches") $\Omega^{(i)}$. At any point in space, two or more of these patches can overlap, and two or more respective values of the field $\mathbf{E}_h^{(i)}$ can coexist. The field value in the MST integration (6.13) can be defined as some weighted average of the values from the nearby "patches". The simplest choice is just the field value corresponding to the nearest patch – i.e. to the nearest (in some sense) stencil. As the grid is refined, multiple values from different patches are expected to converge, as the numerical experiments in the following section illustrate. Moreover, the discrepancy between these values may serve as an error indicator for adaptive procedures; an example is given in Section 6.2.3.

6.2.2 A Numerical Example: Well-Separated Particles

Numerical experiments in this subsection were performed by Jianhua Dai.

A test problem with ten cylindrical particles is considered in this section as an example of FLAME. Locations of the particles in the rectangular computational domain $[-8,8]\times[-8,8]$ are shown in Fig. 6.4, where the equipotential lines are also displayed to visualize the field distribution.

All particles are taken to be identical, with the radius $r_\mathrm{p} = 1$ and the relative dielectric permittivity $\epsilon_\mathrm{p} = 10$; the dielectric constant of the surrounding medium is one. The particles have zero net charge but are polarized by an external electric field applied in the negative x-direction; the magnitude of this field is normalized to unity, and its potential far away from the particles is simply $u_\mathrm{ext} = x$.

[9] In practice, due to numerical errors inherent in the computation of $\underline{u}$ by linear system solvers, especially iterative ones, the system is "almost," but not exactly, consistent. This does not normally cause any problems.

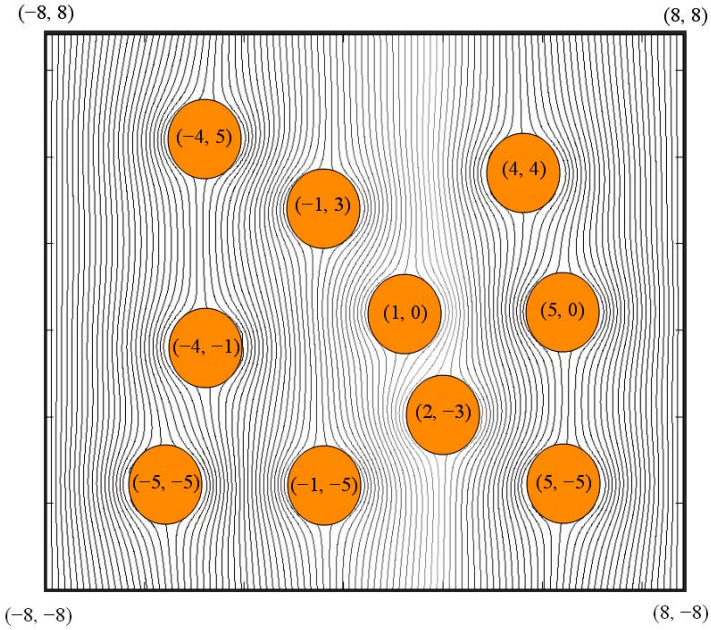


Fig. 6.4. Credit: Jianhua Dai. A test problem with ten particles, with the coordinates of particle centers indicated. The field distribution is characterized by the equipotential lines.

The problem has an analytical solution via the multipole-multicenter expansion. With 20 cylindrical harmonics per particle retained in the expansion, the error turns out to be on the order of 10^{-10}, and for practical purposes this solution is treated as "exact". To eliminate the effects of domain truncation in the testing and verification of FLAME, this "exact" multipole-multicenter solution is imposed as the Dirichlet condition on the exterior boundary of the computational domain.

In this example, the particles are well separated in the sense that no "patch" $\Omega^{(i)}$ (containing grid stencil i) intersects with two or more particles. Consequently, the FLAME basis in each patch can be supplied by the closest particle. The more complicated case of a grid stencil with nodes in two nearby particles is considered in Section 6.2.3.

We first examine convergence of the nodal potential as a function of grid size for the 5-point and 9-point FLAME schemes. The relative rms error is defined as

$$e_u = \frac{\|\underline{u}_h - \mathcal{N} u_{\text{exact}}\|_2}{\|\mathcal{N} u_{\text{exact}}\|_2} \tag{6.19}$$

where u_{exact} is the "exact" (multipole-multicenter) potential as explained above. Fig. 6.5 shows the relative error in the potential vs. mesh size h on

a log-log scale. The error decays approximately as $\mathcal{O}(h^{1.3})$ for the 5-point scheme (see dashed line as a visual aid) and approximately as $\mathcal{O}(h^{3.5})$ for the 9-point scheme.[10]

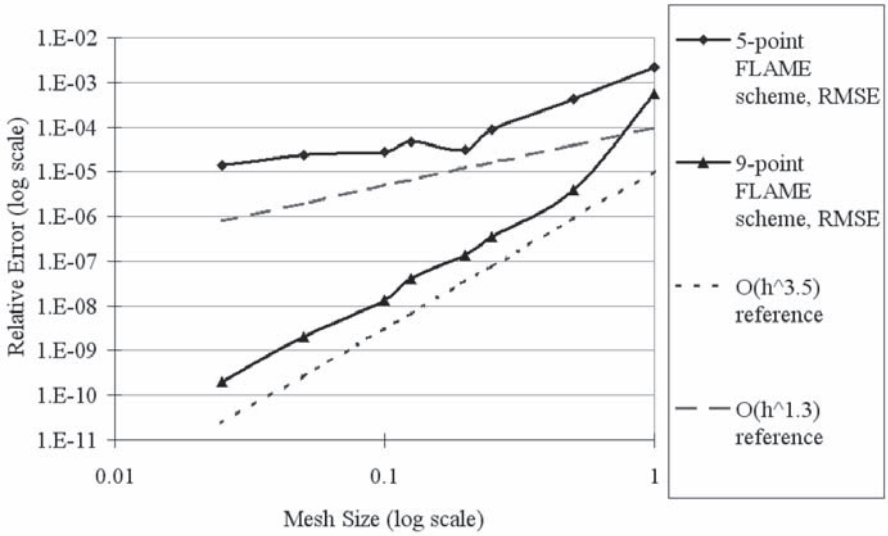


Fig. 6.5. Relative RMS error in the potential for 5-point and 9-point FLAME schemes. (Simulation by Jianhua Dai.)

A similar definition of the relative rms error is used to evaluate the accuracy of the electric field at 100 points chosen randomly in the computational domain. The numerical result is again compared with the multipole-multicenter expansion. Surprisingly, the rate of convergence for the field is not much worse than for the potential: the field error decays approximately as $\mathcal{O}(h^{1.1})$ for the 5-point scheme and as $\mathcal{O}(h^{3.5})$ for the 9-point scheme (Fig. 6.6). In general, differentiation of the potential (to compute the field) almost unavoidably degrades the numerical accuracy. This degradation in the example under consideration turns out to be very moderate.

Finally, force values are computed by numerical integration of the MST over rectangular or circular paths. The edge length of the rectangular path is 10% greater than the diameter of the particle, and the number of integration knots is 100. For verification purposes, the quasi-exact force is calculated using a 40,000-knot numerical quadrature of the "exact" field computed with 40 multipole-multicenter harmonics. The trapezoidal rule is used for numerical integration.

[10] For the 9-point scheme, the slope of the line corresponds to $\sim \mathcal{O}(h^{3.8})$ if all data points are taken into account and to $\sim \mathcal{O}(h^{3.27})$ if the initial sharp decay between the first and second data point is excluded.

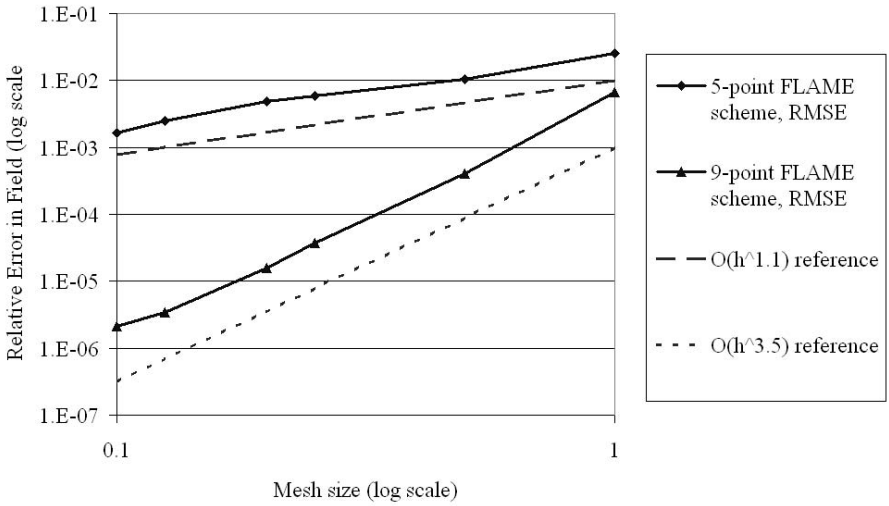


Fig. 6.6. Relative rms error in the electric field for 5-point and 9-point FLAME schemes vs. grid size. (Simulation by Jianhua Dai.)

The radius of the circular integration path is also chosen to be 10% greater than the particle radius. The numerical quadrature for the FLAME force and the "overkill" integration are implemented in the same way as for the rectangular path. The asymptotic behavior of errors in the force is $\sim \mathcal{O}(h^{1.45})$ for the 5-point scheme and $\sim \mathcal{O}(h^{3.46})$ for the 9-point scheme (Fig. 6.7).

6.2.3 A Numerical Example: Small Separations

Numerical experiments in this subsection were performed by Jianhua Dai.

Ideally, Trefftz–FLAME incorporates local *analytical* solution of the governing equation into the difference scheme. However, when analytical approximations are too complicated or unavailable, numerical ones can be used instead. In multiparticle problems, this is the case when several particles are in close proximity to one another or when particles have complex shapes.

J. Dai [DT06] uses local *numerical* and semi-analytical solutions as FLAME basis functions in multiparticle simulations. More specifically, the FLAME basis is constructed either by solving small local finite element problems or, alternatively, by a local multipole-multicenter expansion.

The formulation of the problem is the same as before: the Laplace equation both inside and outside the particles, with standard boundary conditions (6.1) and (6.2). If an external field with potential $u_0(r)$ is applied, the Dirichlet boundary condition at infinity is

$$u(r) \;\to\; u_0(r) \quad \text{as } r \to \infty \tag{6.20}$$

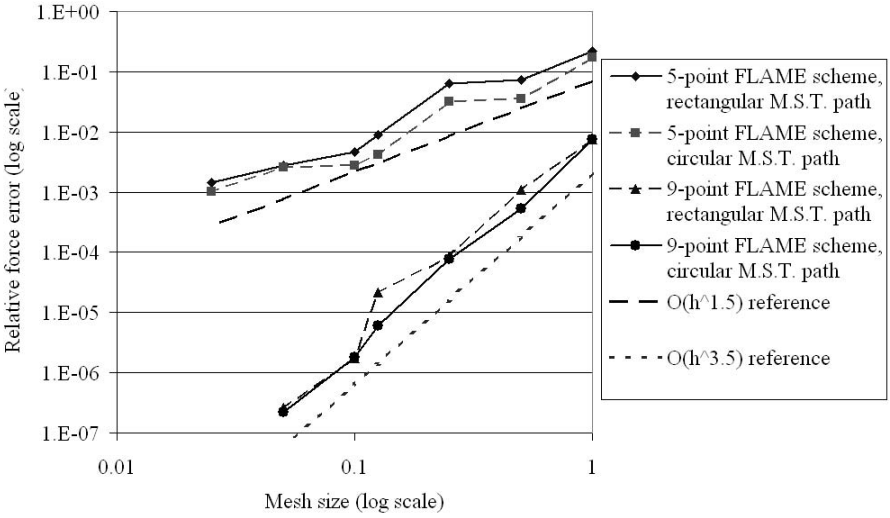


Fig. 6.7. Relative rms error in the electrostatic forces acting on the particles for two FLAME schemes and two MST integration paths vs. grid size. (Simulation by Jianhua Dai.)

In Section 6.2, FLAME bases in the vicinity of a given particle were obtained analytically, by matching harmonic expansions inside and outside the particle. The area of applicability of this approach has limitations, however. If the shape of particles (or other dielectric objects) is not cylindrical or spherical, it is substantially more difficult to construct local analytical approximations of the potential. Furthermore, if two or more particles are separated by distances comparable or smaller than the grid size, the nodes of a stencil may "belong" to different particles (Fig. 6.8), and efficient ways of constructing a Trefftz–FLAME basis in such cases need to be found.

Let us consider a pair of spherical particles of the same radius and examine the dependence of numerical errors on the separation distance between the particles. A uniform external field along the x-coordinate is applied. The relative rms error (rRMSE) in the potential is measured by its average value over more than 1000 random sampling points.

Suppose that the FLAME bases are computed analytically (Section 6.2) by taking into account only one particle closest to the midpoint of the grid stencil. This works well as long as the particles are well separated, i.e. the gap between them is substantially greater than the mesh size. For example, with the gap between a pair of particles equal to $3r_{\mathrm{p}}$ (where r_{p} is the radius of each particle), and with the mesh size equal to one-quarter of the particle radius in each of the three directions, the relative rms error for the potential over the sampling points is about 0.6%. However, when the gap diminishes to

$0.1r_{\mathrm{p}}$ (with the same mesh size), the error increases by more than an order of magnitude, to about 6.7%. In this latter situation, the particles are too close to one another for the solution based on just one of them to be physically meaningful.

To rectify the situation, two approaches for generating FLAME bases are explored. The first one – *local* multipole-multicenter expansions – is applicable to cylindrical or spherical particles and yields an analytical solution even if the particles are in close proximity to one another. Since the relevant techniques and mathematical formulas are very well known, especially in the context of Fast Multipole Methods (see e.g. H. Cheng *et al.* [CGR99]), they are not described here but are used in numerical experiments. Note that only local expansions, involving a small number of nearby particles, are needed to generate the FLAME basis.

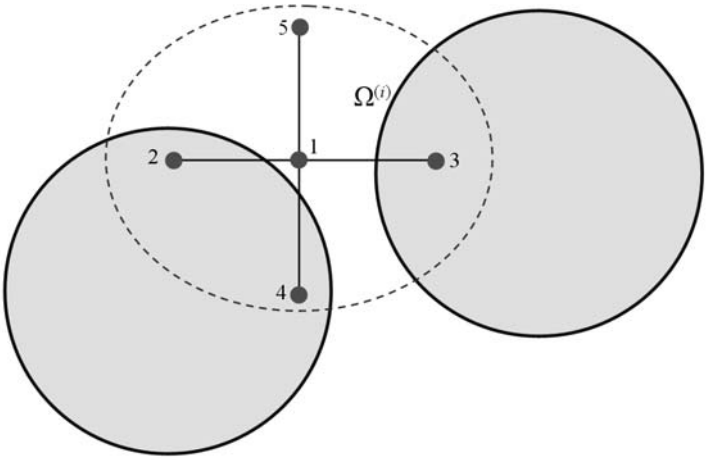


Fig. 6.8. Patch $\Omega^{(i)}$ (dashed line) intersects two nearby particles, which complicates the analytical approximation within this patch.

Another way of constructing the bases is quite useful in cases where local analytical approximations are unavailable. This approach relies on an accurate *numerical*, rather than analytical, solution of a local field problem in patches $\Omega^{(i)}$. While any numerical technique can in principle be applied for this purpose, the Finite Element Method (FEM) is the most general and powerful tool. Note that the local problem does not require construction of globally conforming FE meshes and is in all respects *much* simpler than the global problem would be.

This is further illustrated by the following test examples with four dielectric particles in free space. The particles in air have the relative dielectric constant of $\epsilon_{\mathrm{p}} = 10$. A uniform external field is applied. As before, an analytical solution is available via the (global) multipole-multicenter expansion

– in practice, truncated at the terms with the magnitude below 10^{-11}. As in previous tests, this quasi-exact solution is applied on the domain boundary as a Dirichlet condition, to eliminate the numerical error associated with the approximation of boundary conditions. The layout is shown in Fig. 6.9. The

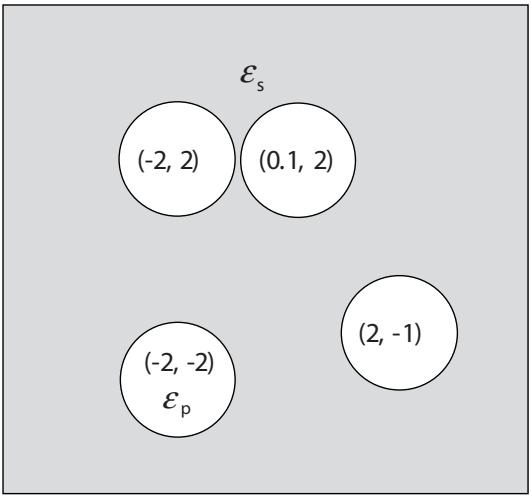


Fig. 6.9. A 2D model problem with four particles. (Credit: J. Dai.)

radii of all particles are $r_p = 1$, and there is a pair of particles with a gap of only 0.1 between them. Two kinds of FLAME bases are used: one from the local multipole-multicenter expansion and the other one, purely numerical, from FE analysis.

The overall accuracy of FLAME with numerical (finite-element) basis functions depends on two main factors. One source of error is the finite-difference discretization by FLAME itself; this error depends primarily on the mesh size of the global Cartesian grid in FLAME. The other source of error is the accuracy of the numerical bases – this error is governed by the usual FE parameters such as the FE mesh size, the order of finite elements, and the geometric shape of the elements.

Fig. 6.10 shows the FLAME simulation results for bases constructed by local multipole-multicenter expansions. The accuracy of FLAME is easily seen to be much higher than that of the standard FD (sFD)–flux balance schemes. When the mesh size is greater than the smallest gap between the particles, sFD provides a very crude approximation at best. Only after the grid size falls below the smallest gap does the accuracy of sFD begin to improve.

For the 5-point FLAME scheme with multipole-multicenter bases, the accuracy improves as the mesh is refined, provided that sufficiently many (in this example, 40) harmonics are used to generate the FLAME basis. For a smaller

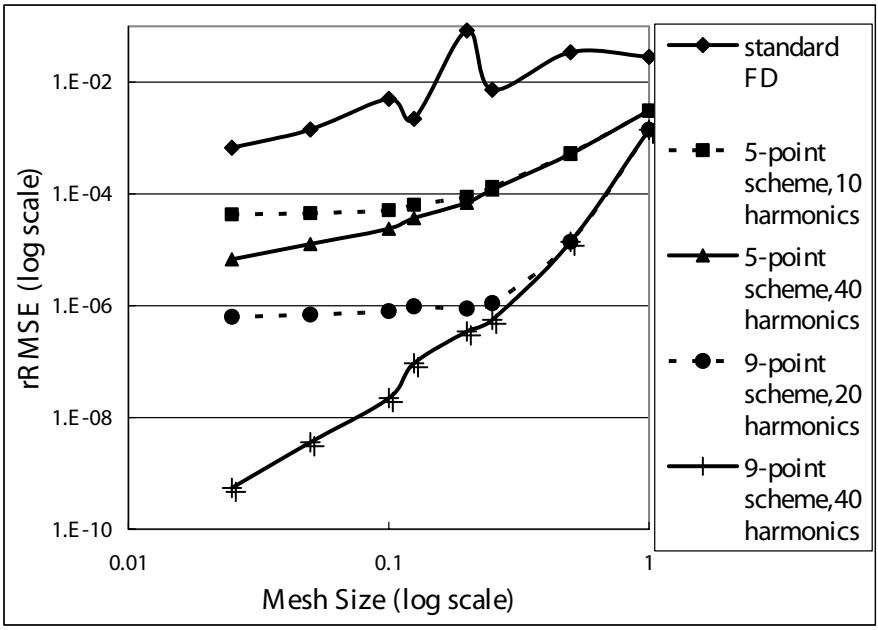


Fig. 6.10. Errors in the potential for standard FD and for FLAME with analytical bases by multipole-multicenter expansion. A 2D example. (Simulation by J. Dai.)

number of harmonics (10), the FLAME error decays only to some saturation level commensurate with the accuracy of the basis functions themselves. Similar observations are valid for the 9-point FLAME scheme (compare the error plots in Fig. 6.10 for 20 and 40 harmonics in the construction of the basis).

The accuracy of the 9-point scheme is obviously much higher than that of the 5-point scheme. From the numerical data, the asymptotic behavior of the error in the potential is approximately $\mathcal{O}(h^{1.6})$ for the 5-point scheme and $\mathcal{O}(h^{3.7})$ for the 9-point scheme.

Two FLAME basis functions computed by FEMLAB^{TM} (COMSOL Multiphysics) are plotted in Fig. 6.11. The functions correspond to the two particles with a small gap (0.1) between them.

Fig. 6.12 shows the FLAME simulation results with this kind of a basis. The number of FE degrees of freedom (d.o.f.) is a simulation parameter that affects the accuracy of the FE solution for the numerical FLAME basis. For the 5-point scheme, 5401 and 59,371 d.o.f. yield similar accuracy, which shows that the numerical error in this case is primarily due to the finite-difference (FLAME), rather than the finite-element, discretization.

The error plot for the 9-point scheme with 59,371 d.o.f. exhibits an anomaly. When the FLAME mesh size falls below 0.025, the accuracy deteriorates. This is because, due to the limited accuracy of the FE solution for the FLAME bases, the null space of matrix N^T (see Chapter 4) has dimension greater than

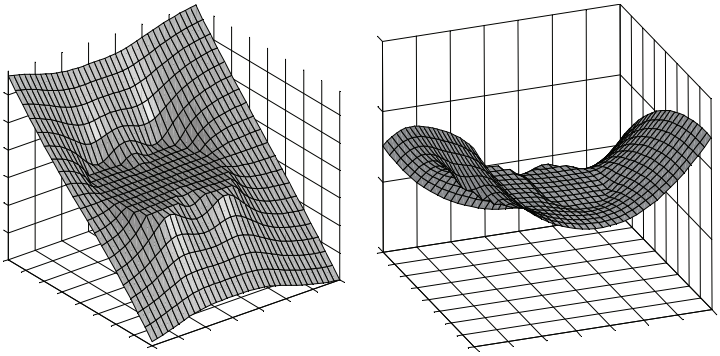


Fig. 6.11. Examples of FLAME basis functions, plotted vs. coordinates x, y. The functions are generated by FEM for a pair of nearby cylindrical particles. Left: basis function corresponding to an external applied field with potential $u_{\text{ext}} = y$. Right: $u_{\text{ext}} = x^2 - y^2$.

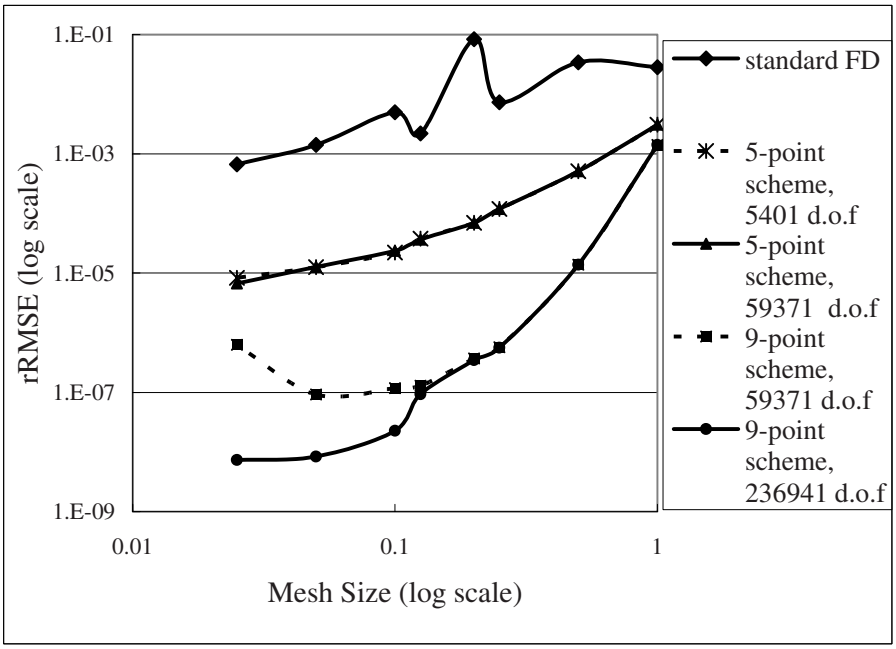


Fig. 6.12. Errors in the potential for standard FD and for FLAME with numerical (finite-element) bases. A 2D example. (Simulation by J. Dai.)

one in some patches. Fortunately, the dimension of the null space is easy to monitor; if it becomes greater than one, the accuracy of the local FE solution needs to be increased (via h- or p-refinement).

An interesting alternative for obtaining the local solutions could be the boundary element method. Although the matrices in this case are full, they can easily be handled due to their small size for each local problem. The advantage is that the local meshes are needed only on the interface boundaries. Possible applications of this type of technique are currently being explored.

Adaptive Refinement

The simulation results in this section are due to Jianhua Dai.

FLAME approximates the solution "patch"-wise (Chapter 4). In the areas where different patches overlap, the discrepancy between the corresponding values of the numerical solution may serve as a natural error indicator. Additional nodes can then be introduced in the regions where the error indicator is highest. This approach in FLAME is only beginning to be explored [DT07], but some computational examples in 2D can already be given.

A few cylindrical dielectric particles at randomly chosen locations in free space are immersed in a uniform external field. A quasi-analytical solution is obtained by the multipole-multicenter expansion and used for verification of the FLAME results.

Figs. 6.13 and 6.14 illustrate the geometric setup and the FLAME nodes after a step of adaptive refinement for two typical problems of this kind. The relative permittivity of all particles is 10. Note that the FLAME grid does not have to be regular Cartesian.

For 5-point FLAME schemes, the respective errors in the potential are given in Tables 6.1 and 6.2. It is encouraging that the adaptive refinement occurred at the "right" places – in the smaller gaps between the particles where the actual numerical error should definitely be expected to be higher. Results for 9-point schemes are qualitatively similar.

The discrepancy between the potential values at edge midpoints is used as an error indicator. For each midpoint, there are two such values from the two patches corresponding to the nodes of that edge. Further, the error indicator for each grid cell is taken to be the average of the indicators for its four edges. Although several grid sizes are seen in the figures, the actual refinement occurred in one step: grid cells with the highest error indicator are subdivided into 8×8 subcells, while their neighboring cells are subdivided into 4×4, and the next layer of neighbors into 2×2. Allowing more abrupt changes in the grid size would lead, as numerical experiments have shown, to much higher numerical errors.

Further results on adaptive FLAME for electrostatic problems and for electromagnetic wave scattering from multiple dielectric particles are reported by J. Dai & myself in [DT07].

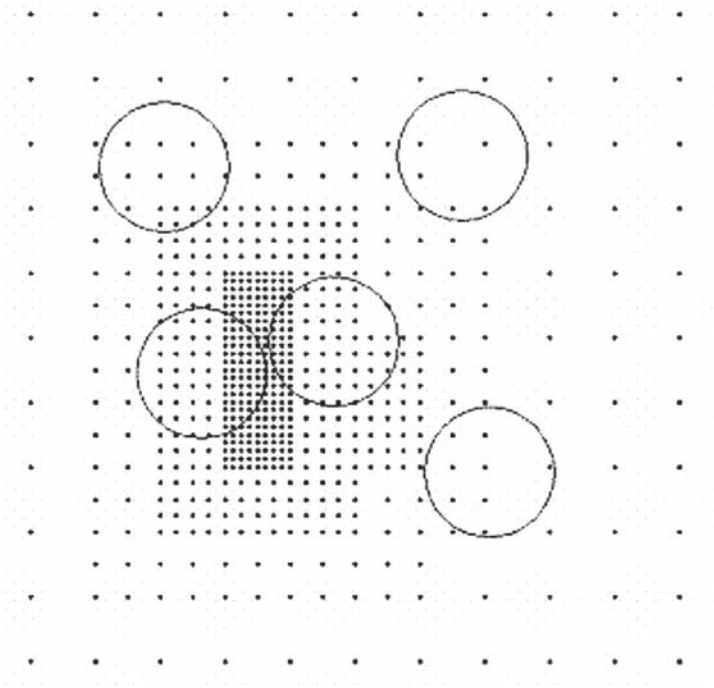


Fig. 6.13. FLAME nodes after refinement. (Credit: J. Dai.)

Table 6.1. Relative error in the potential before and after refinement, for the problem of Fig. 6.13. (Credit: J. Dai.)

	Before refinement	After refinement
Number of nodes	169	684
Relative rms error	7.01×10^{-2}	4.97×10^{-4}

Table 6.2. Relative error in the potential before and after refinement for the problem of Fig. 6.14. (Credit: J. Dai.)

	Before refinement	After refinement
Number of nodes	169	1123
Relative rms error	0.0882	2.8×10^{-4}

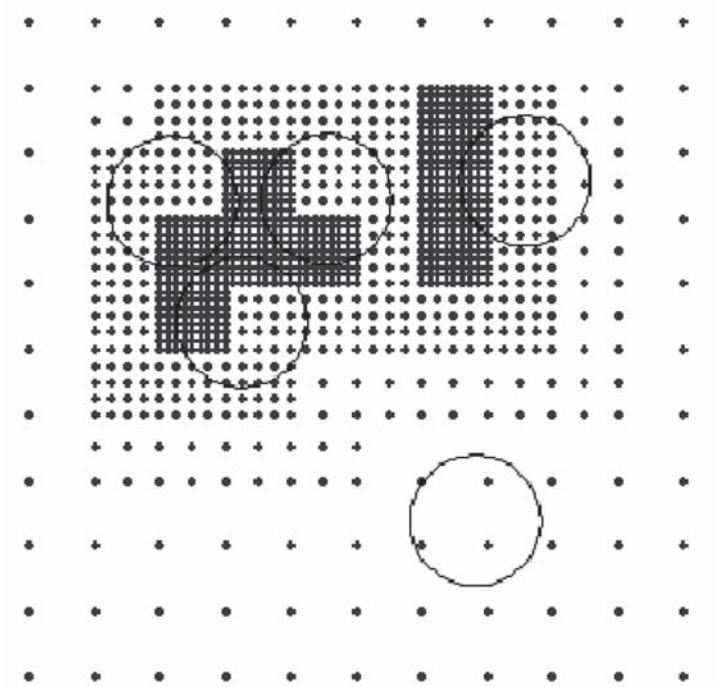


Fig. 6.14. FLAME nodes after refinement. (Credit: J. Dai.)

Summary

For electrostatic or magnetostatic problems with spherical particles, construction of analytical basis functions for FLAME is straightforward via spherical harmonics (Section 6.3.1; [Tsu05a, Tsu06]), provided that the particles are well separated. For particles in close proximity to one another, there are at least two ways of computing the basis functions. The first approach employs local multipole-multicenter expansions. The second way is purely numerical: the local FLAME bases are generated by the Finite Element Method. Note that solving a number of local FE problems is *much* less expensive computationally than solving the global problem, as no complicated meshes and no large FE systems of equations are involved.

Numerical examples demonstrate the high rate of convergence of five- and 9-point FLAME schemes in 2D and 7- and 19-point schemes in 3D. With the same mesh, the accuracy of FLAME is much higher than that of the standard FD–flux balance scheme. This may pave the way for solving problems with a large number of particles on relatively coarse grids, with mesh sizes comparable to or even greater than the radii of the particles and than the

separation distances between them. Thus *numerical* bases can be successfully used in FLAME when analytical ones are not available.

In FLAME, discrepancies between the numerical values of the potential in two overlapping "patches" may serve as a natural error indicator for grid refinement. In the numerical examples (p. 300), this indicator is effective: narrow gaps between particles are selected for refinement and the accuracy is increased by orders of magnitude as a result.

6.3 Static Fields of Spherical Particles in a Homogeneous Dielectric

6.3.1 FLAME Basis and the Scheme

Problems involving dielectric particles in an external dielectric medium arise, in particular, in the simulation of colloidal systems (J. Dobnikar *et al.* [DHM+04], M. Deserno *et al.* [DHM00]). Colloidal particles usually carry a surface electric charge that produces an electrostatic field. In some cases, particles can also be magnetic; controlling them by external magnetic fields may have interesting applications in some emerging areas of nanoscale technology (B. Yellen *et al.* [YF04, YFB04], A. Plaks *et al.* [PTFY03]). The material properties of the particles are usually quite different from those of the solvent. Computationally the problem is quite challenging due to many-body interactions and the heterogeneities.

In this section, 3D FLAME schemes are derived for particles in free space or a homogeneous dielectric. This is analogous to the 2D case considered previously. Solvent effects are dealt with in the following section.

For particles in a homogeneous dielectric, the electrostatic potential is again governed by the Laplace equation both inside and outside the particles, with the standard conditions at particle boundaries:

$$u_{\mathrm{in}}(r_{\mathrm{p}}) \;=\; u_{\mathrm{out}}(r_{\mathrm{p}}) \tag{6.21}$$

$$\epsilon_{\mathrm{in}}\,\frac{\partial u_{\mathrm{in}}}{\partial r} \;=\; \epsilon_{\mathrm{out}}\,\frac{\partial u_{\mathrm{out}}}{\partial r} \;+\; \rho_S, \qquad r = r_{\mathrm{p}} \tag{6.22}$$

These equations are almost the same as for cylindrical particles, (6.1), (6.2), except for the obvious differences in the geometric meaning of the the radial coordinate in the 2D and 3D cases.

The Trefftz–FLAME basis functions in the vicinity of a particle are obtained via spherical harmonics that satisfy the Laplace equation both inside and outside the particle:

$$\psi_{\alpha}^{(i)}(r,\theta,\phi) \;=\; r^n P_n^m(\cos\theta)\exp(\mathrm{i}m\phi) \tag{6.23}$$

inside the particle, and

$$\psi_\alpha^{(i)}(r,\theta,\phi) \;=\; (f_{mn}r^n + g_{mn}r^{-n-1})\, P_n^m(\cos\theta)\exp(\mathrm{i}m\phi) \tag{6.24}$$

outside the particle. Here index α is a single number corresponding to the (n,m) index pair; for example, α can be defined as $\alpha = n(n+1)+m$, $n = 0,1,\ldots$; $m = -n,\ldots,n$; $\alpha = 0,1,\ldots,n^2-1$. The r^n term *is* retained outside the particle because the harmonic expansion is considered locally, in a finite (and small) patch $\Omega^{(i)}$.

Remark 18. One may note a lack of symmetry between the inside and outside regions in this definition of the basis set. If the particle radius is much greater than the mesh size, it may indeed be desirable to restore the symmetry and add the harmonics with negative powers of r to the FLAME basis near the boundary, as the respective "patch" where the FLAME approximation is introduced in this case is away from $r = 0$. Another asymmetry is the lack of coefficients similar to f_{mn} inside the particles; this is just a convenient normalization of the basis functions.

In the above expressions for the basis functions, the standard notation for the associated Legendre polynomials P_n^m is used:

$$P_n^m(x) \;=\; (-1)^m(1-x^2)^{m/2}\,\frac{d^m P_n(x)}{dx^m}, \qquad -1 \le x \le 1 \tag{6.25}$$

The (regular) Legendre polynomials can be expressed, say, by the Rodrigues formula:

$$P_n(x) \;=\; \frac{1}{2^n n!}\,\frac{d^n}{dx^n}(x^2-1)^n, \qquad -1 \le x \le 1 \tag{6.26}$$

For reference, the first few of these polynomials are

$$P_0(x) = 1;\quad P_1(x) = x;\quad P_2(x) = \frac{1}{2}\,(3x^2-1);\quad P_3(x) = \frac{1}{2}\,(5x^3-3x)$$

$$P_0^0(x) = 1;\quad P_1^0(x) = x;\quad P_1^1(x) = -(1-x^2)^{1/2};$$

$$P_2^0(x) = \frac{1}{2}\,(3x^2-1);\quad P_2^1(x) = -3x(1-x^2)^{1/2};\quad P_2^2(x) = 3(1-x^2)$$

The coefficients of the FLAME scheme are derived for the *homogeneous* equation[11] – in the physical problem under consideration, for uncharged particles. The conditions at the particle boundary are satisfied for a suitable choice of coefficients f_{mn} and g_{mn}. Straightforward computation yields the first six basis functions of Table 6.3 [Tsu05a]. The coefficients in the table are

$$c_1 = \frac{\epsilon_{\mathrm{in}}+\epsilon_{\mathrm{out}}}{3\epsilon_{\mathrm{out}}};\quad c_2 = -r_{\mathrm{p}}^3\,\frac{\epsilon_{\mathrm{in}}-\epsilon_{\mathrm{out}}}{3\epsilon_{\mathrm{out}}};\quad b_1 = \frac{2\epsilon_{\mathrm{in}}+3\epsilon_{\mathrm{out}}}{5\epsilon_{\mathrm{out}}};\quad b_2 = -2r_{\mathrm{p}}^5\,\frac{\epsilon_{\mathrm{in}}-\epsilon_{\mathrm{out}}}{5\epsilon_{\mathrm{out}}}$$

[11] i.e. equation with a zero right hand side – not to be confused with a homogeneous medium.

Table 6.3. Trefftz–FLAME basis functions for a spherical particle. (The Poisson equation.)

Basis functions	Inside the particle	Outside the particle	Harmonic
ψ_1	1	1	
ψ_2	x	$xr^{-1}(c_1 r + c_2 r^{-2})$	dipole
ψ_3	y	$yr^{-1}(c_1 r + c_2 r^{-2})$	dipole
ψ_4	z	$zr^{-1}(c_1 r + c_2 r^{-2})$	dipole
ψ_5	$z^2 - x^2$	$(z^2 - x^2)r^{-2}(b_1 r^2 + b_2 r^{-3})$	quadrupole
ψ_6	$z^2 - y^2$	$(z^2 - y^2)r^{-2}(b_1 r^2 + b_2 r^{-3})$	quadrupole

For practical convenience, expressions for the basis functions were converted from spherical to Cartesian coordinates. For example, basis functions ψ_5 and ψ_6 are easily seen to be linear combinations of the following two spherical harmonics:

$$P_2^0(\cos\theta) \;=\; \frac{1}{2}\,(3\cos^2\theta - 1) \;=\; \frac{1}{2}\,r^{-2}\,(3\,z^2 - (x^2 + y^2 + z^2))$$

$$= \; r^{-2}\,(z^2 - \frac{1}{2}\,x^2 - \frac{1}{2}\,y^2)$$

and

$$P_2^2(\cos\theta)\cos 2\phi \;=\; 3\sin^2\theta\cos 2\phi \;=\; 3\sin^2\theta(2\cos^2\phi - 1)$$

$$= \; 3r^{-2}\,\big(2x^2 - (x^2 + y^2)\big) \;=\; 3r^{-2}\,(x^2 - y^2)$$

To construct a FLAME scheme, assume for definiteness that the standard 7-point stencil is used and that the set of six basis functions is chosen as specified in Table 6.3. The Trefftz–FLAME scheme $\underline{s}^{(i)} \in \mathbb{R}^7$ is then computed as the null space of the matrix of nodal values of these basis functions on a given stencil.

Remark 19. The choice of two quadrupole functions, $\psi_{5,6}$ of Table 6.3, out of possible five, is arbitrary. Numerical experience has shown that the particular choice of functions is not critical. Alternatively, one may drop the quadrupole functions altogether and keep only four functions ψ_{1-4} in the FLAME basis set. The null space of the nodal matrix is then three-dimensional, i.e. there are potentially three independent FLAME schemes available. One may be tempted to look for a linear combination of these schemes that would in some sense be optimal – for instance, would produce maximum diagonal dominance. However, this complicates the algorithm and does not lead, in my experience, to higher numerical accuracy of the solution.

If the particles are *charged*, the coefficients of the FLAME scheme (and hence the system matrix overall) remain unchanged, but the right hand side becomes nonzero. The difference equation in FLAME has the form (see e.g. (6.17) on p. 290):

$$\underline{s}^{(i)T}\underline{u}^{(i)} = \underline{s}^{(i)T}\mathcal{N}^{(i)}u_f^{(i)} \tag{6.27}$$

This is completely analogous to the construction of FLAME schemes for charged cylindrical particles – see Example 15 on p. 288. The particular solution $u_f^{(i)}$ is just the Coulomb potential

$$u_f^{(i)} = \begin{cases} q\,(4\pi\epsilon_{\text{out}}r_{\text{p}})^{-1}, & r \le r_{\text{p}} \\ q\,(4\pi\epsilon_{\text{out}}r)^{-1}, & r \ge r_{\text{p}} \end{cases} \tag{6.28}$$

where q is the charge of the particle and r is the distance from the center of the particle. More generally, $u_f^{(i)}$ could be a superposition of potentials (6.28) for several particles in the vicinity of the "patch" $\Omega^{(i)}$ containing stencil i. If a charged particle intersects with $\Omega^{(i)}$, the Coulomb potential of that particle *must* be included in $u_f^{(i)}$ – otherwise the field generated by that particle would not be accounted for. If a particle is near the stencil but does not intersect with its respective "patch," including the potential of that particle into $u_f^{(i)}$ is optional and in general constitutes a trade-off between accuracy and the computational cost.

The actual computation of the right hand side (6.27) is analogous to the numerical example for cylindrical particles (Example 15 on p. 288).

6.3.2 A Basic Example: Spherical Particle in Uniform Field

In this classical example, an uncharged polarizable spherical particle is immersed in a uniform external field. A simple analytical solution is readily available, and so the numerical errors of Trefftz–FLAME and its convergence can be easily analyzed. To eliminate the effects of domain truncation, the exact analytical solution is imposed as a Dirichlet condition in the Trefftz–FLAME system.

In the numerical example below, the computational domain is a unit cube and the radius of the particle is $r_{\text{p}} = 0.07$. The relative dielectric constants of the particle and surrounding dielectric are one and 80, respectively.

If the FLAME scheme is used everywhere in the computational domain, the numerical solution is exact (up to the roundoff error). Indeed, the exact solution contains only the dipole harmonic which lies in the functional space spanned by the FLAME basis functions; consistency error of the FLAME scheme is therefore zero. In practice for multiparticle problems, FLAME schemes are used in the vicinity of each particle, and any standard schemes for the Laplace equation can be used away from the particles. To make the one-particle example consistent with the multiparticle case, the FLAME scheme

is applied only within a certain threshold distance from the center of the particle.

In the numerical experiments, the standard 7-point stencil is used throughout the computational domain. If the midpoint of the stencil is within the threshold distance from the center of the particle, then the FLAME scheme is applied; otherwise the standard 7-point scheme for the Laplace equation is used. The standard scheme limits the overall convergence of the solution to $\mathcal{O}(h^2)$ asymptotically.

The relative numerical error in the potential is defined as in equation (6.19). Fig. 6.15 shows this error as a function of mesh size. The observed convergence rate is $\mathcal{O}(h^2)$ – as noted above, this asymptotic behavior is due to the bottleneck imposed by the standard difference scheme away from the particle. For comparison, convergence of the conventional flux-balance scheme is also shown; the FLAME solution is clearly superior. The figure also demonstrates that the field computed inside the particle exhibits very rapid convergence due to the exact representation of the potential in and near the particle by spherical harmonics.

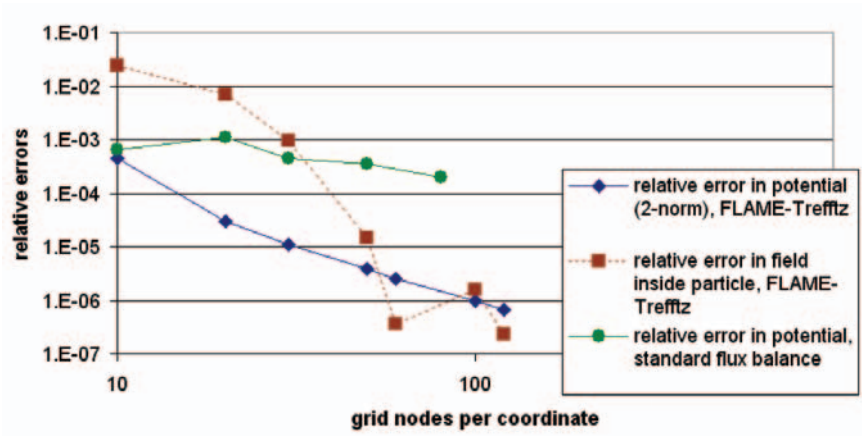


Fig. 6.15. Superior performance of FLAME for the test problem with a polarized spherical particle. The error in the potential in FLAME (diamonds) is much lower than for the standard flux-balance scheme (circles). Convergence of the field inside the particle (squares) is remarkably rapid. (Reprinted by permission from [Tsu05a] ©2005 IEEE.)

A 3D Test with Several Particles

The application of FLAME schemes to 3D *multi*particle problems in homogeneous dielectrics is conceptually analogous to the 2D case (Sections 6.2.2,

6.2.3 on pp. 291, 294). A 3D example includes four particles with the same radius $r_{\mathrm{p}} = 1$ and the dielectric constant $\epsilon_{\mathrm{p}} = 2$. The particles are immersed in a host medium with $\epsilon_{\mathrm{s}} = 80$. This resembles the typical case of colloidal particles in water, with no salt.

Two of the particles are in close proximity to one another, with a gap of 0.1459 between them. A uniform external field is applied. For comparison and verification, the analytical solution is obtained via the multipole-multicenter expansion (truncated at the terms with the magnitude below 10^{-8}). To eliminate the effects of domain truncation, the exact Dirichlet condition is applied at the exterior boundary.

Fig. 6.16 shows the simulation result for FLAME with the bases constructed by the multipole-multicenter expansion with 20 harmonics. The accuracy of FLAME is seen to be much higher than that of standard FD.

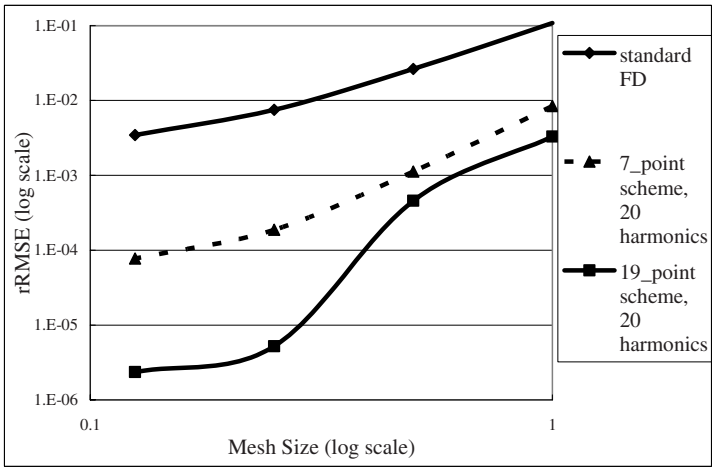


Fig. 6.16. Error in the potential for FLAME with multipole-multicenter basis functions in 3D.

The 19-point scheme[12] yields much higher accuracy than the 7-point scheme if the FLAME bases are computed with sufficient precision. Then the asymptotic error in the potential is $\sim \mathcal{O}(h^{1.5})$ for the 7-point scheme and $\sim \mathcal{O}(h^{3.5})$ for the 19-point scheme.

The remainder of this chapter focuses on a somewhat different, and arguably more interesting and complicated, problem: multiparticle interactions *in solvents*. The microions (e.g. salt ions) in the solvent redistribute themselves in the presence of any external field, which produces a screening effect.

[12] The 19-point stencil is a 3×3×3 cluster of nodes without the corner nodes – the same as for the "Mehrstellen" schemes in Section 4.4.5.

The electrostatic potential can then be described, at least for monovalent ions, by the Poisson–Boltzmann Equation (PBE).

6.4 Introduction to the Poisson–Boltzmann Model

This section reviews a classic problem dating back to the works of G. Gouy [Gou10] in 1910 and D. Chapman [Cha13] in 1913: a charged flat electrode immersed in a solvent.

There are two analogous but somewhat different cases. In the first one, the microions in the solvent are *counterions* dissolved from the electrode; for example, the electrode gives off protons H^+ or other positively charged chemical groups and acquires a negative surface charge ρ_S per unit area (Fig. 6.17). The whole system (electrode + solvent) is electrically neutral. In the second

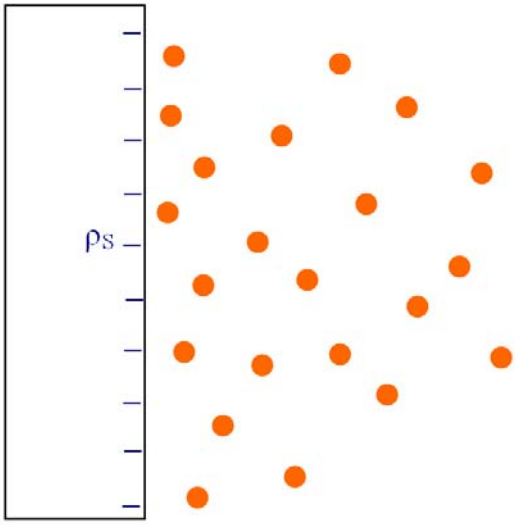


Fig. 6.17. A diffuse layer of cations in solvent near a flat electrode: the Gouy–Chapman problem.

case, the microions in the solvent are due to the presence of an electrolyte – they are salt ions. The solvent itself is electrically neutral as a whole.

In the first case (electrode with counterions), I follow very closely the presentation by A.Yu. Grosberg *et al.* [GNS02]. By assumption, the only counterions in the system are those dissociated from the surface. Since the electrode is large, the problem of the counterion distribution near the surface is treated

as one-dimensional. The electrostatic potential $u(x)$ satisfies the Poisson equation[13]

$$-\nabla^2 u \;=\; \frac{\rho}{\epsilon_{\rm s}} \tag{6.29}$$

where ρ is the volume charge density of the cations, and $\epsilon_{\rm s}$ is the dielectric constant of the solvent ($\epsilon_{\rm s} \approx 80\epsilon_0$ for water under static conditions). The key observation is that charge density, in return, depends on the potential, as the counterions are mobile and their concentration $n(x)$ is affected by the field. The *Boltzmann* distribution for the counterion concentration is assumed:

$$n(x) \;=\; n_S \exp\left(-eZ\,\frac{u(x)-u_S}{k_B T}\right) \tag{6.30}$$

where subscript "S" refers to the surface of the electrode, Ze is the charge of each counterion (e being the proton charge), and $k_B \approx 1.38065 \times 10^{-23}$ $\mathrm{m}^2 \times \mathrm{kg} \times \mathrm{s}^{-2} \times \mathrm{K}^{-1}$ is the Boltzmann constant.

Given this charge distribution, the Poisson equation becomes

$$u''(x) \;=\; -\,\frac{n_S e Z}{\epsilon_S} \exp\left(-eZ\,\frac{u(x)-u_S}{k_B T}\right) \tag{6.31}$$

where the primes denote x-derivatives. This 1D Poisson–Boltzmann equation is manifestly nonlinear (the unknown function u appears in the exponential). Luckily, this equation has an analytical solution that can be obtained in (at least) two different ways. A somewhat more systematic way is to multiply the equation by $2u'$, after which both sides turn into full derivatives – the left hand side becoming equal to $(u'^2)'$ – and the equation can be integrated.

A shortcut is to guess the form of the solution as

$$u(x) \;=\; a\log(x+\lambda) \;+\; b \tag{6.32}$$

substitute it into the equation and find parameters a, λ, b for which the equation and the boundary conditions are satisfied.

The boundary condition at $x = 0$ follows from the fact that the field vanishes inside the electrode:

$$u'(x) \;=\; -\,\frac{\rho_S}{\epsilon_s} \qquad \text{at } x = 0 \tag{6.33}$$

The second condition is that of global electroneutrality: the integral of concentration $n(x)$ per unit area must be normalized to $\rho_S/(Ze)$ ions. With all this in mind, ion concentration is found to be

$$n(x) \;=\; \frac{2\epsilon k_B T}{(eZ)^2}\,(x+\lambda)^{-2} \tag{6.34}$$

[13] The more conventional notation ϕ for the potential could be confused in this chapter with the angular coordinate in the spherical or cylindrical system.

The relevant physical parameters are the Gouy–Chapman length

$$\lambda = \frac{2\epsilon_s\, k_B T}{\rho_S\, eZ} \tag{6.35}$$

and the Bjerrum length

$$l_B = \frac{e^2}{4\pi\epsilon_s\, k_B T} \tag{6.36}$$

The Bjerrum length is the distance at which the energy of electrostatic interaction of two elementary charges is equal to thermal energy $k_B T$ ($l_B \approx 0.7$ nm in water at room temperature).[14]

Let us now consider a three-dimensional problem for an electrolyte, with positive and negative salt ions carrying equal and opposite charges. In general, there may be several species of ions, and consequently the Poisson–Boltzmann equation in general may contain several exponentials:

$$\epsilon_{\mathrm{s}} \nabla^2 u = -\sum_\alpha n_\alpha q_\alpha \exp\left(-\frac{q_\alpha u}{k_B T}\right) \tag{6.37}$$

where summation is over all species of ions present in the solvent, n_α is volume concentration of species α in the bulk, $q_\alpha = Z_\alpha e$ is the charge of species α; other parameters have the same meaning as before. The right hand side of (6.37) reflects the Boltzmann distribution of microions in the mean field with potential u. (More details are given in Section 6.11 and Appendix 6.14.)

For a 1:1 electrolyte, when all ions appear in pairs of opposite but equal-magnitude charges, the exponentials in the PBE can be paired up accordingly to produce the hyperbolic sine functions:

$$\epsilon_{\mathrm{s}} \nabla^2 u = 2 \sum_\beta n_\beta q_\beta \sinh\left(\frac{q_\beta u}{k_B T}\right) \tag{6.38}$$

where summation is now over all *pairs* of ions, and the summation index has been changed to β as a cue. This form of the Poisson–Boltzmann equation is slightly less general than (6.37).

If the electrostatic energy $q_\alpha u$ is (much) smaller than thermal energy $k_B T$, PBE can be approximately linearized around $u = 0$ to yield[15]

$$\epsilon_{\mathrm{s}} \nabla^2 u = -\sum_\alpha n_\alpha q_\alpha + \sum_\alpha n_\alpha q_\alpha \frac{q_\alpha u}{k_B T} \tag{6.39}$$

The first sum vanishes due to the global electroneutrality of the solution; hence

[14] The Bjerrum length is often defined without the factor of 4π in the denominator.

[15] For a detailed and systematic account of "optimal" linearization procedures, see M. Deserno *et al.* [DvG02], M. Bathe *et al.* [BGTR04] and references therein.

$$\epsilon_s \nabla^2 u \;=\; \sum_\alpha n_\alpha q_\alpha \frac{q_\alpha u}{k_B T} \tag{6.40}$$

or, in more compact form,

$$\nabla^2 u \;-\; \kappa^2 u \;=\; 0 \tag{6.41}$$

with

$$\kappa \;=\; (\epsilon_s k_B T)^{-\frac{1}{2}} \left(\sum_\alpha n_\alpha q_\alpha^2 \right)^{\frac{1}{2}} \tag{6.42}$$

κ is called the *Debye–Hückel parameter*.

It is useful to estimate the order of magnitude of the potential for which linearization is acceptable. Equating electrostatic energy qu of monovalent ions ($q = e$) to thermal energy $k_B T$, one obtains the threshold $u_{kT} = k_B T/e \approx 25$ mV at room temperature.

Equation (6.41) is known as the Debye–Hückel approximation. The potential satisfying this equation will typically exhibit an exponential decay with the characteristic length (the *Debye–Hückel length*) equal to the inverse of κ.

Example 16. Solution of the linearized PBE for an isolated charged ball in a solvent.

Due to spherical symmetry, it is natural to write the linearized PBE (6.40) in the solvent in the spherical coordinate system:

$$\frac{1}{r^2} \left(r^2 u' \right)' \;=\; \kappa^2 u \tag{6.43}$$

where the prime stands for the radial derivative. Anticipating the exponential decay, we write the unknown potential as

$$u(r) \;=\; \tilde{u}(r) \exp(-\kappa r) \tag{6.44}$$

where $\tilde{u}$ is a yet unknown function. Substituting this into equation (6.43) and simplifying, we find that

$$\tilde{u}(r) \;=\; c/r \tag{6.45}$$

where c is a constant to be determined. Thus

$$u(r) \;=\; \frac{c}{r} \exp(-\kappa r) \tag{6.46}$$

The constant can be found from Gauss's law on the surface of the ball:

$$-4\pi r^2 \epsilon_s u'(r) \;=\; q \qquad \text{at } r = r_0 \tag{6.47}$$

where q is the total charge of the ball and r_0 is its radius. The derivative of (6.44) is

$$u'(r) \;=\; -\, c r^{-2} \exp(-\kappa r) \;-\; c\kappa r^{-1} \exp(-\kappa r)$$

which yields

$$c = \frac{q\exp(\kappa r_0)}{4\pi\epsilon_s(1+\kappa r_0)} \tag{6.48}$$

The result is thus the Yukawa potential[16]

$$u(r) = \frac{q\ \exp(-\kappa(r-r_p))}{4\pi\epsilon_s r(\kappa r_p+1)} \tag{6.49}$$

6.5 Limitations of the PBE Model

This section follows the excellent exposition by T.T. Nguyen, A.Yu. Grosberg and B.I. Shklovskii [NGS00].

The main physical assumption behind the PBE is that each mobile charge is effectively in the *mean field* of all other charges, and has the Boltzmann probability of acquiring any given energy. This probability is assumed to be *unconditional*, i.e. not depending on possible redistribution of other ions in response to the motion of a given ion. In other words, mean field theory disregards any correlations between the positions and movement of the ions.

The following physical considerations illustrate that such correlations may in fact exist [NGS00]. A very simple arrangement of charges shown in Fig. 6.18

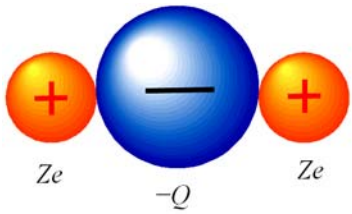


Fig. 6.18. (After T.T. Nguyen *et al.* [NGS00] ©2000 by the American Physical Society, with permission.) This simple system of charges may remain stable (at sufficiently low temperatures) even if the total charge is positive.

may remain stable even if the total charge of the two positive counterions exceeds the absolute value of the charge of the macroion ($2Ze > |Q|$). Indeed, the energy of each counterion Ze "attached" to the macroion $-Q$ is (omitting the $4\pi\epsilon$ factors for brevity)

$$\frac{(Ze)^2}{2(r_Q+r_Z)} - \frac{QZe}{r_Q+r_Z} = \frac{(Ze)^2 - 2QZe}{2(r_Q+r_Z)}$$

which remains negative as long as $Ze < 2Q$. That is, the charge of each counterion could be close to $2Q$, with the total charge of the system being

[16] Hideki Yukawa (1907–1981), winner of the Nobel Prize in physics (1949) for the theoretical prediction in 1934 of mesons – carriers of the nuclear force.

close to $2Q + 2Q - Q = +3Q$, and the system could still remain stable at sufficiently low temperatures.

Counterintuitively, the amount of charge condensed on a macroion can exceed, in some cases substantially, the charge of the macroion itself, leading to a possible "inversion" of charge. This is one of the effects that are not possible in the mean field theory – PBE framework.

However, there is now a consensus that at least for monovalent ions the correlations are weak enough for the PB model to be valid. In the remainder of this chapter, PBE is indeed assumed as the governing equation for the electrostatic potential in the solvent.

6.6 Numerical Methods for 3D Electrostatic Fields of Colloidal Particles

The typical sources of the electrostatic field in colloidal suspensions are surface charges on the particles. The boundary condition on the surface of each particle is

$$u_{\mathrm{s}} \;=\; u_{\mathrm{p}}; \qquad -\,\epsilon_{\mathrm{s}}\frac{\partial u_{\mathrm{s}}}{\partial r} \;+\; \epsilon_{\mathrm{p}}\frac{\partial u_{\mathrm{p}}}{\partial r} \;=\; \rho_S \quad \text{at } r = r_{\mathrm{p}} \tag{6.50}$$

where the surface charge density ρ_S is assumed to be known, r is the radial coordinate with respect to the center of the particle, and r_{p} is the radius of the particle.

Another boundary condition is $u = 0$ at infinity. In practice, this Dirichlet condition is imposed on the domain boundary taken sufficiently far away from the particles. Alternative boundary conditions (e.g. periodic or a superposition of Yukawa potentials) are possible but are not considered here.

In principle, several routes are available for the numerical simulation.

- First, if particle sizes are neglected and the governing Poisson–Boltzmann equation is linearized, the solution is simply the sum of the Yukawa potentials of all particles (Section 6.4). If the characteristic length of the exponential field decay (the Debye length) is small, the electrostatic interactions are effectively short-range and therefore inexpensive to compute. For weak ionic screening (long Debye lengths) Ewald-type methods can be used (G. Salin & J.-M. Caillol [SC00]).
- The Fast Multipole Method (FMM) is applicable under the same assumptions as above: the Yukawa potential of particles of negligible size. The FMM for this case is described in detail by L.F. Greengard & J. Huang [GH02].
- The Finite Element Method (FEM, Chapter 3) and the Generalized Finite Element Method (GFEM) (A. Plaks *et al.* [PTFY03], Section 4.5.2).
- A two-grid approach from computational fluid mechanics adapted to colloidal simulation (M. Fushiki [Fus92], J. Dobnikar *et al.* [DHM$^+$04]): a

spherical mesh around each particle and a common Cartesian background grid.

- The Flexible Local Approximation MEthod (FLAME, Chapter 4, Section 6.7; [Tsu05a, Tsu06]).

The focus of this section is on algorithms that would be applicable to finite-size particles and extendable to nonlinear problems. Ewald-type methods and FMM are not effective in such cases. FEM requires very complex meshing and re-meshing even for a modest number of moving particles (say, on the order of a hundred) and quickly becomes impractical when the number of particles grows. In addition, re-meshing is known to introduce a spurious numerical component in force calculation (see e.g. [Tsu95] and references therein).

GFEM relaxes the restrictions of geometric conformity in FEM by allowing suitable non-polynomial approximating functions to be included in the approximating set. This has been extensively discussed in the literature (M. Griebel & M.A. Schweitzer [GS00, GS02a], T. Strouboulis *et al.* [SBC00], I. Babuška & J.M. Melenk [BM97], I. Babuška *et al.* [BBO03], L. Proekt & I. Tsukerman [PT02], A. Plaks *et al.* [PTFY03], A. Basermann & I. Tsukerman [BT05]). Unfortunately, GFEM has a substantial overhead due to numerical quadratures in geometrically complex domains (such as intersections of spheres and hexahedra) as well as to a higher number of degrees of freedom in generalized finite elements around the particles (A. Plaks *et al.* [PTFY03]).

This leaves two main contenders: the two-grid approach and FLAME. For the former, the potential has to be interpolated back and forth between the local mesh of each particle and the common Cartesian grid; the numerical loss of accuracy in this process is unavoidable. FLAME has only one global Cartesian grid but produces an accurate difference scheme by incorporating *local* approximations of the potential near each particle into the scheme. The Cartesian grid can remain relatively coarse – on the order of the particle radius or even coarser. In contrast, classical FD schemes need the grid size much smaller than the particle radius to avoid the spurious "staircase" effects.

6.7 3D FLAME Schemes for Particles in Solvent

In the presence of an electrolyte, a Trefftz–FLAME basis can also be generated by matching spherical harmonic expansions inside and outside the particle. This is done by analogy with Section 6.3. (See also Remark 19 on p. 305.) The difference is that the FLAME basis in the electrolyte involves spherical Bessel functions rather than the powers of r as in a simple dielectric. Namely, expressions for the FLAME basis functions are

$$\psi_\alpha(r,\theta,\phi) \;=\; \begin{cases} r^n Y_{nm}(\theta,\phi), & r \le r_\mathrm{p} \\ (f_{nm} j_n(\mathrm{i}\kappa r) + g_{nm} n_n(\mathrm{i}\kappa r)) Y_{nm}(\theta,\phi), & r \ge r_\mathrm{p} \end{cases} \tag{6.51}$$

where Y_{nm} are the spherical harmonics and r_p is the radius of the particle. The spherical Bessel functions $j_n(\mathrm{i}\kappa r)$ and $n_n(\mathrm{i}\kappa r)$ in (6.51) are expressible in

terms of hyperbolic sines/cosines and hence relatively easy to work with. As in Section 6.3.1, index α is a single number corresponding to the index pair (n, m).

The coefficients f_{nm}, g_{nm} can be determined from the boundary conditions (6.50), by analogy with a similar calculation in Section 6.3.1. Expressions for coefficients f_{nm}, g_{nm} are summarized in Table 6.4 [Tsu05a].

Table 6.4. Trefftz–FLAME basis functions for a spherical particle. (The Poisson–Boltzmann equation.)

Basis functions	Inside the particle	Outside the particle	Harmonic
ψ_1	1	$w^{-1}[w_0 \cosh(w - w_0) + \sinh(w - w_0)]$	
ψ_2	x	$x(rw^2)^{-1}[c_1(w \cosh w - \sinh w)$ $-c_2(w \sinh w - \cosh w)]$	dipole
ψ_3	y	$y(rw^2)^{-1}[c_1(w \cosh w - \sinh w)$ $-c_2(w \sinh w - \cosh w)]$	dipole
ψ_4	z	$z(rw^2)^{-1}[c_1(w \cosh w - \sinh w)$ $-c_2(w \sinh w - \cosh w)]$	dipole dipole
ψ_5	$z^2 - x^2$	$(z^2 - x^2)(r^2 w^3)^{-1}[-b_1((3 + w^2) \sinh w -$ $3w \cosh w)$ $+ b_2((3 + w^2) \cosh w - 3w \sinh w)]$	quadrupole
ψ_6	$z^2 - y^2$	$(z^2 - y^2)(r^2 w^3)^{-1}[-b_1((3 + w^2) \sinh w -$ $3w \cosh w)$ $+ b_2((3 + w^2) \cosh w - 3w \sinh w)]$	quadrupole

In the Table, $w = \kappa r$, $w_0 = \kappa r_{\mathrm{p}}$. The coefficients b, c are as follows [Tsu05a]: $c_1 = (\epsilon_s w_0^2 \cosh w_0 + \epsilon_{\mathrm{p}} \cosh w_0 + 2\epsilon_s \cosh w_0 - w_0 \epsilon_{\mathrm{p}} \sinh w_0 - 2w_0 \epsilon_s \sinh w_0)/(\epsilon_s \kappa)$; $c_2 = (\epsilon_{\mathrm{p}} \sinh w_0 - \epsilon_{\mathrm{p}} w_0 \cosh w_0 + \epsilon_s w_0^2 \sinh w_0 + 2\epsilon_s \sinh w_0 - 2\epsilon_s w_0 \cosh w_0)/(\epsilon_s \kappa)$; $b_1 = (6\epsilon_{\mathrm{p}} w_0 \sinh w_0 - 4\epsilon_s w_0^2 \cosh w_0 - 2\epsilon_{\mathrm{p}} w_0^2 \cosh w_0 - 6\epsilon_{\mathrm{p}} \cosh w_0 - 9\epsilon_s \cosh w_0 + \epsilon_s w_0^3 \sinh w_0 + 9\epsilon_s w_0 \sinh w_0)/(\epsilon_s \kappa^2)$; $b_2 = (6\epsilon_{\mathrm{p}} w_0 \cosh w_0 - 4\epsilon_s w_0^2 \sinh w_0 - 2\epsilon_{\mathrm{p}} w_0^2 \sinh w_0 - 6\epsilon_{\mathrm{p}} \sinh w_0 - 9\epsilon_s \sinh w_0 + \epsilon_s w_0^3 \cosh w_0 + 9\epsilon_s w_0 \cosh w_0)/(\epsilon_s \kappa^2)$

For the 7-point stencil, one gets a valid FLAME scheme by adopting six basis functions: one "monopole" term ($n = 0$), three dipole terms ($n = 1$) and any two quadrupole harmonics ($n = 2$). Away from the particles, the classical 7-point scheme for the Helmholtz equation is used, even though a Trefftz–FLAME scheme can also be obtained using six local exponentially decaying solutions of the linearized PBE as the FLAME basis.

As explained in Chapter 4 (see p. 203), for inhomogeneous equations of the form

$$Lu = f \quad \text{in } \Omega^{(i)} \tag{6.52}$$

the FLAME scheme is constructed by splitting the potential up into a particular solution $u_f^{(i)}$ of the inhomogeneous equation and the remainder $u_0^{(i)}$ satisfying the homogeneous one:

$$u \;=\; u_0^{(i)} + u_f^{(i)}; \qquad Lu_0^{(i)} = 0; \qquad Lu_f^{(i)} \;=\; f \tag{6.53}$$

Superscript (i) indicates that the splitting is *local*, i.e. it needs to be introduced only within its respective subdomain $\Omega^{(i)}$ containing the grid stencil around node i. The difference scheme for the inhomogeneous equation is (Chapter 4)

$$L_h^{(i)} \underline{u}_h \;=\; L_h^{(i)} \mathcal{N}^{(i)} u_f \tag{6.54}$$

where $\mathcal{N}^{(i)}$ denotes the vector of nodal values of a continuous function on stencil i. For the linearized PBE, u_f can be taken as the Yukawa potential

$$u_f \;=\; \begin{cases} q\,\left[4\pi\epsilon_{\mathrm{s}} r_p(\kappa r_p + 1)\right]^{-1} & r \le r_p \\ q\,\exp(-\kappa(r - r_p))\,\left[4\pi\epsilon_{\mathrm{s}} r(\kappa r_p + 1)\right]^{-1} & r \ge r_p \end{cases} \tag{6.55}$$

Indeed, it is straightforward to verify that this potential satisfies the PBE in the solvent, the Laplace equation (in a trivial way as a constant) inside the particle, and the boundary conditions.

To summarize, the FLAME scheme in the vicinity of charged particles is constructed as follows:

1. Compute the FLAME coefficients for the homogeneous equation. For each grid stencil, this gives the nonzero entries of the corresponding row of the global system matrix.
2. Apply the scheme to the Yukawa potential to get the entry in the right hand side, as prescribed by (4.25) on p. 204.

Away from the particle (in practice, 2–3 grid layers from its surface), splitting (6.53) does not have to be introduced. If it isn't, it does not mean that the source field of the particle is somehow ignored; this field is just not *explicitly built into* the scheme. The following simple 1D example may help to clarify the matter.

Example 17. Consider the following one-dimensional analog of the Poisson–Boltzmann problem:

$$\begin{aligned} u''(x) \;&=\; 0, \quad x \le a \\ u''(x) \;-\; \kappa^2 u \;&=\; 0, \quad a \le x \le L \\ u'(a^+) \;-\; u'(a^-) \;&=\; -\,\rho \\ u'(0) \;=\; 0; \quad u(L) \;&=\; 0 \end{aligned} \tag{6.56}$$

The computational domain is $[0, L]$; potential u is governed by the Laplace equation inside the "particle" $[0, a]$ and by the Helmholtz equation in the rest of the domain. The derivative jump condition at $x = a$ is analogous to the boundary condition on the surface of a charged particle.

Let the FLAME scheme be constructed on the standard three-point stencil of a uniform grid with size h. The coefficient vector $\underline{s} \in \mathbb{R}^3$ of the scheme is (see equation (4.38) in Section 4.4.1, p. 207)

$$\underline{s} \;=\; (1,\; -2\cosh\kappa h,\; 1)^T \tag{6.57}$$

Hence the FLAME difference equation away from the "particle," and with no potential splitting, is

$$u_{i-1} \;-\; 2\cosh(\kappa h)\, u_i \;+\; u_{i+1} \;=\; 0 \tag{6.58}$$

for three stencil points $i-1$, i, and $i+1$.

Let us now introduce, over stencil i, the splitting

$$u \;=\; u_0^{(i)} \;+\; u_f^{(i)} \tag{6.59}$$

where the inhomogeneous part

$$u_f^{(i)} \;=\; u^{\text{1D}}_{\text{Yukawa}} \;\equiv\; \begin{cases} \rho\kappa^{-1}, & x \le a \\ \rho\kappa^{-1}\exp(-\kappa(x-a)), & x \ge a \end{cases} \tag{6.60}$$

The FLAME scheme *with the potential splitting* is

$$u_{i-1} \;-\; 2\cosh(\kappa h)\, u_i \;+\; u_{i+1} \;=\; u_f(x_{i-1}) \;-\; 2\cosh(\kappa h)\, u_f(x_i) \;+\; u_f(x_{i+1}) \tag{6.61}$$

where superscript (i) has been dropped, as u_f in this example is taken the same for all stencils.

Now compare the schemes with and without the potential splitting. Scheme (6.58) – without the splitting – is valid only for the homogeneous equation, i.e. for stencils away from the "particle". Scheme (6.61) is valid everywhere.

If the stencil is completely outside the particle, both schemes (6.58), (6.61) are consistent and either one of them can be used – in fact, in this 1D example these two schemes happen to be identical because the right hand side of (6.61) is zero. This can be verified directly by substituting u_f (6.60) into (6.61), but the fundamental reason for the zero right hand side is that u_f in this case lies in the functional space spanned by the FLAME basis functions $\exp(\pm\kappa x)$.

This accidental feature of the 1D example should be brushed aside, as the goal is to illustrate the idea of potential splitting. In 3D problems, the space of (local) solutions of the homogeneous equation is infinite-dimensional, and therefore u_f in source-free regions cannot be expected to lie in the finite-dimensional FLAME space. The right hand side of the scheme analogous to (6.61) is then in general nonzero, and the schemes with and without the potential splitting are different. Both schemes are consistent in source-free regions; the scheme with the potential splitting is consistent everywhere.

6.8 The Numerical Treatment of Nonlinearity

In this section, the general Newton–Raphson–Kantorovich procedure for nonlinear FLAME schemes (see Section 4.3.4, p. 203) is specialized to the Poisson–Boltzmann equation. It will still be convenient, up to a point, to use the generic operator notation

$$\mathcal{L}u \;=\; f \tag{6.62}$$

It is helpful to treat u and f as generalized functions (distributions, Appendix 6.15) to account for surface charges; this eliminates the need to consider surface boundary conditions as separate equations. The P–B operator, in its hyperbolic sine version, is (see (6.38))

$$\mathcal{L}u \;=\; \epsilon_{\mathrm{s}}\nabla^2 u \;-\; 2\sum_{\beta} n_\beta q_\beta \sinh\left(\frac{q_\beta u}{k_B T}\right) \tag{6.63}$$

and the right hand side is

$$f \;=\; -\,\rho_S \delta_S \tag{6.64}$$

Here n is the exterior normal to the surface; δ_S is the Dirac-type surface δ-function defined formally as the linear functional

$$\langle \delta_S,\ \psi\rangle \;=\; \int_S \psi\, dS \tag{6.65}$$

for any smooth "test" function ψ (see Appendix 6.15).

The scene is now set for the N–R–K iterations. Given approximation u_m of the exact solution at iteration m, one constructs the subsequent approximation $u_{m+1} = u_m + \delta u_m$ using the linearization

$$\mathcal{L}(u_m + \delta u_m) \;\approx\; \mathcal{L}u_m + \mathcal{L}'(u_m)\,\delta u_m \tag{6.66}$$

where $\mathcal{L}'$ is the Fréchet derivative (see Appendix 4.9). Equating the right hand side of (6.66) to f, one finds the approximate increment δu_m by solving

$$\mathcal{L}'(u_m)\,\delta u_m \;=\; R_m \;\equiv\; f \;-\; \mathcal{L}u_m \tag{6.67}$$

Residual R_m characterizes the accuracy of the m-th approximation to the solution.

In the colloidal problem, the equation within the particles is linear, which simplifies the implementation of the N–R–K procedure. To elaborate, let us write out a natural splitting of the P–B operator into its linear and nonlinear parts:

$$\mathcal{L}u \;\equiv\; \mathcal{L}_{\mathrm{lin}}u \;+\; \mathcal{L}_{\mathrm{nonlin}}u \tag{6.68}$$

where

$$\mathcal{L}_{\mathrm{lin}}u \;\equiv\; \nabla\cdot\epsilon_{\mathrm{s}}\nabla u \tag{6.69}$$

$$\mathcal{L}_{\text{nonlin}} u \;\equiv\; -2 \sum_{\beta} n_\beta q_\beta \sinh\left(\frac{q_\beta u}{k_B T}\right) \tag{6.70}$$

Importantly, the nonlinear part vanishes inside each particle:

$$\mathcal{L}_{\text{nonlin}} u(r) \;=\; 0, \quad r \in \Omega_{\text{p}}^{(k)}, \quad \text{for each particle } k \tag{6.71}$$

Inside the particles, where the operator is linear, the N–R–K residual gets annihilated after the first iteration. Indeed, for $m \geq 0$,

$$R_{m+1} \;=\; f - \mathcal{L} u_{m+1} \;=\; f - \mathcal{L}(u_m + \delta u_m) \;=\; R_m - \mathcal{L}\delta u_m \;=\; 0 \text{ in } \Omega_{\text{p}}^{(k)} \tag{6.72}$$

This chain of equalities relies on the linearity of $\mathcal{L}$ inside the particles. In particular, the very last equality is due to the definition (6.67) of δu_m and due to the fact that the Fréchet derivative of a linear operator is that operator itself, so $\mathcal{L}' = \mathcal{L}$ inside the particles.

The N–R–K residual is thus nonzero only strictly within the solvent, due to the nonlinearity of the P–B operator there. Notably, the residual does *not* contain the Dirac-delta term on the particle surface.[17] This implies that the increment δu_m satisfies the *homogeneous* condition on the surface; that is, δu_m does not "see" the surface charge. The right hand side of (6.67) thus contains only regular derivatives and no δ-functions:

$$R_m \;=\; \begin{cases} -\{\mathcal{L} u_m\}, & \text{in the solvent} \\ 0, & \text{inside the particles} \end{cases}, \quad m = 1, 2, \ldots \tag{6.73}$$

where the curly brackets stand for the classical (nondistributional) derivative.[18]

Thus each N–R–K iteration involves a linearized PBE with equivalent "sources" R_m (6.73) *in the solvent only.*[19] To apply FLAME to this equation, one splits the unknown function δu_m up into a homogeneous part $\delta u_m^{(0)}$ that can be approximated by the FLAME basis functions and a particular solution $\delta u_m^{(\text{p})}$ that satisfies the inhomogeneous equation:[20]

$$\delta u_m \;=\; \delta u_m^{(0)} \;+\; \delta u_m^{(\text{p})} \tag{6.74}$$

where

[17] The very first N–R–K iteration may be an exception, if the initial approximation u_0 does not satisfy the boundary condition for the jump of the normal derivative on the surface.

[18] This notation is due to V.S. Vladimirov [Vla84]. See Appendix 6.15.

[19] This linearization is purely *local.*

[20] FLAME schemes are always constructed "patch-wise," and the potential splitting is considered within a single patch containing a given node stencil. This is implicitly understood but not explicitly indicated for brevity. The local nature of the potential splitting also implies that this splitting is unaffected by the conditions on the exterior boundary of the domain.

$$\mathcal{L}'(u_m)\,\delta u_m^{(0)} \;=\; 0, \quad \mathcal{L}'(u_m)\,\delta u_m^{(\mathrm{p})} \;=\; -\{\mathcal{L}u_m\} \tag{6.75}$$

The particular solution can be defined by a Yukawa-like expression

$$\delta u_m^{(\mathrm{p})} \;=\; q\,\exp(-\kappa(r-r_p))\,\left[4\pi\epsilon_{\mathrm{s}} r(\kappa r_p+1)\right]^{-1} \tag{6.76}$$

In contrast with the usual Yukawa potential, parameter κ may now be different for different "patches" (which, however, is not explicitly indicated in the expression, to keep the notation simple).

Thus u_m is a combination of the Yukawa-like potential and FLAME basis functions. Because of that, and due to the nonlinearity of operator $\mathcal{L}$, the actual expression for the residual $\{\mathcal{L}u_m\}$ is complicated. Although the exact analytical representation for $\delta u_m^{(\mathrm{p})}$ can in principle be found with any degree of accuracy by, say, local expansion into spherical harmonics, let us retain only the zero-order term in $\{\mathcal{L}u_m\}$ for practical simplicity:

$$\mathcal{L}(u_m) \;=\; \nabla\cdot\epsilon_{\mathrm{s}}\nabla u_m - 2\sum_\beta n_\beta q_\beta \sinh\left(\frac{q_\beta u_m}{k_B T}\right) \;\approx\; L_{m0} = \mathrm{const} \tag{6.77}$$

This zero-order (i.e. constant) approximation can be found by evaluating $\mathcal{L}(u_m)$ at any given point in the solvent within the patch (e.g. at the central node of the stencil if it happens to lie in the solvent).

The derivative $\mathcal{L}'$ has the form (Appendix 4.9, p. 237)

$$\mathcal{L}'(u_m) \;=\; \nabla\cdot\epsilon_{\mathrm{s}}\nabla \;-\; \kappa^2 \tag{6.78}$$

where

$$\kappa^2 \;=\; \frac{1}{k_B T}\sum_\beta n_\beta q_\beta^2 \cosh\left(\frac{q_\beta u_m}{k_B T}\right)$$

Parameter κ depends on the potential and hence on coordinates. With the approximation limited to zero order, $\kappa \approx \kappa_0 = \mathrm{const}$ within the given patch; then the particular solution is also a constant and is equal to, due to the continuity of the solution across the particle boundary,

$$\delta u_m^{(\mathrm{p})} \;\approx\; L_{m0}\kappa_0^{-2} \;\; \text{everywhere within a given patch} \tag{6.79}$$

With the particular solution so defined, construction of the FLAME scheme at each N–R–K iteration follows the guidelines of Chapter 4, Section 4.3.4 (p. 203).

6.9 The DLVO Expression for Electrostatic Energy and Forces

The classical Derjaguin–Landau–Verwey–Overbeek (DLVO) [DL41, VO48] theory describes colloidal interactions and stability of colloidal systems. DLVO

has been used widely and successfully for many years. This section outlines the treatment of electrostatic interactions in the DLVO model. Short-range attractive forces are briefly commented on in the following subsection. In Section 6.9, the analytical formulas for the electrostatic potential and forces between colloidal particles (E.J.W. Verwey & J.Th.G. Overbeek [VO48], Chapter X) are used for comparison and validation of the FLAME results. The following physical assumptions are made in the DLVO analysis:

- The electrolyte is a simple dielectric with a given constant permittivity.
- The Boltzmann distribution applies to the microions in the electrostatic field. Hence the potential is governed by the Poisson–Boltzmann equation. Furthermore, the potential is sufficiently small so that the PBE can be linearized.
- The potential is constant over the surface of each particle.

The linearity assumption is essential for any analytical study, because a closed-form solution for the nonlinear PBE is only available for the simplest geometry: an infinite plane electrode (Section 6.4). A semi-analytical solution exists for a charged rod (M. Deserno *et al.* [DHM00]). J.E. Sader [Sad97] derived an approximate analytical solution for a charged sphere in an electrolyte.

While the assumption of constant surface potential simplifies the problem, the analytically more complicated case of constant surface charge can also be handled: potential in the solvent is sought as a superposition of two multipole expansions centered at the first and the second particle, respectively. For the Laplace equation (i.e. no electrolyte) this procedure is relatively straightforward. For the linearized Poisson–Boltzmann equation outside the particles and the Laplace equation inside, the spherical harmonics are more complicated (spherical Bessel functions of the radial coordinate within the solvent), and the relevant translation formulas for the harmonic expansion from one center to another are quite involved (M. Danos & L.C. Maximon [DM65], L.F. Greengard & J. Huang[21] [GH02], N. Gumerov & R. Duraiswami [GD03]). An analytical solution without the full formalism of multipole translations, both for constant surface charge and constant surface potential, was worked out by H. Ohshima [Ohs94a, Ohs94b, Ohs95].

Verwey & Overbeek [VO48] argue that in practice the difference between the constant potential and constant surface charge cases is small. They derive the (now classical) analytical result for energy and forces between two colloidal particles under the assumption of constant surface potential. Then, since the potential is governed by the Laplace equation inside the particles, it must be constant within each particle. In the solvent, the potential is assumed to satisfy the linearized Poisson–Boltzmann equation with known Dirichlet boundary conditions on particle surfaces.

Let the z axis pass through the centers of the two particles. Since the potential distribution is axially symmetric, each multipole (MP) expansion

[21] I am grateful to Jingfang Huang for his very helpful comments.

has the form

$$u_{\mathrm{MP}}^{(\alpha)} = \sum_{n=0}^{\infty} c_n^{(\alpha)} P_n(\cos\theta_\alpha)\, k_n(\kappa r_\alpha), \quad \alpha = 1, 2 \tag{6.80}$$

where r_α, θ_α are the spherical coordinates *with respect to the center of particle* α (α = 1,2); $c_n^{(\alpha)}$ are some coefficients; P_n is the Legendre polynomial and

$$k_n(z) = \left(\frac{2}{\pi z}\right)^{\frac{1}{2}} K_{n+\frac{1}{2}}(z)$$

is the modified spherical Bessel function.[22] The first three of these functions (n = 0, 1, 2) are

$$\begin{aligned} k_0(z) &= \frac{1}{z}\exp(-z) \\ k_1(z) &= \frac{z+1}{z^2}\exp(-z) \\ k_2(z) &= \frac{z^2+3z+3}{z^3}\exp(-z) \end{aligned}$$

The two-center multipole approximation of the (total) potential in the solvent is simply

$$u_{\mathrm{MP}} = u_{\mathrm{MP}}^{(1)} + u_{\mathrm{MP}}^{(2)} \tag{6.81}$$

If the two particles are identical, the coefficients $c_n^{(1)}$ and $c_n^{(2)}$ are equal for all n and superscript (α) can therefore be dropped. These coefficients must be such that the Dirichlet conditions on the particle boundaries are satisfied. The Galerkin method can be applied to find the coefficients:

$$\int_S u_{\mathrm{MP}} P_n(\cos\theta_\alpha)\, dS = u_S \int_S P_n(\cos\theta_\alpha)\, dS, \quad n = 0, 1, \ldots \tag{6.82}$$

where u_S is the known potential on the surface S of either of the particles. In the Galerkin method, by definition, the test functions are the same as the basis functions – in this case, the Legendre polynomials. The multipole potential u_{MP} (6.81) includes contributions from both particles and contains all unknown coefficients c_n.[23] Integration of spherical harmonics associated with one of the particles over this particle's surface is very simple. Integration of harmonics associated with the other particle is more technical. Today such computation is routine in Fast Multipole Methods (see e.g. H. Cheng *et al.* [CGR99]); Verwey & Overbeek derived their result directly.

[22] See G. Arfken [Arf85] or Eric W. Weisstein. "Modified Spherical Bessel Function of the Second Kind." From MathWorld – A Wolfram Web Resource. http://mathworld.wolfram.com/ModifiedSphericalBesselFunctionoftheSecond-Kind.html

[23] Verwey & Overbeek [VO48] denote these coefficients with λ.

The system of Galerkin equations (6.82) is infinite, and a practical approximation is obtained in DLVO by truncating the expansion to three terms ($n = 0, 1, 2$). Once the algebraic details are worked out, the DLVO expression for the energy of electrostatic interaction of two colloidal particles is found to be ([VO48], pp. 149–159)

$$W(\tilde{r}) \approx \psi_0^2 \, \epsilon_\mathrm{s} r_\mathrm{p} \, \frac{\exp(-\kappa r_\mathrm{p}(\tilde{r} - 2))}{4\pi\tilde{r}} \, \beta, \qquad \tilde{r} \equiv \frac{r}{r_\mathrm{p}} \tag{6.83}$$

Here, as before, r_p is the radius of each particle, ψ_0 is the surface potential of each particle (with its variation over the surface neglected), κ is the Debye–Hückel parameter (6.42), ϵ_s is the (absolute) dielectric permittivity of the solvent, and β is a coefficient (not to be confused with $1/k_B T$). The factor of 4π is present in (6.83) but not in [VO48] due to a difference in the system of units.

Unfortunately, both ψ_0 and β depend on the spherical-harmonic coefficients c_n. These coefficients are obtained by solving the Galerkin system and are not described by simple analytical formulas. Parameter β, tabulated in [VO48], always lies in the range $0.6 \leq \beta \leq 1$, being close to unity for large separations between the particles and approaching 0.6 for small separations. If, as a practical approximation, this factor is dropped, the interaction energy is *over*estimated by a coefficient not much greater than one. If further simplification is made by replacing ψ_0 with the Yukawa potential on the surface of a single particle (thereby neglecting the contribution of the other particle), the energy is *under*estimated by a similar factor. Taken together, the two simplifications produce a very useful and accurate expression for the energy of electrostatic interaction:

$$W(r) \approx \frac{\exp(2\kappa r_\mathrm{p})}{(1 + \kappa r_\mathrm{p})^2} \, \frac{q^2}{4\pi\epsilon_\mathrm{s} r} \exp(-\kappa r_\mathrm{p}), \qquad q = eZ \tag{6.84}$$

where q and Z are the charge of each particle in absolute units and in the units of the elementary charge, respectively. The electrostatic force is obtained by differentiating this expression with respect to r:

$$F(r) \approx \frac{\exp(2\kappa r_\mathrm{p})}{1 + \kappa r_\mathrm{p}} \, \frac{q^2}{4\pi\epsilon_\mathrm{s} r^2} \exp(-\kappa r_\mathrm{p}) \tag{6.85}$$

6.10 Notes on Other Types of Force

Although the focus of this chapter is on electrostatic interactions, other types of force need to be mentioned for completeness. First, in particle dynamics dissipative and stochastic forces of Brownian motion play a major role; see H.C. Ottinger [Ott96], M. Fushiki [Fus92] and J. Dobnikar *et al.* [DHM+04].

Second, for small separations between the particles van der Waals forces may become important. These are attractive forces caused by dipole (or, more

generally, multipole) interactions between molecules. The dipole moments are inherently nonzero in polar molecules; in nonpolar ones, there are fluctuating dipole moments due to, in the classical picture, the orbital motion of electrons. The fluctuating moments of one molecule induce reaction moments in the neighboring ones, the end result being attractive forces (called *London*[24] *dispersion forces*). At very small separation, the electron clouds of two molecules overlap, which leads to a strong repulsion force that outweighs the attractive effects. In practice, both attraction and repulsion are frequently approximated by the Lennard–Jones potential

$$V_{\mathrm{LJ}}(r) \;=\; C\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$$

σ being a parameter. Theory of molecular forces is presented in a very lucid way by J. Israelachvili [Isr92]. Dispersion forces are studied very thoroughly by J. Mahanty & B.W. Ninham [MN76].[25] V.A. Parsegian has written a comprehensive monograph on van der Waals forces [Par06].

From the computational perspective, van der Waals forces between particles are inexpensive to evaluate if some analytical approximation (such as e.g. the Lennard–Jones potential) is adopted. The reason for this computational efficiency is that these forces are short-range and effectively involve no more than just a few neighbors of each particle.

Although simplified analytical approximations are adequate for many practical purposes, a more precise and rigorous computation of van der Waals forces between particles of different shapes and with different material parameters is a very interesting and challenging computational problem in its own right. The fundamental theory of finding dispersion forces from first principles of classical electrodynamics (with elements of a quantum-mechanical treatment) was laid out in the early 1950s by S.M. Rytov[26] [Ryt53]. In his seminal paper of 1955, E.M. Lifshitz[27] successfully carried out the Rytov theory calculation for the force between two semi-infinite slabs [Lif56]. Later, M.L. Levin & S.M. Rytov, in their book [LR67] that received less attention than I believe it deserves (there is apparently no English translation), streamlined the Rytov–Lifshitz calculation by taking advantage of the reciprocity principle. A brief account of these developments follows.

In the Rytov theory, phenomenological stochastic terms are introduced in the right hand side of Maxwell's equations, to reflect the fluctuating currents

[24] Fritz Wolfgang London (1900–1954), a German–American physicist.

[25] For the related subject of *Casimir forces*, see the review papers by M. Bordag *et al.* [BMM01], S.K. Lamoreaux [Lam99], and the monograph by P.W. Milonni [Mil94].

[26] Sergei Mikhailovich Rytov (1908–1997) – an outstanding physicist and radiophysicist.

[27] Evgenii Mikhailovich Lifshitz (1915–1985) – one of the most versatile physicists of all time and a co-author (with Lev Landau) of the famous comprehensive *Course of Theoretical Physics*.

in the media. As an exception to the use of the SI system of units throughout this book, the Rytov formulas are written here in the Gaussian system for consistency with the original work of Rytov, Lifshitz and others.

The Maxwell equations with stochastic sources are considered in the frequency domain. (The use of Fourier transforms for random processes is nontrivial and is justified heuristically by S.M. Rytov [Ryt53] and in a mathematically rigorous way by A.M. Iaglom [Iag62].)

$$\nabla \times \mathbf{E} \;=\; -\,\mathrm{i}\,\frac{\omega}{c}\,\mathbf{H} \tag{6.86}$$

$$\nabla \times \mathbf{H} \;=\; \mathrm{i}\,\frac{\omega}{c}\,(\epsilon\mathbf{E} + (\epsilon - 1)\mathbf{K}) \tag{6.87}$$

The stochastic current $\mathbf{K}$ is characterized by the correlation function that, according to Rytov's analysis, can be written as[28]

$$\mathcal{K}_{\alpha\beta}(r',\, r'') \;\equiv\; \langle K_\alpha(r')\, K^*_\beta(r'')\rangle \;=\; C\,\delta_{\alpha\beta}\,\delta(r' - r''), \quad \alpha, \beta = x, y, z \tag{6.88}$$

where r', r'' are two points in space, the angle brackets denote the ensemble average, $\delta_{\alpha\beta}$ is the Kronecker delta, and C is a constant equal to

$$C \;=\; \frac{4\hbar}{\operatorname{Im}\epsilon}\left(\frac{1}{2} + \frac{1}{\exp\frac{\hbar\omega}{T} - 1}\right) \tag{6.89}$$

($\hbar$ is the modified Planck constant and T is the temperature). The Kronecker delta indicates that different Cartesian components of the stochastic sources are uncorrelated. The "1/2" term in the big brackets corresponds to zero-point energy, i.e. the lowest energy level of the sources allowed by the uncertainty principle of quantum mechanics.

For two bodies in proximity to one another, the electromagnetic field produced by the fluctuating sources $\mathbf{K}$ in one of them leads to a force exerted on the other body. The van der Waals force in the Rytov–Lifshitz theory is nothing other than the statistical average of this electromagnetic force; it can be computed using the Maxwell Stress Tensor (MST, also statistically averaged).

The MST contains products of electric and magnetic field components; their correlation functions are

$$\langle E_\alpha(r_1) E_\beta(r_2)\rangle = \left\langle \left(\int K_\gamma(r')\, g_{\gamma\alpha}(r', r_1)\, dr'\right)\left(\int K_\gamma(r")\, g_{\gamma\beta}(r", r_2)\, dr"\right)\right\rangle \tag{6.90}$$

where $g_{\gamma\alpha}$ are components of the dyadic Green's function $\overleftrightarrow{g}$. Integration is carried out over the stochastic sources in the first body. Expressions for the magnetic field are completely similar.

[28] Lifshitz uses a slightly different definition of $\mathbf{K}$ and, accordingly, a somewhat different correlation function.

Converting the product of integrals into a double integral, noting that the only quantities subject to statistical averaging are the stochastic sources $\mathbf{K}$, and applying the correlation function (6.88), one obtains

$$\langle E_\alpha(r_1)E_\beta(r_2)\rangle \;=\; C\sum_{\gamma=1}^{3}\int_{\text{body }\#1} K_\gamma(r')\,g_{\gamma\alpha}(r',r_1)\,g_{\gamma\beta}(r',r_2)\,dr' \qquad (6.91)$$

The quantity $\langle E_\alpha(r)E_\beta(r)\rangle$ that enters the MST (and similar averages for the magnetic field) can be obtained from the above expression simply by setting $r_1 = r_2 = r$. A straightforward numerical implementation of this approach would involve setting up a set of integration knots in body #1 and performing the respective set of field computations with point sources at each of the knots to find Green's functions.

The procedure can be made much more efficient by taking advantage of the reciprocity – Hermitian symmetry of Green's dyadic. Its entries in (6.91) can be found by placing an elementary source at point r and a receiver at point r'; alternatively, the source and the receiver can be swapped. Although the results are equivalent theoretically, there is a great difference computationally.[29] The reason is that "sources" and "receivers" do not appear on an equal footing in computation: finding the field for one source yields its values everywhere, i.e. for all possible "receivers".

Instead of computing the field at some point r due to *distributed sources*, it is easier to compute the field distribution due to a single point-like source at r. For any given point r, this replaces a large set of field computations for sources at variable locations r' with just one field computation for the source at r.

If the MST is used, then the above procedure can be embedded into surface integration over points r on a surface enclosing body #2. An outline of the numerical algorithm for computing dispersion forces is hence as follows:

1. For two given bodies, choose an MST integration surface enclosing body #2 and a set of knots for a numerical quadrature over this surface.
2. Compute Green's functions for electric and magnetic fields at each integration knot on the surface. (This requires solution of six separate field problems for oscillating electric/magnetic dipole sources in three different directions.)
3. Compute the correlation $\langle E_\alpha(r)E_\beta(r)\rangle$ by integrating the product of two fields over body #1. Compute a similar correlation for the magnetic field.
4. Carry out numerical integration over the MST surface to obtain the contribution to the dispersion force at frequency ω.
5. Carry out numerical integration over all frequencies to find the total dispersion force acting on body #2.

[29] *Computation* here is understood in a broad sense and includes both analytical and numerical methods. (Levin & Rytov had only analytical computation in mind.)

To my knowledge, this proposal has not yet been implemented numerically, and clearly there are very serious computational challenges. The procedure relies on extremely accurate computation of the electromagnetic field due to elementary dipole sources on the MST surface, so that the integration of the MST can also be done accurately. Further, precise numerical integration with respect to frequency, especially in the vicinity of absorption lines of the media, is required, and the (phenomenological) complex dielectric permittivity has to be accurately represented over a wide frequency range. If these obstacles are overcome, the "numerical Rytov–Lifshitz" algorithm could lead to interesting results for forces between particles and molecular structures of different materials and shapes.

6.11 Thermodynamic Potential, Free Energy and Forces

Very helpful comments and discussion with Alain Bossavit and Markus Deserno on the material of this section are gratefully acknowledged.

Once the electrostatic potential has been found, derivative quantities, most notably forces on colloidal particles, need to be determined. This matter is taken up in the present section.

To the electrostatic equation

$$-\nabla \cdot \epsilon \nabla u \;=\; \rho \qquad \text{in } \Omega \tag{6.92}$$

there corresponds the Lagrangian

$$\mathcal{L}\{u\} = \rho u - \frac{1}{2}\,(\epsilon \nabla u) \cdot \nabla u \tag{6.93}$$

such that the *action*

$$G(u,\rho) \;=\; \int_\Omega \mathcal{L}\{u\}\, d\Omega \;\equiv\; \int_\Omega \left[\rho u - \frac{1}{2}(\epsilon \nabla u) \cdot \nabla u\right] d\Omega \tag{6.94}$$

has a stationary point at the solution u^* of the electrostatic equation (6.92). This can be verified by computing the variation of G with respect to potential u. Indeed, $G(u + \delta u, \rho)$, where u is *not* required to be the solution of the electrostatic equation, is

$$G(u+\delta u,\rho) \;=\; G(u) \;+\; \int_\Omega \left[\rho \delta u - (\epsilon \nabla u) \cdot \nabla \delta u \;+\; \frac{1}{2}(\epsilon \nabla \delta u) \cdot \nabla \delta u\right] d\Omega$$

Integrating the second term by parts and noting that the component linear in δu must vanish at the stationary point, one indeed obtains the electrostatic equation.

The stationary point of the action is in fact a maximum, which can be rigorously shown by computing the second variation of G but is also suggested by the fact that the stationary point is unique and $G(u^*, \rho) > G(0, \rho) = 0$ (see equation (6.95) below).

Remark 20. G can be viewed as a mathematical function of different sets of independent variables. In (6.94), the variables are u and ρ; however, when $u = u^* \equiv u^*(\rho)$, $G(\rho) \equiv G(u^*(\rho), \rho)$ can be considered as a function of ρ only. Furthermore, in the computation of forces via virtual work, we shall need to introduce the displacement of the body on which the electrostatic force is acting; then, clearly, G also depends on that displacement. Mathematically, these cases correspond to different functions, defined in different mathematical domains. Nevertheless for simplicity, but with some abuse of notation, the same symbol G will be used for all such functions, the distinguishing feature being the set of arguments.[30]

It is well known that for $u = u^*$ the expression for $G(u, \rho)$ simplifies because

$$\int_\Omega \rho u^* \, d\Omega \;=\; \int_\Omega (\epsilon \nabla u^*) \cdot \nabla u^* \, d\Omega$$

(To prove this, integrate the right hand side by parts and take into account the electrostatic equation for u^* and the boundary conditions.)

Hence, for $u = u^*$, action is in fact equal to the energy of the electrostatic field:

$$G(\rho) \;\equiv\; G(u^*(\rho),\, \rho) \;=\; \frac{1}{2} \int_\Omega (\epsilon \nabla u^*) \cdot \nabla u^* \, d\Omega \tag{6.95}$$

Remark 21. It is for this reason that G is often considered in the physical literature to be the *free energy* functional for the field; see K.A. Sharp & B. Honig [SH90], E.S. Reiner & C.J. Radke [RR90], M.K. Gilson *et al.* [GDLM93]. (In these papers, an additional term corresponding to microions in the electrolyte is included, as explained below.) However, the unqualified identification of G with energy is misleading, for the following reasons. First, G is mathematically defined for *arbitrary* u but its physical meaning for potentials not satisfying the electrostatic equation is unclear. (What is the physical meaning of a quantity that cannot physically exist?) Second, as already noted, G is *maximized*, not minimized, by $u = u^*$, which is rather strange if G is free energy.[31] F. Fogolari & J.M. Briggs [FB97] make very similar observations.

"Action" is a term from theoretical mechanics; in thermodynamics, G is commonly referred to as *thermodynamic potential*. An accurate physical interpretation and treatment of potential G is essential for computing electrostatic forces via virtual work, as forces are directly related to free energy rather than to the more abstract Lagrangian. More precisely, if a (possibly charged) body,

[30] In modern programming languages, such overloading of "functions" or "methods" is the norm.

[31] One could reverse the sign of $\mathcal{F}$, in which case the stationary point would be a *minimum*; however, this functional would no longer have the meaning of field energy, as its value at the exact solution u would be negative. See the same comment in footnote 10 on p. 83.

such as a colloidal particle, is subject to a ("virtual") displacement $\boldsymbol{d\xi}$, the electrostatic force $\boldsymbol{F}$ acting on the body satisfies

$$\boldsymbol{F} \cdot \boldsymbol{d\xi} \;=\; -\,dG^*(\boldsymbol{\xi}) \;\equiv\; -\,dG(\boldsymbol{\xi}, u^*(\rho(\boldsymbol{\xi})), \rho(\boldsymbol{\xi})) \tag{6.96}$$

where G^*, the thermodynamic potential evaluated at $u = u^*$, is the energy of the field according to (6.95). The definition of G is overloaded (see Remark 20): it now includes an additional parameter $\boldsymbol{\xi}$, the displacement of the body.[32] Importantly, the notation for G in (6.96) makes it explicit that solution u^* is a function of charge density ρ, which in turn depends on the position of the body. Then

$$dG(\boldsymbol{\xi}, u^*(\rho(\boldsymbol{\xi})), \rho(\boldsymbol{\xi})) \;=\; \left(\frac{\partial G}{\partial \boldsymbol{\xi}} \;+\; \frac{\partial G(u^*)}{\partial u} \frac{\partial u^*}{\partial \rho} \frac{\partial \rho}{\partial \boldsymbol{\xi}} \;+\; \frac{\partial G}{\partial \rho} \frac{\partial \rho}{\partial \boldsymbol{\xi}} \right) \cdot \boldsymbol{d\xi}$$

where

$$\frac{\partial u}{\partial \boldsymbol{\xi}} \;\equiv \left(\frac{\partial u}{\partial \xi_x}, \frac{\partial u}{\partial \xi_y}, \frac{\partial u}{\partial \xi_z} \right)$$

Since u^* is a stationary point of the thermodynamic potential, $\partial G(u^*)/\partial u = 0$, the second term in the right hand side vanishes and the differential becomes

$$dG(\boldsymbol{\xi}, u^*(\rho(\boldsymbol{\xi})), \rho(\boldsymbol{\xi})) \;=\; \left(\frac{\partial G}{\partial \boldsymbol{\xi}} \;+\; \frac{\partial G}{\partial \rho} \frac{\partial \rho}{\partial \boldsymbol{\xi}} \right) \cdot \boldsymbol{d\xi} \tag{6.97}$$

Let us now consider an alternative interpretation of G, where u is *not*, from the outset, constrained to be u^*. In this case,

$$dG(\boldsymbol{\xi}, u, \rho(\boldsymbol{\xi})) \;=\; \left(\frac{\partial G}{\partial \boldsymbol{\xi}} \;+\; \frac{\partial G}{\partial \rho} \frac{\partial \rho}{\partial \boldsymbol{\xi}} \right) \cdot \boldsymbol{d\xi} \tag{6.98}$$

In this interpretation, u is an *independent variable* in function G, and hence its partial derivative with respect to the displacement does not appear. Evaluation of *this version* of dG at $u = u^*$ thus yields the same result as in the previous case (6.97), where u was constrained to be u^* from the beginning.

In summary, one can compute the electrostatic energy first, by fixing $u = u^*$ in the thermodynamic potential or by any other standard means, and then apply the virtual work principle for forces. Alternatively, it is possible to apply virtual work directly to the thermodynamic potential (even though it is not energy for an arbitrary u) and *then* set $u = u^*$; the end result is the same.

Potential $G_{\mathrm{PB}}(u, \rho)$ for the PBE includes, in addition to (6.94), an entropic term related to the distribution of microions in the solvent. Theoretical analysis and derivation of G_{PB} goes back to the classical DLVO theory

[32] For a deformable structure, there exists a deeper and more general mathematical description of motion as a diffeomorphism $\xi_t : \Omega \to \Omega$, parameterized by time t; see e.g. A. Bossavit [Bos92]. For the purposes of this section, a simpler definition will suffice.

and the subsequent work of G.M. Bell & S. Levine [BL58] (1958). A systematic analysis is given by M. Deserno & C. Holm [DH01] and M. Deserno & H.-H. von Grünberg [DvG02] (2001–2002). In the context of macromolecular simulation, thermodynamic functionals, free energy, electrostatic and osmotic forces were studied by K.A. Sharp & B. Honig [SH90] (1990) and by M.K. Gilson *et al.* [GDLM93] (1993). These developments are considered in more detail below. A much more advanced treatment that goes beyond mean field theory, and beyond the scope of this book, is due to R.D. Coalson & A. Duncan [CD92], R.R. Netz & H. Orland [NO99, NO00], Y. Levin [Lev02b], A. Yu. Grosberg *et al.* [GNS02], T.T. Nguyen *et al.* [NGS00].

There are several equivalent representations of the thermodynamic potential. The following expression for the canonical ensemble (fixed total number of ions N, volume V and temperature T) is essentially the same as given by Deserno & von Grünberg [DvG02] and by Dobnikar *et al.* [DHM+04]:

$$G_{\mathrm{PB}}(u,\rho) \;=\; \int_{\mathbb{R}^3} \left(\frac{1}{2}\rho u \;+\; k_B T \sum_{\alpha} n_\alpha \log\left(n_\alpha \lambda_T^3\right) \right) dV \tag{6.99}$$

where n_α is the (position-dependent) volume concentration of species α of the microions and ρ is the total charge density equal to the sum of charge densities ρ^f of macroions ("fixed" ions) and ρ^m of microions (mobile ions). The normalization factor λ_T – the thermal de Broglie wavelength – renders the argument of the logarithmic function dimensionless and makes the classical and quantum mechanical expressions compatible:

$$\lambda_T \;=\; \sqrt{\frac{2\pi}{m k_B T}}\;\hbar$$

For the canonical ensemble, this factor adds a non-essential constant to the entropy.

If $u = u^*$ is the solution of the Poisson equation[33] with charge density ρ, then $G_{\mathrm{PB}}(u^*,\rho)$ is equal to the Helmholtz free energy of the system. Indeed, the right hand side of (6.99) has in this case a natural interpretation as electrostatic energy minus temperature times the entropy of the microions. Details are given in Appendix 6.14.

Solution u_{PB} of the Poisson–Boltzmann equation is in fact a stationary point of G_{PB}, under two constraints: (i) u is the electrostatic potential corresponding to ρ (that is, u satisfies the Poisson equation with ρ as a source), and (ii) electroneutrality of the solvent. This is verified in Appendix 6.14 by computing the variation of G_{PB} with respect to u.

The osmotic pressure force is given by the following expression [GDLM93, DHM+04] (Appendix 6.14)

[33] Equivalent to the solution of the Poisson–Boltzmann equation if, and only if, ρ^m obeys the Boltzmann distribution.

$$\mathbf{F}_{\text{osm}} = -k_B T \oint_S \sum_\alpha n_\alpha \, \mathbf{dS} \tag{6.100}$$

This is not surprising: since correlations are ignored, the microions behave as an ideal gas with pressure $n_\alpha k_B T$ for each species. Naturally, gas pressure depends on the density, and a nonuniform distribution of the microions around a colloidal particle in general produces a net force on it. In the numerical implementation, surface integral (6.100) is a simple amendment to the Maxwell Stress Tensor integral over a surface enclosing the particle under consideration (M. Fushiki [Fus92], J. Dobnikar [DHM+04]).

6.12 Comparison of FLAME and DLVO Results

In this numerical example of two charged colloidal particles in a solvent, the following parameters are used: particle radius normalized to unity; the solvent and solute dielectric constants are 80 and 2, respectively; the size of the computational domain is $10 \times 10 \times 10$; charges of the two particles are equal and normalized to unity. The linearized PBE, with the Debye length of 0.5, is applied in the solvent.

For comparison and verification, the problem is solved both with FEM and FLAME. In addition, an approximate analytical solution is available as a superposition of two Yukawa potentials.[34]

Finite Element simulations were run using FEMLAB™ (COMSOL Multiphysics), a commercial finite element package.[35] Two FE meshes with second-order tetrahedra are generated: a coarser one with 4,993 nodes, 25,195 elements, 36,696 degrees of freedom, and a finer one (Fig. 6.19) with 18,996 nodes, 97,333 elements, 138,053 degrees of freedom.

Two FLAME grids are used: $32 \times 32 \times 32$ and $64 \times 64 \times 64$. The FLAME scheme is applied on 7-point stencils in the vicinity of each particle – more precisely, if the midpoint of the stencil is within the distance $r_\text{p} + h$ from the center of the particle with radius r_p (as usual, h is the mesh size). Otherwise the standard 7-point scheme is used.

Fig. 6.20 shows the potential distribution along the line connecting the centers of the two particles. The FEM and FLAME results, as well as the approximate analytical solution, are all in good agreement.

As in the 2D case of Section 6.2.1, electrostatic forces can be computed via the Maxwell Stress Tensor (MST). The 3D analysis in this section also includes *osmotic pressure* forces due to the "gas" of microions.

The electrostatic energy for linear dielectric materials is (J.D. Jackson [Jac99], W.K.H. Panofsky & M. Phillips [PP62])

[34] The Yukawa potential is the exact solution for a single particle in a homogeneous solvent, not perturbed by the presence of any other particles.

[35] http://www.comsol.com

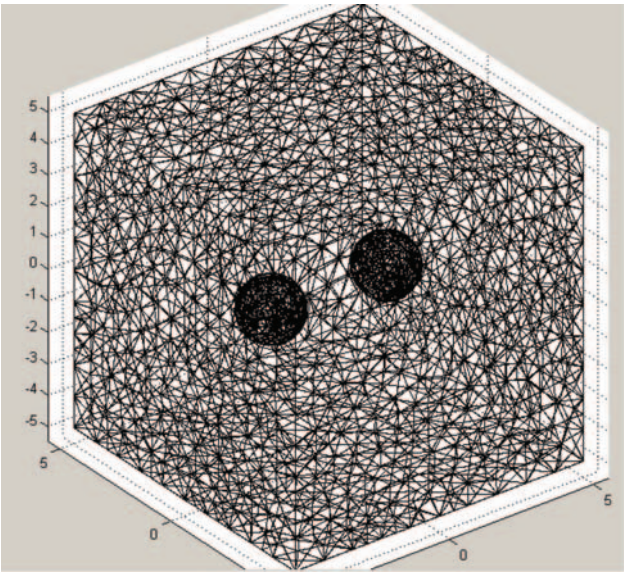


Fig. 6.19. A sample FE mesh for two particles.

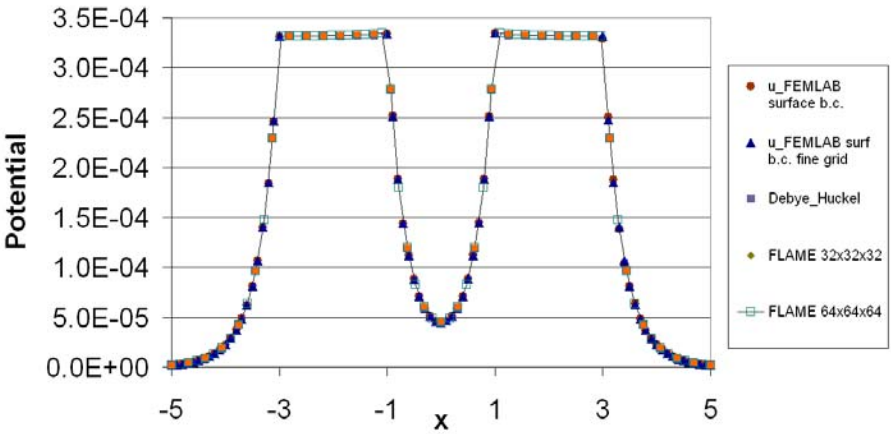


Fig. 6.20. Electrostatic potential along the line going through the centers of two particles. FLAME and FEM results are almost indistinguishable.

$$W^{\mathrm{el}} = \frac{1}{2}\int_{\mathbb{R}^3} \mathbf{E}\cdot\mathbf{D}\,dV \tag{6.101}$$

where, as usual, $\mathbf{E}$ and $\mathbf{D}$ are the electric field and displacement vectors, respectively. Noting that $\mathbf{E} = -\nabla u$, $\nabla\cdot\mathbf{D} = \rho$ (where ρ is the *total* electric charge density, including that of colloids and microions), and integrating by parts, one obtains another well known expression for the total energy:

$$W^{\mathrm{el}} = \frac{1}{2}\int_{\mathbb{R}^3} \rho u\,dV \tag{6.102}$$

The electrostatic part $\overleftrightarrow{T}^{\mathrm{el}}$ of the MST is defined as (see (6.12) on p. 289; J.D. Jackson [Jac99], J.A. Stratton [Str41] or W. K.H. Panofsky & M. Phillips [PP62])

$$\overleftrightarrow{T}^{\mathrm{el}} = \epsilon\begin{pmatrix} E_x^2 - \frac{1}{2}E^2 & E_xE_y & E_xE_z \\ E_yE_x & E_y^2 - \frac{1}{2}E^2 & E_yE_z \\ E_zE_x & E_zE_y & E_z^2 - \frac{1}{2}E^2 \end{pmatrix} \tag{6.103}$$

where ϵ is the dielectric constant of the medium in which the particles are immersed, E is the amplitude of the electric field and $E_{x,y,z}$ are its Cartesian components.

The electrostatic force acting on a particle is

$$\mathbf{F}_{\mathrm{el}} = \oint_S \overleftrightarrow{T}^{\mathrm{el}}\cdot\mathbf{dS} = \epsilon\oint_S \left[(\mathbf{E}\cdot\hat{n})\mathbf{E} - \frac{1}{2}E^2\hat{n}\right] dS \tag{6.104}$$

where S is any surface enclosing one, and only one, particle. Theoretically, the value of the force does not depend on the choice of the integration surface, but for the numerical results this is not exactly true.

In the FLAME experiments, the integration surface is usually chosen as spherical and is slightly larger than the particle. Adaptive numerical quadratures in the φ–θ plane are used for the integration. Obviously, the integration knots in general differ from the nodes of the FLAME grid, and therefore interpolation is needed. This involves a linear combination of the FLAME basis functions (six functions in the case of a 7-point scheme), plus the particular solution of the inhomogeneous equation in the vicinity of a charged particle. The interpolation procedure is completely analogous to the 2D one (Section 6.2.1).

It is interesting to compare FLAME results for the electrostatic force between two particles with the DLVO values from (6.85) on p. 324.[36] For this comparison, the main quantities are rendered dimensionless by scaling: $\tilde{r} = r/r_{\mathrm{p}}$, $\tilde{\mathbf{F}} = r_{\mathrm{p}}\mathbf{F}/kT$. FLAME is applied to the linearized PBE, with periodic boundary conditions. Typical surface plots of the potential distribution are shown in Fig. 6.21 and Fig. 6.22 for illustration.

FLAME vs. DLVO forces are plotted in Fig. 6.23 and Fig. 6.24. The first of these figures corresponds to the Debye length equal to the diameter of the

[36] FLAME simulations were performed by E. Ivanova and S. Voskoboynikov.

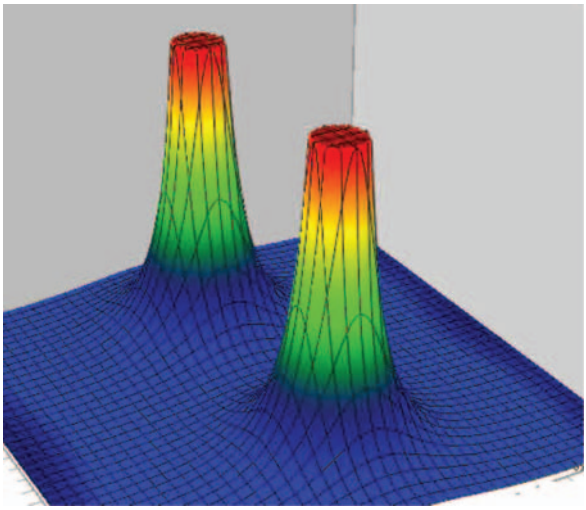


Fig. 6.21. An example of potential distribution (in arbitrary units) near two colloidal particles. The potential is plotted in the symmetry plane between the particles. (Simulation by E. Ivanova and S. Voskoboynikov.)

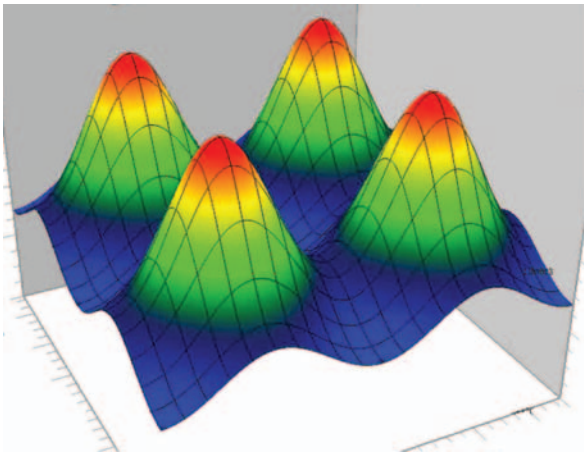


Fig. 6.22. An example of potential distribution (in arbitrary units) around eight colloidal particles. In the plane of the plot, only four of the particles produce a visible effect. (Simulation by E. Ivanova and S. Voskoboynikov.)

particle (or $\kappa r_{\mathrm{p}} = 0.5$). In the second figure, the Debye length is five times greater ($\kappa r_{\mathrm{p}} = 0.1$), so that the electrostatic interactions decay more slowly. Other parameters are listed in the figure captions.

Both the DLVO and FLAME results are approximations, and some discrepancy between them is to be expected. For small separations, the difference between the results can be attributed primarily to the approximations taken in the DLVO formula (6.85) for the ψ_0 and β parameters (p. 324). For intermediate distances between the particles, the agreement between DLVO and FLAME is excellent. For large separations comparable with the size of the computational box, FLAME suffers from the artifacts of periodic boundary conditions: the field and forces are affected by the periodic images of the particles.[37] For example, when the distance between a pair of particles A and B is half the size of the computational cell, the forces on A due to B and due to the periodic image of B on the opposite side of A cancel out. (More remote images have a similar but weaker effect, due to the Debye screening.) Obviously, this undesirable effect can be reduced by increasing the size of the box or by imposing approximate boundary conditions as a superposition of the Yukawa potentials.

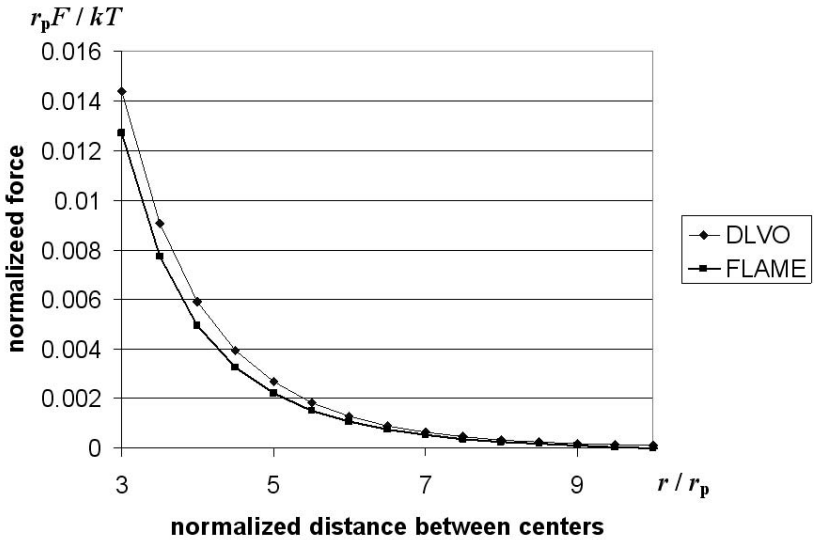


Fig. 6.23. Comparison of FLAME and DLVO forces between two particles. Parameters: $Z = 4$, $\frac{e^2}{\epsilon_{\mathrm{s}} k_B T}/r_{\mathrm{p}} = 0.012$, $\epsilon_{\mathrm{p}} = 1$, $\epsilon_{\mathrm{s}} = 80$, $\kappa r_{\mathrm{p}} = 0.5$, domain size 20. (Simulations by E. Ivanova and S. Voskoboynikov.)

[37] As we know from Chapter 5, a similar "periodic imaging" phenomenon is central in Ewald methods.

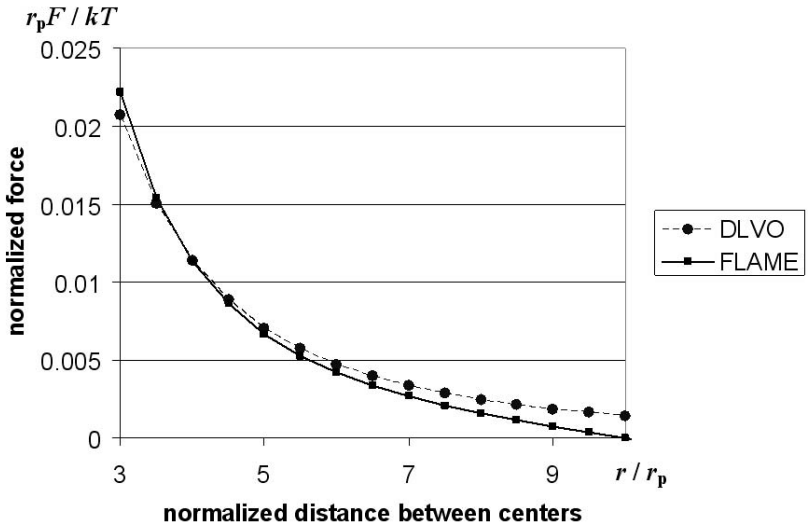


Fig. 6.24. Comparison of FLAME and DLVO forces between two particles. Parameters: same as in Fig. 6.23, except for $\kappa r_p = 0.1$. (Simulations by E. Ivanova and S. Voskoboynikov.)

6.13 Summary and Further Reading

Heterogeneous electrostatic models on the micro- and nanoscale, particularly in the presence of electrolytes, are of critical importance in a broad range of physical and biophysical applications: colloidal suspensions, polyelectrolytes, polymer- and biomolecules, etc. Due to the enormous complexity of these problems, any substantial improvement in the computational methodology is welcome.

Ewald methods that are commonly used in current computational practice (Chapter 5) work very well for homogeneous media. While in colloidal simulation the dielectric contrast between the solvent and solute can be neglected with an acceptable degree of accuracy, in macromolecular simulation this contrast cannot be ignored. From this perspective, the Flexible Local Approximation MEthods (FLAME) appear to be a step in the right direction. In FLAME, the numerical accuracy is improved – in many cases significantly – by incorporating accurate local approximations of the solution into the difference scheme.

The literature on colloidal, polyelectrolyte and molecular systems is vast. The following brief, and certainly incomplete, list includes only publications that are closely related to the material of this chapter: H.C. Ottinger [Ott96], M.O. Robbins *et al.* [RKG88], M. Fushiki [Fus92], J. Dobnikar *et al.* [DHM$^+$04], M. Deserno *et al.* [DHM00, DH01], B. Honig & A. Nicholls

[HN95], W. Rocchia *et al.* [RAH01], N.A. Baker *et al.* [BSS+01], T. Simonson [Sim03], D.A. Case *et al.* [CCD+05].

6.14 Appendix: Thermodynamic Potential for Electrostatics in Solvents

In this Appendix, thermodynamic potential (6.99) (p. 331, repeated here for convenience)

$$G_{\mathrm{PB}}(u,\rho) \;=\; \int_{\mathbb{R}^3} \left(\frac{1}{2}\,\rho u \;+\; k_B T \sum_\alpha n_\alpha (\log(n_\alpha \lambda_T^3) - 1) \right) dV \tag{6.105}$$

is considered in more detail. The total charge density $\rho = \rho^f + \rho^m$ is the sum of charge densities of macro- and microions, and λ_T is the thermal de Broglie wavelength

$$\lambda_T \;=\; \frac{h}{\sqrt{2\pi m k_B T}} \tag{6.106}$$

Although the integral in (6.105) is formally written over the whole space, in reality the integration can of course be limited just to the finite volume of the solvent. Alternative forms of the thermodynamic functional (M. Deserno & C. Holm [DH01], M. Deserno & H.-H. von Grünberg [DvG02], K.A. Sharp & B. Honig [SH90], M.K. Gilson *et al.* [GDLM93], J. Dobnikar *et al.* [DHM+04]) are considered later in this Appendix.

If $u = u^*$ is the solution of the Poisson equation with the total charge density ρ as the source, then the first term $\int_{\mathbb{R}^3} \frac{1}{2}\,\rho u^*\,dV$ is, as is well known from electromagnetic theory, equal to the energy of the electrostatic field. Free energy – the amount of energy available for reversible work – is different from the electrostatic energy due to heat transfer between the microions and the "heat bath" of the solvent. The Helmholtz free energy is

$$F \;=\; \langle E \rangle \;-\; TS$$

where the angle brackets indicate statistical averaging. This coincides with expression (6.105) for $G_{\mathrm{PB}}(u^*,\rho)$ because the entropy of the "gas" of microions is

$$S \;=\; k_B \int_{\mathbb{R}^3} \sum_\alpha n_\alpha (\log(n_\alpha \lambda_T^3) - 1)\,dV$$

Let us now show that the solution u_{PB} of the Poisson–Boltzmann equation is a stationary point of the thermodynamic potential $G_{\mathrm{PB}}(u^*,\rho)$, subject to two constraints. The first one is electroneutrality:

$$\int_{\mathbb{R}^3} \left(\sum_\alpha q_\alpha n_\alpha - \rho^f \right) dV \;=\; 0 \tag{6.107}$$

The second constraint (or more precisely, a set of constraints – one for each species of the microions) in the canonical ensemble is a fixed total number N_α of ions of species α:

$$\int_{\mathbb{R}^3} n_\alpha \, dV \;=\; N_\alpha \tag{6.108}$$

To handle the constraints, terms with a set of Lagrange multipliers λ and λ_α are included in the functional:

$$G_{\mathrm{PB}}(u^*, \rho, \lambda, \lambda_\alpha) \;=\; \int_{\mathbb{R}^3} \left(\frac{1}{2}\, \rho u^* \;-\; \lambda \left(\sum_\alpha q_\alpha \, n_\alpha \;-\; \rho^f \right) \right.$$

$$\left. + \, k_B T \sum_\alpha n_\alpha (\log(n_\alpha \lambda_T^3) - 1) \right) dV - \sum_\alpha \lambda_\alpha \left(\int_{\mathbb{R}^3} n_\alpha \, dV \;-\; N_\alpha \right) \tag{6.109}$$

Note that the functional is evaluated at $u = u^*$, the solution of the Poisson equation; clearly, u^* is the only electrostatic potential that can physically exist for a given charge density ρ.

The stationary point of this functional is found by computing the variation δG_{PB}. The integration-by-parts identity

$$\int_{\mathbb{R}^3} \delta\rho \, u^* \, dV \;=\; \int_{\mathbb{R}^3} \rho \, \delta u^* \, dV$$

helps to simplify the electrostatic part of δG_{PB}:

$$\delta G_{\mathrm{PB}}(u^*, \rho, \lambda, \lambda_\alpha) \;=\; \int_{\mathbb{R}^3} \left[\sum_\alpha u^* q_\alpha \, \delta n_\alpha \;-\; \lambda \sum_\alpha q_\alpha \, \delta n_\alpha - \sum_\alpha \lambda_\alpha \delta n_\alpha \right.$$

$$\left. +\; k_B T \sum_\alpha \log(n_\alpha \lambda_T^3) \, \delta n_\alpha \right] dV$$

(The obvious relationship $\rho_\alpha = q_\alpha n_\alpha$ between charge density and concentration has been taken into account.) Since the variations δn_α are arbitrary, the following conditions emerge:

$$u^* q_\alpha \;+\; k_B T (\log(n_\alpha \lambda_T^3) + 1) \;-\; \lambda q_\alpha \;-\; \lambda_\alpha \;=\; 0$$

This immediately yields the Boltzmann distribution for the ion density:

$$n_\alpha \;=\; n_{\alpha 0} \, \exp\left(-\frac{q_\alpha u^*}{k_B T} \right) \tag{6.110}$$

Thus the Poisson–Boltzmann distribution of the microions is indeed the stationary point of the thermodynamic potential, under the constraints of electroneutrality and a fixed number of ions.

It was already argued, on physical grounds, that the thermodynamic functional (6.105), evaluated at $u = u_{\mathrm{PB}}$ – the solution of the Poisson–Boltzmann

equation – yields the free energy of the colloidal system. Since this result is fundamental and has important implications (in particular, for the computation of forces as derivatives of free energy with respect to [virtual] displacement), it is desirable to derive it in a systematic and rigorous way. The classical work on this subject goes back to the 1940s and 1950s (E.J.W. Verwey & J. Th. G. Overbeek [VO48], G.M. Bell & S. Levine [BL58]). Here I review more recent contributions that are most relevant to the material of the present chapter: K.A. Sharp & B. Honig [SH90], E.S. Reiner & C.J. Radke [RR90], M.K. Gilson *et al.* [GDLM93], and M. Deserno & C. Holm [DH01].

Sharp & B. Honig [SH90] note that a thermodynamic potential similar to G_{PB} above is minimized by the solution of the Poisson–Boltzmann equation. Therefore, they argue, this potential represents the free energy of the system. While the conclusion itself is correct, the argument leading to it lacks rigor. First, it is not difficult to verify that the functional is actually *maximized*, not minimized, by the PB solution. More importantly, there are infinitely many different functionals that are stationary at u_{PB}. This was already noted in Remark 21 on p. 329.

Reiner & Radke [RR90] address this latter point by postulating that free energy must be a function $\mathcal{F}$ of the action functional and that $\mathcal{F}$ must have additive properties with respect to the volume and surfaces of the system. They then proceed to show that $\mathcal{F}$ may alter G_{PB} only by an unimportant additive term and a scaling factor – in other words, G_{PB} is essentially a unique representation of free energy. However, the initial postulate is not justified: the fact that two functionals share the same stationary point does not imply that one of them can be expressed as a function of the other. For example, all functionals of the form

$$U_m \;=\; \int_{\mathbb{R}^3} |u|^m \, dV, \qquad m = 1, 2, \ldots$$

have the same obvious minimization point $u = 0$. Yet it is impossible to express, say, U_{100} as a function of just U_1 – much more information about the underlying function u is needed.[38]

Deserno & Holm's derivation [DH01] is based on the principles of statistical mechanics and combines rigor with relative simplicity. Their analysis starts with the system Hamiltonian for N microions (only one species for brevity) treated as point charges:

$$H(\mathbf{r}, \mathbf{p}) \;=\; \sum_{i=1}^{N} \frac{p_i^2}{2m} \;+\; \sum_{1 \le i < j \le N} \frac{q^2}{4\pi\epsilon \, |\mathbf{r}_i - \mathbf{r}_j|} \;+\; \int_{\mathbb{R}^3} \sum_{i=1}^{N} \frac{q \rho^f(\mathbf{r})}{4\pi\epsilon \, |\mathbf{r}_i - \mathbf{r}|} \, dV \tag{6.111}$$

[38] In case the reader is unconvinced, here is a simple 1D illustration. Let a family of rectangular pulses u_ϵ be defined as equal to ϵ^{-1} on $[0, \epsilon]$ ($\epsilon > 0$) and zero otherwise. These pulses have the same U_1 but very different U_{100}. It is therefore impossible to determine U_{100} based on U_1 alone.

where q and m are the charge and mass of each microion; $\mathbf{r}_i$ and $\mathbf{p}_i$ are the position and momentum vectors of the i-th microion. Mutual interactions of fixed charges are not included in the Hamiltonian, as that would only add an inessential constant.

The Hamiltonian can be rewritten using potentials u^m and u^f of the microions and fixed ions, respectively:

$$H(\mathbf{r},\mathbf{p}) \;=\; \sum_{i=1}^{N} \frac{p_i^2}{2m} \;+\; q \sum_{i=1}^{N} \left(\frac{1}{2}\, u^m(\mathbf{r}_i;\mathbf{r}) \;+\; u^f(\mathbf{r}_i) \right) \tag{6.112}$$

Remark 22. In this last form, the Hamiltonian includes self-energies of the microions, and so the expression should strictly speaking be adjusted (as done in Chapter 5) to eliminate the singularities. However, anticipating that the micro-charges will eventually be smeared and treated as a continuum, we turn a blind eye to this complication and opt for simpler notation.

Remark 23. The microion potential $u^m(\mathbf{r}_i;\mathbf{r})$, is "measured" at point $\mathbf{r}_i$ but depends on the $3N$-vector $\mathbf{r}$ of coordinates of *all* charges. This coupling of all coordinates makes precise statistical analysis extremely difficult. In the mean field approximation, the situation is simplified dramatically by averaging out the contribution to $u^m(\mathbf{r}_i)$ of all charges other than i.

As is well known from thermodynamics, the partition function Z is obtained, in the classical limit, by integrating the exponentiated Hamiltonian:[39]

$$Z \;=\; \frac{1}{N!\, h^{3N}} \int \exp(-\beta H)\, d\mathbf{r}\, d\mathbf{p}, \qquad \beta \equiv \frac{1}{k_B T} \tag{6.113}$$

where the integral is over the whole $6N$-dimensional phase space. Z serves as a normalization factor for the probability density of finding the system near a given energy value H:

$$f(\mathbf{r}_1,\ldots,\mathbf{r}_N,\mathbf{p}_1,\ldots,\mathbf{p}_N) \;=\; Z^{-1} \exp(-\beta H) \tag{6.114}$$

The Helmholtz free energy is, as is also well known,

$$F \;=\; -\, k_B T \,\log Z \tag{6.115}$$

The momentum part of Z gets integrated out of (6.113) quite easily and yields

$$F_{\mathbf{p}} \;=\; k_B T \log(N!\, \lambda_T^{3N}) \;\approx\; N \left[\log(N \lambda_T^3) \;-\; 1) \right] \tag{6.116}$$

where the Stirling formula for the factorial has been used.

[39] *Partition function* is arguably a misnomer: it is in fact the result of *integration* or *summation*, which is the opposite of partitioning. "Sum over states" (a direct translation from the original German *Zustandssumme*) is a more appropriate but less frequently used term.

The position part of Z, unlike the momentum part, is impossible to evaluate exactly, due to the pairwise coupling of the coordinates of all microions via the $|\mathbf{r}_i - \mathbf{r}_j|$ terms in the Hamiltonian. The mean field approximation decouples these coordinates (see Remark 23), thereby splitting the system Hamiltonian into a sum of the individual Hamiltonians of all microions. Consequently, the joint probability density (6.114) becomes a product of the individual probability densities of the ions, implying that the correlations between the ions are neglected. The limitations of this assumption are summarized in Section 6.5 on p. 313.

Once the coordinates are (approximately) decoupled, the N-fold integration of $\exp(-\beta H)$ in Z (6.113) yields the following expression for thermodynamic potential (M. Deserno & C. Holm [DH01]):

$$\tilde{G}_{\mathrm{PB}} = \int_{\mathbb{R}^3} \left(qn(\mathbf{r}) \left(\frac{1}{2} u^m(\mathbf{r}) + u^f(\mathbf{r}) \right) + k_B T n(\mathbf{r}) \left(\log(n(\mathbf{r})\lambda_T^3) - 1 \right) \right) dV \tag{6.117}$$

where both the momentum part (6.116) and the mean-field coordinate part are included. In addition, the continuum limit has been taken, so that the microions are now represented by the equivalent volume density $n(\mathbf{r})$. The tilde sign in $\tilde{G}_{\mathrm{PB}}$ is used to recognize that the electrostatic energy part in this functional is different from a more natural expression

$$\int_{\mathbb{R}^3} \frac{1}{2} \rho u \, dV$$

appearing in (6.105). However, the difference is not essential. Indeed, splitting the total charge density ρ and the total electrostatic potential u up into the microion and fixed-charge parts, we get

$$\begin{aligned} \int_{\mathbb{R}^3} \frac{1}{2} \rho u \, dV &= \frac{1}{2} \int_{\mathbb{R}^3} (\rho^m u^m + \rho^m u^f + \rho^f u^m + \rho^f u^f) \, dV \\ &= \int_{\mathbb{R}^3} \left(\frac{1}{2} \rho^m u^m + \rho^m u^f + \frac{1}{2} \rho^f u^f \right) dV \end{aligned}$$

where the reciprocity principle (or, mathematically, integration by parts) was used to reveal two equal terms. The last term, involving only the fixed charges, is constant and can therefore safely be dropped from the potential. This immediately makes the expression equivalent to the electrostatic part of G_{PB} (6.105).

Alternative forms of the thermodynamic functional can be obtained under an additional constraint: potential u satisfies the electrostatic equation *for the Boltzmann distribution of the microions* (6.110). An equivalent expression for the Boltzmann distribution is

$$\log n_\alpha = -\frac{q_\alpha u}{k_B T} + \text{const}$$

Hence the entropic term in the functional – for the Boltzmann distribution of the ion density – can be rewritten as

$$\int_{\mathbb{R}^3} k_B T \sum_\alpha n_\alpha (\log(n_\alpha \lambda_T^3) - 1)\, dV \;=\; -\int_{\mathbb{R}^3} \sum_\alpha n_\alpha q_\alpha u\, dV \;+\; \text{const}$$

$$= \; -\int_{\mathbb{R}^3} \rho^m u\, dV \;+\; \text{const} \tag{6.118}$$

6.15 Appendix: Generalized Functions (Distributions)

The first part of this Appendix is an elementary introduction to generalized functions, or distributions. The second part outlines their applications to boundary value problems and to the treatment of interface boundary conditions.

The history of mathematics is full of examples where the existing notions and objects work well for a while but then turn out to be insufficient and need to be extended to make further progress. That is, for example, how one proceeds from natural numbers to integers and then to rational, real and complex numbers. In each case, there are desirable operations (such as e.g. division of integers) that cannot be performed within the existing class, which calls for an extension of this class.

A different example that involves an extension of the exponential function from numbers to matrices and operators is outlined in Appendix 2.10 on p. 65.

Why would functions in standard calculus need to be generalized? What features are they lacking? One notable problem is differentiation. As an example, the Heaviside unit step function[40] $H(x)$, equal to one for $x \geq 0$ and zero otherwise, in regular calculus does not not have a derivative at zero. In an attempt to generalize the notion of derivative and make it applicable to the step function, one may consider an approximation H_ϵ to $H(x)$ (Fig. 6.15).

The derivative of $H_\epsilon(x)$ is a rectangular pulse equal to $1/\epsilon$ for $|x| < \epsilon/2$ and zero for $|x| > \epsilon/2$. (In standard calculus, this derivative is undefined for $x = \pm\epsilon/2$.) As $\epsilon \to 0$, H_ϵ tends to the step function, but the limit of the derivative $H'_\epsilon(x)$ in the usual sense is not meaningful. Indeed, this pointwise limit $H'_{\epsilon\to 0}(x)$ is equal to infinity at $x = 0$ and zero everywhere else. In contrast with the usual integration/differentiation operations that are inverses of one another, in this irregular case the original unit step $H(x)$ cannot be recovered from $H'_{\epsilon\to 0}(x)$. Indeed, although the *existence* of the step can be inferred from $H'_{\epsilon\to 0}(x)$, the information about the *magnitude* of the step is lost.

[40] Oliver Heaviside (1850–1925) is a British physicist and mathematician, the inventor of operational calculus, whose work profoundly influenced electromagnetic theory and analysis of transmission lines. The modern vector form of Maxwell's equations was derived by Heaviside (Maxwell had 20 equations with 20 unknowns).

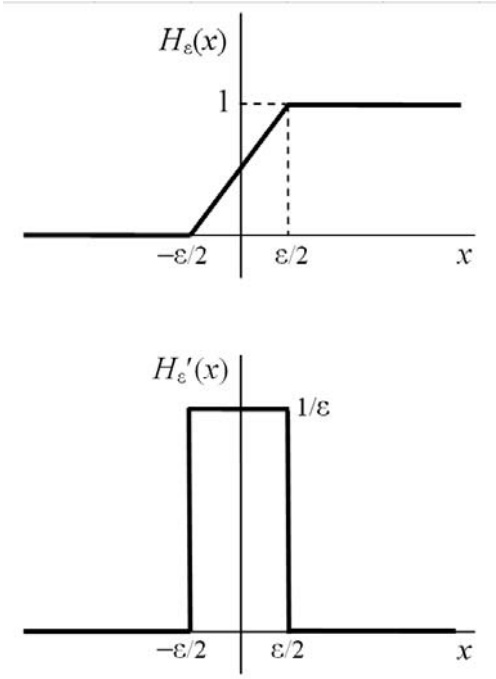


Fig. 6.25. A steep ramp (top) approximates the Heaviside step function. The derivative of this ramp function is a sharp pulse (bottom). However, as $\epsilon \to 0$, the pointwise limit of this derivative is not meaningful.

A critical observation in regard to the sequence of narrow and tall pulses with $\epsilon \to 0$ is that the precise pointwise values of these pulses are unimportant; what matters is the "action" of such pulses on some system to which they may be applied. A mathematically meaningful definition of this action is the integral

$$\int_R H_\epsilon'(x)\psi(x)\,dx \tag{6.119}$$

where $\psi(x)$ is any smooth function that can be viewed as a "test" function to which $H_\epsilon'(x)$ is applied.[41]

It is easy to see that for $\epsilon \to 0$ the integral in (6.119), unlike H_ϵ' itself, has a simple limit:

$$\int_R H_\epsilon'(x)\psi(x)\,dx = \int_{-\epsilon/2}^{\epsilon/2} \epsilon^{-1}\psi(x)\,dx \to \psi(0)$$

[41] For technical reasons, in the usual definition of generalized functions it is assumed that $\psi(x)$ is differentiable infinitely many times and has a compact support. For the mathematical details, see the monographs cited at the end of this Appendix.

Thus the "action" of $H'_\epsilon(x)$ on any smooth function $\psi(x)$ is just $\psi(0)$. The proper mathematical term for this action is a *linear functional*: it takes a smooth function ψ and maps it to a number, in this particular case to $\psi(0)$.

This insight ultimately leads to the far-reaching notion of *generalized functions*, or *distributions*: linear functionals defined on smooth "test" functions.

Example 18. The above functional that maps any smooth funciton ψ to its value at zero is the famous *Dirac delta*:

$$\langle \delta,\ \psi \rangle \ = \ \psi(0) \tag{6.120}$$

where the angle brackets denote a linear functional. For instance, $\langle \delta,\ \exp(x) \rangle = \exp(0) = 1$, $\langle \delta,\ x^2 + 3 \rangle = 3$, etc.[42]

There is an inconsistency between the proper mathematical treatment of the Dirac delta (and other distributions) as a linear functional and the popular informal notation $\delta(x)$ (implying that the Dirac delta is a function of x) and $\int \delta(x)\psi(x)dx$. The integral sign, strictly speaking, should be understood only as a shorthand notation for a linear functional.

Example 19. Any regular function $f(x)$ can be viewed also as a distribution by associating it with the linear functional

$$\langle f,\ \psi \rangle \ = \ \int_R f(x)\psi(x)\, dx \tag{6.121}$$

It can be shown that the distributions corresponding to different integrable functions are indeed different, and so this definition is a valid one. For example, the sinusoidal function $\sin x$ is associated with the generalized function $\int_R \sin x\, \psi(x)\, dx$.

Example 20. While any regular function can be identified with a distribution, the opposite is not true. The Dirac delta is one example of a generalized function that does not correspond to any regular one. Another such example is the *Cauchy principal value* distribution

$$\langle \text{p.v.} \left(\frac{1}{x} \right),\ \psi \rangle \ = \ \lim_{\epsilon \to 0+} \int_{|x|>\epsilon} \frac{\psi(x)}{x}\, dx \tag{6.122}$$

This distribution cannot be identified, in the sense of (6.121), just with the function $1/x$, as the integral

$$\int_R \frac{1}{x}\, \psi(x)\, dx$$

does not in general exist if $\psi(0) \neq 0$.

[42] Strictly speaking, since $\exp(x)$ and $x^2 + 3$ do not have a compact support, these expressions are not valid without additional elaboration.

Generalized functions have very vast applications to differential equations: suffice it to say that Green's functions are, by definition, solutions of the equation with the right hand side equal to the Dirac delta. The remainder of this Appendix covers the most essential features and notation relevant to the content of Chapter 6.

While functions in classic calculus are not always differentiable, *generalized* functions are. To see how the notion of derivative can be generalized, start with a differentiable (in the calculus sense) function $f(x)$ and consider the "action" of its derivative on any smooth test function $\psi(x)$:

$$\int_R f'(x)\psi(x)\,dx \;=\; -\int_R f(x)\psi'(x)\,dx \tag{6.123}$$

This is an integration-by-parts identity, where the term outside the integral vanishes because the test function ψ, by definition, has a compact support and therefore must vanish at $\pm\infty$. Since differentiation has been removed from f, the right hand side of (6.123) has a wider range of applicability and can now be taken as a definition of the generalized derivative of f even if f is not differentiable in the calculus sense. Namely, the generalized derivative of f is defined as the linear functional

$$\langle f',\ \psi\rangle \;=\; -\int_R f(x)\psi'(x)\,dx \tag{6.124}$$

Example 21. Applying this definition to the Heaviside step function H, we have

$$\langle H',\psi\rangle \;=\; -\int_R H(x)\psi'(x)\,dx \;=\; -\int_0^\infty \psi'(x)\,dx \;=\; \psi(0) \;=\; \langle\delta,\psi\rangle \tag{6.125}$$

In more compact notation, this is a well-known identity

$$H' \;=\; \delta$$

The derivative of the unit step function (in the sense of distributions) is the delta function.

Example 22. As a straightforward but practically very useful generalization of the previous example, consider a function $f(x)$ that is smooth everywhere except for a few discrete points x_i, $i = 1,\ldots,n$, where it may have jumps $[f]_i \equiv f(x_i+) - f(x_i-)$. Then the distributional derivative of f is

$$f' \;=\; \{f'\} \;+\; \sum_{i=1}^{n} [f]_i\,\delta(x-x_i) \tag{6.126}$$

where $\delta(x-x_i)$ is, by definition, the functional[43]

[43] There is an inconsistency between the popular notation $\delta(x-x_i)$, suggesting that δ is a function of x, and the mathematical meaning of δ as a linear functional. More proper notation would be $\delta(x_i,\psi)$.

$$\langle \delta(x - x_i),\ \psi \rangle \ = \ \psi(x_i)$$

In (6.126), the braces denote regular derivatives[44] viewed as generalized functions. The generalized derivative of f is thus equal to the regular one, plus a set of Dirac deltas corresponding to the jumps of f. The derivation of (6.126) is a straightforward extension of that of (6.125).

Example 23. For $f(x) = H(x)\cos x$, where $H(x)$ is the Heaviside step function, $f'(x) = \{f'(x)\} + \delta(x)$, with $\{f'(x)\} = -H(x)\sin x$.

Example 24. We now make the leap over to three dimensions. In 3D, distributions are also defined as linear functionals acting on smooth "test" functions with a compact support. For instance, the Dirac delta in 3D is

$$\langle \delta, \psi \rangle \ = \ \psi(0) \tag{6.127}$$

which is formally the same definition as in 1D, except that now ψ is a function of three coordinates and zero in the right hand side means the origin $x = y = z = 0$. Generalized partial derivatives are defined by analogy with the 1D case; for example,

$$\langle \frac{\partial f}{\partial x}, \psi \rangle \ = \ -\int_{\mathbb{R}^3} f(x) \frac{\partial \psi}{\partial x}\, dx \tag{6.128}$$

Of particular interest in Chapter 6 is generalized divergence. The divergence equation $\nabla \cdot \mathbf{D} = \rho$ is valid for *volume* charge density ρ; however, if divergence is understood in the sense of distributions, this equation becomes applicable to surface charges as well. If $\mathbf{D}$ is a *smooth* field, then for any "test" function ψ integration by parts yields[45]

$$\int_{\mathbb{R}^3} \psi \nabla \cdot \mathbf{D}\, dV \ = \ -\int_{\mathbb{R}^3} \mathbf{D} \cdot \nabla \psi\, dV \tag{6.129}$$

The extra term outside the integral vanishes because ψ has a compact support and is therefore zero at infinity. The above identity suggests, by analogy with generalized derivative, a definition of generalized divergence as a linear functional

$$\langle \nabla \cdot \mathbf{D}, \psi \rangle \ = \ -\int_{\mathbb{R}^3} \mathbf{D} \cdot \nabla \psi\, dV \tag{6.130}$$

Consider now the generalized derivative for the case where the normal component of $\mathbf{D}$ may have a jump across a surface S enclosing a domain Ω. (In electrostatic problems, Ω may be a body with a dielectric permittivity different from that of the outside medium, and S may carry a surface charge.) Then the generalized derivative is transformed, by splitting the integral into regions inside and outside Ω and again using integration by parts, to

[44] This is V.S. Vladimirov's notation [Vla84].
[45] Test functions are smooth by definition.

$$\langle \nabla\cdot\mathbf{D},\, \psi\rangle = -\int_{\mathbb{R}^3} \mathbf{D}\cdot\nabla\psi\, dV = \int_{\Omega} \psi\nabla\cdot\mathbf{D}\, dV + \int_{\mathbb{R}^3-\Omega} \nabla\cdot\mathbf{D}\, \psi\, dV + \int_S \psi[\mathbf{D}_n]\, dS \tag{6.131}$$

where $[\mathbf{D}_n]$ is the jump of the normal component of $[\mathbf{D}]$ across the surface:

$$[\mathbf{D}_n] \;=\; (\mathbf{D}_{\text{out}} - \mathbf{D}_{\text{in}})\cdot\mathbf{n}$$

and $\mathbf{n}$ is the outward normal to the surface of Ω. In more compact form, generalized divergence (6.131) can be written as

$$\nabla\cdot\mathbf{D} \;=\; \{\nabla\cdot\mathbf{D}\} \;+\; [\mathbf{D}_n]\,\delta_S \tag{6.132}$$

where the curly brackets again denote "calculus-style" divergence in the volume and δ_S is the surface-delta defined formally as the functional

$$\langle \delta_S,\, \psi\rangle \;=\; \int_S \psi\, dS$$

The physical meaning of expression (6.132) is transparent: generalized divergence is equal to regular divergence (that can be defined via the usual derivatives everywhere except for the surface), plus the surface-delta term corresponding to the jump. This result is analogous to the 1D expression for generalized derivative (6.126) in the presence of jumps.

The last example shows, as a consequence of (6.132), that Maxwell's divergence equation $\nabla\cdot\mathbf{D} = \rho$ is valid for both volume and surface charges (or any combination thereof) if divergence is understood in the generalized sense. This point of view is very convenient, as it allows one to treat interface boundary conditions as a natural part of the differential equations rather than as some extraneous constraints. In particular, zero generalized divergence of the $\mathbf{D}$ field in electrostatics implies zero volume charges *and* zero surface charges – the continuity of the normal component of $\mathbf{D}$ across the surface.

Further reading

The original book by L. Schwartz [Sch66] is a very good introduction to the theory of distributions, at the mathematical level accessible to engineers and physicists. V.S. Vladimirov's book [Vla84] focuses on applications of distributions in mathematical physics and is highly relevant to the content of this chapter. A simpler introduction, with the emphasis on electromagnetic problems, is given by D.G. Dudley [Dud94]. There is also a vast body of advanced mathematical literature on the theory of distributions, but that is well beyond the scope of this book.

7

Applications in Nano-Photonics

7.1 Introduction

Visible light is electromagnetic waves with submicron wavelengths – between ∼400 nm (blue light) and ∼700–750 nm (red light) in free space. Therefore propagation of light through materials is affected greatly by their submicron features and structures. Moreover, the ability to create and control such small features has led to amazing new physical effects, technologies and devices, as discussed later in this chapter.

Truly nanoscale features, *much* smaller than the wavelength, can also be crucial. In particular, one of the recent exciting directions in photonics involves nanoscale (5–50 nm) "plasmon" particles and structures that exhibit very peculiar resonance behavior in the optical frequency range (Section 7.11).

This chapter is not a comprehensive review of nano-photonics; rather, it covers selected intriguing applications and related methods of computer simulation. For a broader view, see P.N. Prasad's monographs [Pra03, Pra04]. References on more specific subjects (photonic crystals, plasmonics, nano-optics, etc.) are given in the respective sections of this chapter.

The indispensable starting point in a discussion of photonics is Maxwell's equations that describe electromagnetic fields in general and propagating electromagnetic waves in particular. After a brief review of Maxwell's equations, the chapter gives an introduction to band structure and the Photonic BandGap (PBG) phenomenon in photonic crystals, plasmonic particles and plasmon-enhanced Scanning Near-field Optical Microscopy (SNOM), backward waves, negative refraction and nanofocusing, with related simulation examples.

7.2 Maxwell's Equations

The system of Maxwell's equations contains the "curl part"

$$\nabla \times \mathbf{E} \;=\; -\,\partial_t \mathbf{B} \tag{7.1}$$

$$\nabla \times \mathbf{H} \;=\; \partial_t \mathbf{D} \;+\; \mathbf{J} \tag{7.2}$$

and the "divergence part"

$$\nabla \cdot \mathbf{D} \;=\; \rho \tag{7.3}$$

$$\nabla \cdot \mathbf{B} \;=\; 0 \tag{7.4}$$

In these equations, $\mathbf{E}$ and $\mathbf{H}$ are the electric and magnetic field, respectively; $\mathbf{D}$ and $\mathbf{B}$ are the electric and magnetic flux densities, respectively; ρ is the electric charge density, and $\mathbf{J}$ is the electric current density. For physical definitions of these vector quantities and a detailed physical discussion see well-known textbooks by L.D. Landau & E.M. Lifshitz [LL84], J.A. Stratton [Str41], R.P. Feynman *et al.* [FLS89], W.K.H. Panofsky & M. Phillips [PP62], R. Harrington [Har01].

The physical meaning of Maxwell's equations becomes more transparent if they are rewritten in integral form using the standard vector calculus identities. The first two equations become

$$\oint_{\partial S} \mathbf{E} \cdot \mathbf{dl} \;=\; -\,\frac{d}{dt} \int_S \mathbf{B} \cdot \mathbf{dS} \tag{7.5}$$

$$\oint_{\partial \Omega} \mathbf{H} \cdot \mathbf{dl} \;=\; \frac{d}{dt} \int_\Omega \mathbf{D} \cdot \mathbf{dS} \;+\; \int_\Omega \mathbf{J} \cdot \mathbf{dS} \tag{7.6}$$

These relationships are valid for any open surface S with its closed-contour boundary ∂S oriented in the standard way. Equation (7.5) – known as Faraday's Law – means that the electromotive force (emf) over a closed contour is induced by the changing magnetic flux passing through that contour. (The emf is defined as the line integral of the electric field.)[1] Unlike the emf equation (7.5), equation (7.6) for the magnetomotive force (mmf, the contour integral of the magnetic field) contains two terms in the right hand side. The mmf is due to the changing electric flux and to the electric current passing through the closed contour.

The lack of complete symmetry between the emf and mmf equations (7.5) and (7.6) is due to the apparent absence of magnetic charges (monopoles).[2]

[1] An alternative approach, where – loosely speaking – the emf is taken as a *primary* quantity and the field is defined via the emf, is arguably more fundamental but requires the notions of differential geometry and differential forms that are beyond the standard engineering curriculum. See monographs by P. Monk [Mon03] and A. Bossavit [Bos98] as well as the section on edge elements (Section 3.12, p. 139).

[2] On February 14, 1982 a monopole-related event may have been registered in the laboratory of Blas Cabrera (B. Cabrera, First results from a superconductive detector for moving magnetic monopoles, *Phys. Rev. Lett.*, vol. 48, pp. 1378–1381, 1982). An abrupt change in the magnetic flux through a superconducting loop was recorded (the magnetic flux is known to be quantized). A magnetic monopole passing through the loop would cause a similar flux jump. However, nobody has been able to reproduce this result.

If monopoles are ever discovered, presumably the Faraday Law will have to be amended, as magnetic currents would contribute to the emf over a closed contour.

Next, the integral form of the divergence equations (7.3) and (7.4) is, for any 3D domain Ω bounded by a closed surface $\partial\Omega$,

$$\oint_{\partial\Omega} \mathbf{D} \cdot \mathbf{dS} \;=\; Q, \qquad Q \;=\; \int_{\Omega} \rho \, d\Omega \tag{7.7}$$

$$\oint_{\partial\Omega} \mathbf{B} \cdot \mathbf{dS} \;=\; 0 \tag{7.8}$$

The first of these equations, known as Gauss's Law, relates the flux of the $\mathbf{D}$ vector through any closed surface to the total electric charge *inside* that surface. The second equation, for the flux of the $\mathbf{B}$ field, is analogous, except that there is no magnetic charge (see footnote 2).

As it stands, the system of four Maxwell's equations is still underdetermined. Generally speaking, a vector field in the whole space (and vanishing at infinity) is uniquely defined by *both* its curl and divergence, whereas Maxwell's equations specify the curl of $\mathbf{E}$ and the divergence of $\mathbf{D}$, not $\mathbf{E}$. The same is true for the pair of magnetic fields $\mathbf{H}$ and $\mathbf{B}$. To close the system of equations, one needs to specify the relationships, known as constitutive laws, between $\mathbf{E}$, $\mathbf{D}$, $\mathbf{H}$ and $\mathbf{B}$. In linear isotropic materials,

$$\mathbf{D} \;=\; \epsilon \mathbf{E}, \quad \epsilon \;=\; \epsilon(x, y, z) \tag{7.9}$$

$$\mathbf{B} \;=\; \mu \mathbf{H}, \quad \mu \;=\; \mu(x, y, z) \tag{7.10}$$

In other types of media, however, relationships between the fields can be substantially more complicated – they can be nonlinear and can include the time history of the electromagnetic process. The dependence on the history is called *hysteresis* (I.D. Mayergoyz [May03]). Moreover, the magnetic and electric fields can be coupled (e.g. symmetrized Condon or Drude–Born–Fedorov relations for chiral media; see J. Lekner [Lek96]). Our discussion and examples, however, will be limited to the linear isotropic case (7.9), (7.10).

There is a connection between the curl and divergence equations. Indeed, since divergence of curl is zero, by applying the divergence operator to both sides of the curl equations (7.1) and (7.2) one obtains

$$\partial_t \nabla \cdot \mathbf{B} \;=\; 0 \tag{7.11}$$

and

$$\nabla \cdot (\partial_t \mathbf{D} \;+\; \mathbf{J}) \;=\; 0 \tag{7.12}$$

The first equation implies the zero-divergence condition (7.4) for $\mathbf{B}$ *if, in addition*, zero divergence is imposed as the initial condition at any given moment of time. Alternatively, zero divergence can be easily deduced from Faraday's Law if the fields are time-harmonic (i.e. sinusoidal in time – more about this

case below). Without such additional assumptions, zero divergence does not in general follow from Faraday's Law.

Similar considerations show a close connection, but not complete equivalence, between the divergence equation (7.12) and conservation of charge. Substituting $\nabla \cdot \mathbf{D} = \rho$ (7.3) into (7.12) gives

$$\nabla \cdot \mathbf{J} \;=\; -\,\partial_t \rho \tag{7.13}$$

which is a mathematical expression of charge conservation.[3]

This logic cannot be completely reversed to produce the divergence equation for $\mathbf{D}$ from charge conservation and the curl equation for $\mathbf{H}$. Indeed, substituting conservation of charge (7.13) into (7.12), one obtains

$$\partial_t(\nabla \cdot \mathbf{D} \;-\; \rho) \;=\; 0 \tag{7.14}$$

which makes the divergence equation $\nabla \cdot \mathbf{D} = \rho$ true at *all* moments of time, *provided that* it holds at any given moment of time.

Time-harmonic fields can be described by complex phasors. It will always be clear from the context whether a time function or a phasor is being considered, and I shall therefore for simplicity of notation denote phasors with the same symbols as the corresponding time dependent fields ($\mathbf{H}$, $\mathbf{D}$, etc.), with little danger of confusion.

At the same time, we are facing a dilemma with regard to notational conventions on complex phasors themselves. Physicists usually assume that the actual $\mathbf{E}$-field can be obtained from its phasor as $\mathrm{Re}\{\mathbf{E}\exp(-\mathrm{i}\omega t)\}$, and similarly for other fields. Electrical engineers take the plus sign, $\exp(+\mathrm{i}\omega t)$, in the complex exponential. This notational difference is equivalent to replacing all phasors with their complex conjugates. Unfortunately, material parameters also get replaced with their conjugates, and confusion may arise, say, if engineers take the dielectric permittivity from the physical data measured in the "wrong" quadrant. In addition, physicists and mathematicians typically use symbol i for the imaginary unit, while engineers prefer j.

All these conventions are of course equally valid, but a notational mismatch could easily lead to sign errors. A little trick may prove helpful. Throughout the book, symbol i is used for the imaginary unit. The reader accustomed to the electrical engineering convention for phasors, $\exp(+\mathrm{i}\omega t)$, should simply assume that $\mathrm{i} \equiv j$; the physicist should set $\mathrm{i} \equiv -i$.

$$\text{Electrical engineers}: \quad \mathrm{i} \;\equiv\; j$$

$$\text{Physicists}: \quad \mathrm{i} \;=\; -i$$

[3] Charge conservation is more easily noted if this equation is put into integral form, $\oint_{\partial\Omega} \mathbf{J} \cdot \mathbf{dS} = -d_t Q$. The current flowing out of a closed volume is equal to the rate of depletion of electric charge inside that volume.

With these reservations in mind, Maxwell's equations for the phasors of time-harmonic fields are

$$\nabla \times \mathbf{E} = -\mathrm{i}\omega\mathbf{B} \tag{7.15}$$

$$\nabla \times \mathbf{H} = \mathrm{i}\omega\mathbf{D} + \mathbf{J} \tag{7.16}$$

Maxwell's "divergence equations" (7.3), (7.4) do not involve time derivatives and are therefore unchanged in the frequency domain.

For time-harmonic fields, zero divergence for $\mathbf{B}$ follows directly and immediately from (7.15), and conservation of charge follows from (7.16).

7.3 One-Dimensional Problems of Wave Propagation

7.3.1 The Wave Equation and Plane Waves

The simplest, and yet important and instructive, case for electromagnetic analysis involves fields that are independent of two Cartesian coordinates (say, y and z) and may depend only on the third one (x); the medium is assumed to be source-free ($\rho = 0$, $\mathbf{J} = 0$), isotropic and homogeneous, with parameters ϵ and μ independent of the spatial coordinates and time. Divergence equations (7.3) and (7.4) in this case yield

$$\frac{\partial D_x}{\partial x} = 0, \qquad \frac{\partial B_x}{\partial x} = 0 \tag{7.17}$$

and hence D_x and B_x must be constant. These trivial uniform electro- and magnetostatic fields are completely disassociated from the rest of the analysis and will hereafter be ignored.

In the absence of the x-component of the fields, the curl equations (7.1) and (7.2) become

$$\frac{\partial E_y}{\partial x} = -\mu\frac{\partial H_z}{\partial t}; \qquad \frac{\partial E_z}{\partial x} = \mu\frac{\partial H_y}{\partial t} \tag{7.18}$$

$$\frac{\partial H_y}{\partial x} = \epsilon\frac{\partial E_z}{\partial t}; \qquad -\frac{\partial H_z}{\partial x} = \epsilon\frac{\partial E_y}{\partial t} \tag{7.19}$$

It is not hard to see that the equations have decoupled into two pairs:

$$\frac{\partial E_y}{\partial x} = -\mu\frac{\partial H_z}{\partial t}; \qquad -\frac{\partial H_z}{\partial x} = \epsilon\frac{\partial E_y}{\partial t} \tag{7.20}$$

$$\frac{\partial E_z}{\partial x} = \mu\frac{\partial H_y}{\partial t}; \qquad \frac{\partial H_y}{\partial x} = \epsilon\frac{\partial E_z}{\partial t} \tag{7.21}$$

These pairs of equations correspond to two separate waves: one with the (E_y, H_z) components of the fields and the other one with the (E_z, H_y) components. In optics and electromagnetics, it is customary to talk about different *polarizations* of the wave; by convention, it is the direction of the *electric* field

that defines polarization. Thus the wave of (7.20) is said to be polarized in the y-direction, while the wave of (7.21) is polarized in the z-direction.

We can now focus on one of the waves – say, on the (E_y, H_z) wave (7.20) – because the other one is completely similar. The magnetic field can be eliminated by differentiating the first equation in (7.20) with respect to x, the second one with respect to time and then adding these equations to remove the mixed derivative of the H-field. This leads to the wave equation

$$\frac{\partial^2 E_y}{\partial x^2} \; - \; \mu\epsilon \, \frac{\partial^2 E_y}{\partial t^2} \; = \; 0 \tag{7.22}$$

It is straightforward to verify, using the chain rule of differentiation, that any field of the form

$$E_y(x,t) = g(v_p t \pm x) \tag{7.23}$$

satisfies the governing equation (7.22) if g is an arbitrary twice-differentiable function and v_p is

$$v_p \; = \; \frac{1}{\sqrt{\mu\epsilon}} \tag{7.24}$$

For example, $E_y(x,t) = (v_p t - x)^2$ and $E_y(x,t) = \cos k(v_p t - x)$, where k is a given parameter, are valid waves satisfying the electromagnetic equations.

Physically, (7.23) represents a waveform that propagates in space without changing its shape (the shape is specified by the g function). Let us trace the motion of any point with a fixed value of E_y on the waveform. The fixed value of the field implies zero full differential

$$dE_y \; = \; \frac{\partial E_y}{\partial t} \, dt \; + \; \frac{\partial E_y}{\partial x} \, dx \; = \; g' v_p dt \; \pm \; g' dx \; = \; 0 \tag{7.25}$$

and hence (for a nonzero derivative g')

$$\frac{dx}{dt} \; = \; - \, \frac{\partial E_y}{\partial t} \, / \, \frac{\partial E_y}{\partial x} \; = \; \mp \, v_p \tag{7.26}$$

Thus any point on the wave form moves with velocity v_p; it can also be said that the waveform as a whole propagates with this velocity. Note that for $v_p > 0$ the $g(x - v_p t)$ wave moves in the $+x$-direction, while the $g(x + v_p t)$ wave moves in the $-x$-direction. In the very common particular case where the waveform g is sinusoidal, the point of constant value of the field is also the point of constant phase. For this reason, v_p is known as *phase velocity.*

To solve the wave equation (7.22), let us apply the Fourier transform. The transforms will sometimes be marked by the hat symbol; in many cases, however, for the sake of simplicity no special notation will be used and complex phasors will be identified from the context and/or by the argument ω. In this section, let us also drop the y subscript, as the field has only one component. Then the wave equation becomes

$$E''(x) \; + \; \omega^2 \mu\epsilon E(x) \; = \; 0 \tag{7.27}$$

where the prime indicates the x-derivative. This is the Helmholtz equation whose general solution $E(x)$ is a superposition of two *plane waves* $E_\pm \exp(\pm kx)$, so called because their surfaces of equal phase are planes.

$$E(x) \;=\; E_+ \exp(\mathrm{i}kx) \;+\; E_- \exp(-\mathrm{i}kx) \tag{7.28}$$

where $E_\pm$ are some amplitudes and

$$k \;=\; \omega\sqrt{\mu\epsilon} \tag{7.29}$$

is the *wavenumber*. Since k enters the solution (7.28) with both plus and minus signs, it is at this point unimportant which branch of the square root is chosen to define k in (7.29). This issue will become nontrivial later, in the context of backward waves and negative refraction.

7.3.2 Signal Velocity and Group Velocity

Plane waves cannot be used as "signals"; they do not transfer energy or information because, by definition, they exist forever and everywhere. Thus, unavoidably, information transfer must involve more than one frequency.

Now, the standard textbook argument goes like this: consider a superposition of *two* waves, for simplicity of the same amplitude, with slightly different frequencies $\omega \pm \Delta\omega$ ($\Delta\omega \ll \omega$). Simple algebra gives

$$\begin{aligned} &\exp\left[\mathrm{i}\left((\omega+\Delta\omega)t-(k+\Delta k)x\right)\right] \;+\; \exp\left[\mathrm{i}\left((\omega-\Delta\omega)t-(k-\Delta k)x\right)\right] \\ &\qquad = \; 2\,\exp\left[\mathrm{i}(\omega t-kx)\right]\cos\left(\Delta\omega t-\Delta kx\right) \end{aligned}$$

The cosine term can be viewed as a low-frequency ($\Delta\omega$) "signal" and the complex exponential as a high-frequency (ω) carrier wave. The "signal" $\cos(\Delta\omega t - \Delta kz)$ manifests itself as beats on the carrier wave and propagates with the *group velocity* $v_g = \Delta\omega/\Delta k$ (the "group" consisting of just two waves in this idealized case). The $\Delta\omega \to 0$ limit

$$v_g \;=\; \frac{\partial\omega}{\partial k} \tag{7.30}$$

is then declared to be "signal velocity" – different from the phase velocity $v_p = \omega/k$.

However, if a single monochromatic wave contains zero information, one may wonder how it may be possible for two such waves – or any finite number of plane waves for that matter – to carry a nonzero amount of information.[4] Indeed, the train of beats is no less predictable than a single plane wave and also is present, theoretically, everywhere and forever. It cannot therefore be used as a signal any more than a single plane wave can.

[4] This is why the word "signal" was put in quotes in the previous paragraph.

A completely rigorous analysis must rely on precise definitions of "information" and "signal" – a territory into which I will not attempt to venture here and which would take us too far from the main subjects of this chapter. Instead, following the books by L. Brillouin [Bri60] and P.W. Milonni [Mil04], let us note that an observer can receive a nonzero amount of information only if the future behavior of the wave cannot be determined from its values in the past. This implies, in particular, that an information-carrying wave has to be, in the mathematical sense, non-analytic.

As a characteristic example, consider a pointwise source capable of generating an arbitrary (not necessarily analytic!) field at $x = 0$. Let us use this source to produce amplitude modulation

$$E(0,t) \;=\; \mathcal{E}(0,t)\exp(\mathrm{i}\omega_0 t) \tag{7.31}$$

where $\mathcal{E}(t)$ is a low-frequency waveform that can be used to carry (useful) information and ω_0 is the carrier frequency. To find the field at any $x > 0$, we Fourier-transform the wave equation and assume only outgoing waves $E(0,\omega)\exp(\mathrm{i}\omega t - k(\omega)x)$. The Fourier transform $E(0,\omega)$ is found from the given field at $x = 0$:

$$E(0,\omega) \;=\; \int_{-\infty}^{\infty} \mathcal{E}(0,t)\exp(\mathrm{i}\omega_0 t)\exp(-\mathrm{i}\omega t)dt \;=\; \hat{\mathcal{E}}(\omega-\omega_0)$$

That is, the modulation shifts the spectrum of $\mathcal{E}$ by ω_0, as is well known in signal analysis. The complex field phasor at an arbitrary $x > 0$ then is

$$E(x,\omega) \;=\; \hat{\mathcal{E}}(0,\omega-\omega_0)\,\exp(-\mathrm{i}k(\omega)x) \tag{7.32}$$

If there is no dispersion, i.e. the velocity of the wave is frequency-independent, $k(\omega) = \omega/v_p$ and

$$E(x,\omega) \;=\; \hat{\mathcal{E}}(0,\omega-\omega_0)\,\exp\left(-\mathrm{i}\omega\,\frac{x}{v_p}\right) \quad \text{(no dispersion)}$$

the inverse Fourier transform of which is

$$E(x,t) \;=\; \mathcal{E}\left(t-\frac{x}{v_p}\right) \quad \text{(no dispersion)}$$

The wave arrives at the observation point x unmolested, only with a time delay x/v_p.

We are, however, interested in the general case with dispersion. The time-dependent field can be found from its Fourier transform (7.32) as

$$E(x,t) \;=\; \int_{-\infty}^{\infty} \hat{\mathcal{E}}(0,\omega-\omega_0)\,\exp(-\mathrm{i}k(\omega)x)\exp(\mathrm{i}\omega t)\,d\omega \tag{7.33}$$

which gives the low-frequency "signal" $\mathcal{E}(x,t)$

$$\mathcal{E}(x,t) \;=\; E(x,t)\exp(-\mathrm{i}\omega_0 t) \;=\; \int_{-\infty}^{\infty} \hat{\mathcal{E}}(0,\omega')\,\exp(-\mathrm{i}k(\omega')x)\exp(\mathrm{i}\omega' t)\,d\omega' \tag{7.34}$$

where

$$\omega' \equiv \omega - \omega_0$$

The velocity of this signal can be found from the condition of zero differential $d\mathcal{E}(x,t)$ in full analogy with equations (7.25) and (7.26); this velocity is the ratio of partial differentials of $\mathcal{E}(x,t)$ with respect to t and x. These partial derivatives are

$$\frac{\partial \mathcal{E}}{\partial x} \;=\; \mathrm{i}\int_{-\infty}^{\infty} k(\omega')\,\mathcal{E}(\omega')\,\exp(\mathrm{i}k(\omega')x)\exp(-\mathrm{i}\omega' t)\,d\omega' \tag{7.35}$$

and

$$\frac{\partial \mathcal{E}}{\partial t} \;=\; -\,\mathrm{i}\int_{-\infty}^{\infty} \omega'\,\mathcal{E}(\omega')\,\exp(\mathrm{i}k(\omega')x)\exp(-\mathrm{i}\omega' t)\,d\omega' \tag{7.36}$$

So far the expressions have been exact; now an approximation is needed to find a relationship between the two partial derivatives. Since $\mathcal{E}(t)$ is a low-frequency function, the main contribution to the Fourier transforms comes from the small values of $\omega' = \omega - \omega_0$. Hence, expressing ω' with first-order accuracy with respect to small k as

$$\omega' \;\approx\; k\,\frac{\partial \omega'}{\partial k}(0) \qquad (\text{small } k)$$

one has

$$\frac{\partial \mathcal{E}}{\partial t} \approx \;-\,\mathrm{i}\frac{\partial \omega'}{\partial k}(0)\int_{-\infty}^{\infty} k\mathcal{E}(\omega')\exp(\mathrm{i}kx)\exp(-\mathrm{i}\omega' t)dt$$

Therefore the velocity of the signal is

$$v_{\text{signal}} \;\approx\; \frac{\partial \mathcal{E}}{\partial t}\Big/\frac{\partial \mathcal{E}}{\partial x} \;=\; \frac{\partial \omega}{\partial k} \;\equiv\; v_g \tag{7.37}$$

Thus group velocity $\partial\omega/\partial k$, contrary to what some textbooks may lead one to believe, is only an *approximation* of signal velocity (P.W. Milonni elaborates on this in [Mil04]). As the derivation above shows, the accuracy of this approximation depends on the deviation of the dispersion curve $\omega(k)$ from a straight line within the frequency range $[\omega_0 - \omega_{\mathcal{E}},\, \omega_0 + \omega_{\mathcal{E}}]$, where $[-\omega_{\mathcal{E}},\, \omega_{\mathcal{E}}]$ is the characteristic frequency band for the signal $\mathcal{E}$ (beyond which its amplitude spectrum is zero or can be neglected); it is assumed that $\omega_{\mathcal{E}} \ll \omega_0$.

One may not be satisfied with these approximations and may wish to define signal velocity *exactly*. However, the precise definition is elusive. Indeed, consider a broadband signal such as a sharp pulse. Its high frequency components can, at least in principle, be used to convey information. But at high frequencies the material parameters tend to their free space values ϵ_0 and μ_0, and hence group velocity tends to the speed of light. Thus – as a matter of

principle and disregarding all types of noise – information can be transferred with the velocity of light *in any medium.*

An equivalent and instructive physical interpretation is given by A. Sommerfeld ([Bri60], p. 19), with attribution to W. Voigt:

> "We will show here that the wave front velocity is always identical with the velocity of light in vacuum, c, irrespective of whether the material is normally or anomalously dispersive, whether it is transparent or opaque, or whether it is simply or doubly refractive. The proof is based on the theory of dispersion of light, which explains the various optical properties of materials on the basis of the forced oscillations of the particles of the material, either electrons or ions. ... According to our present knowledge ..., there exists only one isotropic medium for electrodynamic phenomena, the vacuum, and the deviations from vacuum properties can be traced back to the forced oscillations of charges. When the wave front of our signal makes its way through the optical medium, it finds the particles which are capable of oscillating originally at rest ..., (except for their thermal motion which has no effect on propagation, due to its randomness). Originally, therefore, the medium seems optically empty; only after the particles are set into motion, can they influence the phase and form of the light waves. The propagation of the wavefront, however, proceeds undisturbed with the velocity of light in vacuum, independently of the character of the dispersing ions."

7.3.3 Group Velocity and Energy Velocity

The relationship between group velocity and the Poynting vector has substantial physical significance in its own right but even more so in connection with backward waves and negative refraction, to be discussed later in this chapter (Section 7.13). Let us consider a homogeneous source-free isotropic material with frequency-dependent parameters $\epsilon(\omega)$ and $\mu(\omega)$. Losses at a given operating frequency (but not necessarily at other frequencies) will be neglected, so that both ϵ and μ are real.

A y-polarized plane wave propagating in the x-direction is governed by the equation

$$E''(x) \;+\; k^2 E(x) \;=\; 0, \qquad k^2 \equiv \omega^2 \epsilon(\omega)\mu(\omega) \tag{7.38}$$

where $E = E_y$, and has the form

$$E(x) = E_0 \exp(-\mathrm{i}kx) \tag{7.39}$$

The magnetic field $H = H_z$ is

$$H(x) \;=\; H_0 \exp(-\mathrm{i}kx); \qquad H_0 = \frac{k}{\omega\mu} E_0 \;=\; \left(\frac{\epsilon}{\mu}\right)^{\frac{1}{2}} E_0 \tag{7.40}$$

Power flux is characterized by the time-averaged Poynting vector with the x-component only:

$$\langle P \rangle \;\equiv\; \langle P \rangle_x \;=\; \frac{1}{2}\,\mathrm{Re}(EH^*) \;=\; \frac{1}{2}\,|E_0|^2\,\mathrm{Re}\left(\frac{\epsilon}{\mu}\right)^{\frac{1}{2}} \tag{7.41}$$

If one is interested in the wave with power flow in the $+x$ direction, then the real part of k is positive and the square root in (7.41) is the one with a positive real part.

Since group velocity and the Poynting vector are related to the propagation of signals and energy, respectively, there is a connection between them. For the group velocity, we have

$$v_g^{-1} \;=\; \frac{\partial k}{\partial \omega} \;=\; \left(2\epsilon\mu + \omega\mu\,\frac{\partial \epsilon}{\partial \omega} + \omega\epsilon\,\frac{\partial \mu}{\partial \omega}\right)\,\frac{1}{2}\,(\epsilon\mu)^{-\frac{1}{2}} \tag{7.42}$$

The amount of field energy transferred through a surface element $dS = dy\,dz$ over the time interval dt is equal to $w\,dS\,dx \;=\; w\,dS\,v_E dt$, where w is the volume energy density and v_E is energy velocity. On the other hand, the same transferred energy is equal to $P\,dS\,dt$; hence

$$w\,dS\,v_E dt \;=\; PdSdt$$

or simply

$$w\,v_E \;=\; P \tag{7.43}$$

If one assumes that energy, like signals, propagates with group velocity (under the approximation assumptions considered above), i.e. $v_E \;=\; v_g$, then the volume energy density can be obtained from (7.43) and (7.42). After some algebra,

$$w \;=\; Pv_g^{-1} \;=\; \frac{1}{4}\left(\frac{\partial(\omega\epsilon)}{\partial\omega}\,|E|^2 \;+\; \frac{\partial(\omega\mu)}{\partial\omega}\,|H|^2\right) \tag{7.44}$$

where the relationship between the electric and magnetic field amplitudes, as specified in (7.40), has been worked into this expression to make it symmetric with respect to both fields.

This result for dispersive media is well established in the physics literature (L.D. Landau & E.M. Lifshitz [LL84], L. Brillouin [Bri60]) and is notably different from the classical formula for static fields

$$w_{\text{static}} \;=\; \frac{1}{2}\,\left(\epsilon\,|E|^2 \;+\; \mu\,|H|^2\right)$$

The difference between the numerical factors in the "static" and "dynamic" expressions for the energy density is natural, as the additional 1/2 in (7.44) reflects the usual "effective value" of sinusoidally oscillating quantities. More interesting is the dependence of energy in a dispersive medium on the ω-derivatives of ϵ and μ. The physical nature of these additional terms is explained by Brillouin ([Bri60], pp.88–93):

> "The energy ... at the time when E passes through zero is quite different from the zero energy that the dielectric has after being isolated from an electric field for a long time. In order to explain the fact that the permittivity ϵ of the dielectric is different from that of the vacuum, ϵ_0, one must admit that the medium contains mobile charges, electrons or ions in motion or electric dipoles capable of orientation; then, one takes as the zero energy of the system the condition that all the charged particles are at rest in their equilibrium positions. ... all the charged particles may pass by their equilibrium positions at the time $t = 0$ when the field vanishes, but they pass them with nonzero velocity. [The additional term] represents the *kinetic energy of all the charged particles* contained in the dielectric."

7.4 Analysis of Periodic Structures in 1D

Much of research in nano-photonics is related to electromagnetic wave propagation in periodic structures with a characteristic size comparable with, but smaller than, the wavelength. The mathematical side of the analysis is centered at differential equations with periodic coefficients. We therefore start with a summary of the relevant mathematical theory, first for ordinary differential equations, and then generalizations to two and three dimensions.

This section will focus on key ideas and results important from the physical perspective; further mathematical details can be found in the monographs by M.S.P. Eastham [Eas73] and by W. Magnus & S. Winkler [MW79]. In a condensed form, the theory is given in W. Walter's book [Wal98b]. For applications in optics and photonics, books by P. Yeh [Yeh05] and K. Sakoda [Sak05] are recommended.

Very useful insights can be gained from one-dimensional analysis. In media with one-dimensional periodicity along the x-axis, the source-free one-component field satisfies equations (7.134) or (7.136), which are particular cases of *Hill's equation*

$$d_x \left(P(x)\, d_x u\right) \;+\; Q(x) u \;=\; 0 \tag{7.45}$$

Here u is the single Cartesian component of either the electric or magnetic field; d_x denotes the x-derivative. $P(x)$, $Q(x)$ are known functions (possibly complex-valued), periodic in x with a period x_0:

$$P(x + x_0) \;=\; P(x), \qquad Q(x + x_0) \;=\; Q(x), \quad \forall x \in R \tag{7.46}$$

Although much of the analysis below can be generalized to arbitrary second order equations with periodic coefficients and to higher order equations, it is Hill's equation that is most relevant to 1D problems in nano-photonics.

For theoretical analysis of Hill's equation, it is convenient to rewrite this second-order equation as a system of two first-order equations with a vector of unknowns $(u, v)^T$, where $v \equiv P(x) d_x u$:

$$d_x u = P^{-1}(x)\, v \tag{7.47}$$
$$d_x v = -\, Q(x) u \tag{7.48}$$

or in matrix-vector form

$$d_x w = A w, \quad w \equiv \begin{pmatrix} u \\ v \end{pmatrix}, \quad A \equiv \begin{pmatrix} 0 & P^{-1}(x) \\ -Q(x) & 0 \end{pmatrix} \tag{7.49}$$

Under quite general assumptions on the smoothness of $P(x)$, $Q(x)$, solutions of this system exist and form a two-dimensional space. If two solutions $\psi_1(x)$ and $\psi_2(x)$ are a basis in this space (i.e. are linearly independent), it is helpful to combine them into a 2×2 matrix $\Psi(x)$ with columns $\psi_1(x)$ and $\psi_2(x)$. Clearly, this matrix itself satisfies the differential equation (7.49), i.e.

$$d_x \Psi(x) = A \Psi(x) \tag{7.50}$$

because this equation holds true column-wise. Further, let $\psi_1(x)$ and $\psi_2(x)$ be a *special* pair of basis functions that correspond to the initial conditions

$$\Psi(0) = I \tag{7.51}$$

Matrix $\Psi(x)$ is then called the *fundamental* matrix of the system.

Any solution $\tilde{\psi}(x)$ can be expressed as a linear combination of basis functions $\psi_1(x)$, $\psi_2(x)$

$$\tilde{\psi}(x) = \Psi(x)\, c \tag{7.52}$$

where c is some constant column vector in $\mathbb{C}^2$. Consequently, any solution $\tilde{\Psi}(x)$ of *matrix* equation (7.50) is linearly related to the fundamental matrix $\Psi(x)$:

$$\tilde{\Psi}(x) = \Psi(x)\, C \tag{7.53}$$

where C is some time-independent 2×2 matrix.

Let us now take into account the periodicity of the coefficients. It is clear that translation of any solution by the spatial period x_0 is also a solution. In particular, $\tilde{\Psi}(x) \equiv \Psi(x + x_0)$ is a solution. As such, it must be linearly related to the fundamental matrix by (7.53), i.e.

$$\Psi(x + x_0) = \Psi(x)\, C \tag{7.54}$$

Here (with a slight abuse of notation) matrix C is a particular instance of the generic matrix C in (7.53). Setting $x = 0$ in (7.54) yields $C = \tilde{\Psi}(0)$, because $\Psi(0) = I$ by the definition of the fundamental matrix. With this in mind, the translated solution can now be expressed as

$$\tilde{\Psi}(x) \equiv \Psi(x + x_0) = \Psi(x)\, \Psi(x_0) \tag{7.55}$$

At first glance, since the coefficients of the underlying equation are periodic, one may want to look for two linearly independent solutions that would also

be periodic with period x_0. This quickly turns out to be a false trail. In fact, even a single periodic solution in general does not exist. A trivial example is the equation with constant coefficients $y'' - y = 0$ that has only non-periodic exponential solutions.

The "right" idea is to weaken the periodicity condition and look for "scaled-periodic" solutions:

$$u(x + x_0) \;=\; \lambda u(x), \quad \forall x \in \mathbb{R} \tag{7.56}$$

where λ is a yet undetermined parameter – possibly complex, even if the equation itself is real. (Caution: "scaled-periodic" is not a standard term. However, it is descriptive and intuitive enough to be adopted here.)

This condition can be written in an equivalent form if the solution is "unscaled" by introducing

$$u_{\mathrm{PER}}(x) \;=\; \lambda^{-x/x_0} u(x) \tag{7.57}$$

where subscript "PER" connotes periodicity (that will become obvious very soon). In terms of $u_{\mathrm{PER}}(x)$, condition (7.56) simplifies just to

$$u_{\mathrm{PER}}(x + x_0) \;=\; u_{\mathrm{PER}}(x) \quad \forall x \in \mathbb{R} \tag{7.58}$$

That is, function $u_{\mathrm{PER}}(x)$ is periodic with the period x_0. Returning to the original function $u(x)$, one obtains

$$u(x) \;=\; \lambda^{x/x_0}\, u_{\mathrm{PER}}(x), \;\; \text{with} \;\; u_{\mathrm{PER}}(x + x_0) \;=\; u_{\mathrm{PER}}(x) \quad \forall x \in \mathbb{R} \tag{7.59}$$

This result can be rewritten in a more conventional form by introducing a new parameter K_B such that $\lambda = \exp(-\mathrm{i}K_B x_0)$:

$$u(x) \;=\; \exp(-\mathrm{i}K_B x)\, u_{\mathrm{PER}}(x), \;\; \text{with} \;\; u_{\mathrm{PER}}(x + x_0) \;=\; u_{\mathrm{PER}}(x) \quad \forall x \in \mathbb{R} \tag{7.60}$$

Subscript "B" is introduced in honor of Felix Bloch[5] but will occasionally be dropped if there is no possibility of confusion with other possible interpretations of symbol K.

The motivation for introducing the new parameter K_B is that the most interesting practical case occurs when $|\lambda| = 1$ and consequently K_B is purely real (see below). Then the complex exponential has a clear physical meaning as a phase factor. In particular, $\exp(-\mathrm{i}K_B x_0)$ is the phase shift over one lattice cell.

Equation (7.60) represents a "scaled-periodic" solution $u(x)$ as a product of a periodic function and – for real K_B – a traveling *Bloch wave*. Such waves play a central role in the analysis of periodic structures. Note that in general the wavelength $2\pi/K_B$ corresponding to the $\exp(-\mathrm{i}K_B x)$ factor is different

[5] Felix Bloch (1905–1983), Swiss-American physicist, 1952 Nobel Prize winner in Physics; http://nobelprize.org/physics/laureates/1952/bloch-bio.html

from the spatial period x_0. Determining the connection between the two is one of the objectives of the analysis.

To find the "scaled-periodic" function $u(x)$, we first note that, as any solution, it can be expressed as a linear combination of the fundamental solutions of the differential equation:

$$u(x) \;=\; \Psi(x)\, c \tag{7.61}$$

with some coefficient vector c. The condition of scaled periodicity then is

$$\Psi(x + x_0)\, c \;=\; \lambda \Psi(x)\, c \tag{7.62}$$

or, with (7.55) in mind,

$$\Psi(x)\, \Psi(x_0)\, c \;=\; \lambda \Psi(x)\, c \tag{7.63}$$

The fundamental matrix $\Psi(x)$ is nonsingular, and hence

$$\Psi(x_0)\, c \;=\; \lambda c \tag{7.64}$$

Thus λ and c are an eigenvalue and a corresponding eigenvector of $\Psi(x_0)$. The analysis is reversible and scaled-periodicity (7.56) can be deduced from the eigenvalue condition (7.64).

While the eigenvalue problem of type (7.64) is general for linear ODE with periodic coefficients, one feature of matrix $\Psi(x_0)$ is special for Hill's equation:

$$\det \Psi(x) \;=\; 1, \quad \forall x \in R \tag{7.65}$$

This result follows from the Abel–Liouville–Jacobi–Ostrogradskii identity for the Wronskian; see e.g. E. Hairer *et al.* [HrW93], W. Walter [Wal98b]:

$$\det\, W(x) \;=\; \det\, W(0)\, \exp \int_0^x \mathrm{Tr}\, A(\xi)\, d\xi \tag{7.66}$$

This identity is valid for any linear system $d_x w \;=\; A(x) w$; the columns of matrix $W(x)$ form a set of linearly independent solutions of this system; as a reminder, the determinant of W is called the Wronskian.[6]

For Hill's equation, matrix A is defined in (7.49) and has a zero diagonal; hence $\mathrm{Tr}\, A = 0$ and the Abel–Liouville–Jacobi–Ostrogradskii identity yields

$$\det\, W(x) \;=\; \det\, W(0)$$

[6] Josef Hoëné de Wronski (1778–1853) proposed theories of everything in the Universe based on properties of numbers, designed caterpillar-like vehicles intended to replace railroad transportation, tried to square the circle, and attempted to build both a perpetual motion machine and a device to predict the future. He also studied infinite series whose coefficients are the determinants now known as the Wronskians. http://en.wikipedia.org/wiki/Josef_Wronski; http://www.angelfire.com/scifi2/rsolecki/jozef_maria_hoene_wronski.html

In particular, for the fundamental matrix $\Psi(x)$, since $\Psi(0) = I$ by definition, the determinant is equal to one for all x, as stipulated in (7.65).

It immediately follows that, for Hill's equation, the characteristic polynomial for $\Psi(x_0)$ is

$$\lambda^2 \; - \; \mathrm{Tr}\,\Psi(x_0)\,\lambda \; + \; 1 \; = \; 0 \tag{7.67}$$

and consequently

$$\lambda_1 \lambda_2 \; = \; 1 \tag{7.68}$$

where $\lambda_{1,2}$ are the eigenvalues (or possibly one eigenvalue of multiplicity two) of (7.64).

If the coefficients of the differential system, i.e. functions $P(x)$ and $Q(x)$, are real, then matrix $\Psi(x_0)$ is real as well, and the eigenvalues of (7.67) can either be real and reciprocal or, alternatively, complex conjugate and lying on the unit circle.

The characteristic equation has solutions

$$\lambda_{1,2} \; = \; \frac{1}{2}\left(\mathrm{Tr}\,\Psi(x_0) \; \pm \; \sqrt{\mathrm{Tr}^2\,\Psi(x_0) \; - \; 4}\right) \tag{7.69}$$

and hence the type of λ will depend on whether $|\mathrm{Tr}\,\Psi(x_0)|$ is greater or less than two, $|\mathrm{Tr}\,\Psi(x_0)| = 2$ being the borderline case.

If $|\mathrm{Tr}\,\Psi(x_0)| > 2$, the eigenvalues are real and the corresponding Bloch parameter K_B in (7.60) is purely imaginary. Equation (7.60) then shows a trend of exponential increase of the solution for $x \to \infty$ or $x \to -\infty$ (depending on the sign of λ). If the differential equation describes the field behavior in an infinite medium (the main subject of this chapter), such exponentially growing solutions are deemed nonphysical.

In contrast, for $|\mathrm{Tr}\,\Psi(x_0)| < 2$ the eigenvalues are complex conjugate and lie on the unit circle. This physically corresponds to solutions with a phase change, but no amplitude change, over x_0. Such solutions are called *Bloch–Floquet* (or simply *Bloch*) waves and are central in the electromagnetic analysis of periodic structures not only in 1D, but also in 2D and 3D (see subsequent sections).

For the borderline case $|\mathrm{Tr}\,\Psi(x_0)| = 2$, with its subcases, the eigenproblem is analyzed in detail by M.S.P. Eastham [Eas73]. The presentation below is different from, but ultimately equivalent to, Eastham's analysis. Instead of the individual eigenmodes of (7.56), let us consider a pair of fundamental solutions, with matrix $\Psi(x)$ satisfying the "scaled-periodicity" relation (7.55):

$$\Psi(x + x_0) \; = \; \Psi(x)\Psi(x_0) \tag{7.70}$$

where the scaling is effected by matrix $\Psi(x_0)$ rather than by parameter λ as in the scalar case. This equation can now be "unscaled" by the matrix variable change

$$\Psi_{\mathrm{PER}}(x) \; = \; \Psi(x)\,\left[\Psi(x_0)\right]^{-x/x_0} \tag{7.71}$$

This is conceptually similar to (7.57) and upon substitution into (7.70) shows that $\Psi_{\mathrm{PER}}(x)$ is indeed periodic (as its subscript "PER" suggests):

$$\Psi_{\mathrm{PER}}(x + x_0) \;=\; \Psi_{\mathrm{PER}}(x) \tag{7.72}$$

Hence the matrix solution of Hill's equation must have the form

$$\Psi(x) \;=\; \Psi_{\mathrm{PER}}(x)\; [\Psi(x_0)]^{x/x_0} \tag{7.73}$$

where $\Psi_{\mathrm{PER}}(x)$ is x_0-periodic.

Non-integer powers of matrices used in the expressions above need an accurate definition. The best source of information on matrix functions (and on matrix theory in general) is the monograph by F.R. Gantmakher [Gan59, Gan88]. For the purposes of this section, we only need a few facts about matrix functions.

Let matrix $\Psi(x_0)$ be represented in the Jordan form:

$$\Psi(x_0) \;=\; SJS^{-1} \tag{7.74}$$

where S is some transformation matrix. In general, J consists of blocks corresponding to the eigenvalues of the matrix; for each eigenvalue λ of multiplicity one, the corresponding "block" is just the diagonal element (1×1-matrix) λ; for each eigenvalue λ of multiplicity k, the corresponding $k \times k$ block contains λ on the diagonal and ones on the upper subdiagonal,[7] all other matrix elements being zero.

For a 2×2-matrix like $\Psi(x_0)$ in Hill's equation, the Jordan block is particularly simple. If the two eigenvalues are distinct, then

$$J \;=\; \mathrm{diag}(\lambda_1, \lambda_2) \tag{7.75}$$

For one eigenvalue λ of multiplicity two (and hence $\lambda = \pm 1$ due to (7.68)) the Jordan block can either still have the diagonal form (7.75) if two linearly independent eigenvectors exist or, alternatively,

$$J \;=\; \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}, \qquad \lambda = \pm 1 \tag{7.76}$$

Expression (7.73) for the fundamental matrix $\Psi(x)$ includes the power $[\Psi(x_0)]^{x/x_0}$, which is

$$[\Psi(x_0)]^{x/x_0} \;=\; SJ^{x/x_0}S^{-1} \tag{7.77}$$

where either

$$J^{x/x_0} \;=\; \mathrm{diag}(\lambda_1^{x/x_0},\; \lambda_2^{x/x_0}) \tag{7.78}$$

or

[7] Or, alternatively, on the lower subdiagonal – it is a matter of convention.

$$J^{x/x_0} \;=\; \begin{pmatrix} \lambda^{x/x_0} & \frac{x}{x_0}\lambda^{x/x_0-1} \\ 0 & \lambda^{x/x_0} \end{pmatrix} \tag{7.79}$$

It is convenient to denote $\lambda_{1,2} = \exp(-\mathrm{i}K_{1,2}x_0)$, where K is defined modulo 2π. In particular, $K = 0$ for $\lambda = +1$ and $K = \pi/x_0$ for $\lambda = -1$. Upon substitution into the main equation (7.73) for the fundamental matrix, one obtains in the case of distinct eigenvalues $\lambda_{1,2}$

$$\Psi(x)\,S \;=\; \Psi_{\mathrm{PER}}(x)\,S\,\mathrm{diag}\,(\exp(-\mathrm{i}K_1x),\,\exp(-\mathrm{i}K_2x)) \tag{7.80}$$

and for λ of multiplicity two

$$\Psi(x)\,S \;=\; \exp(-\mathrm{i}Kx)\,\Psi_{\mathrm{PER}}(x)\,S\begin{pmatrix} 1 & \frac{x}{x_0}\exp(\mathrm{i}Kx_0) \\ 0 & 1 \end{pmatrix} \tag{7.81}$$

The columns of matrix $\Phi(x) \equiv \Psi(x)S$, being linear combinations of the columns of $\Psi(x)$, form a pair of linearly independent solutions of Hill's equation. In the right hand sides of (7.80) and (7.81), matrix $\Phi_{\mathrm{PER}}(x) \equiv \Psi_{\mathrm{PER}}(x)S$ is x_0-periodic (since S is constant).

Thus we have found a matrix solution $\Phi(x)$ representable either as

$$\Phi(x) \;=\; \Phi_{\mathrm{PER}}(x)\begin{pmatrix} \exp(-\mathrm{i}K_1x) & 0 \\ 0 & \exp(-\mathrm{i}K_2x) \end{pmatrix} \tag{7.82}$$

or, alternatively, as

$$\Phi(x) \;=\; \exp(-\mathrm{i}Kx)\,\Phi_{\mathrm{PER}}(x)\begin{pmatrix} 1 & \frac{x}{x_0}\exp(\mathrm{i}Kx_0) \\ 0 & 1 \end{pmatrix} \tag{7.83}$$

In the diagonal case (7.82), both columns of $\Phi(x)$ are seen to be products of a periodic function and a complex exponential $\exp(-\mathrm{i}K_{1,2}x)$. For the Jordan form (7.83), the first column is a completely analogous product, but the second column is more complicated:

$$\psi_2 \;=\; \exp(-\mathrm{i}Kx)\left(\phi_{1,\mathrm{PER}} + \frac{x}{x_0}\,\phi_{2,\mathrm{PER}}\right), \qquad K = 0 \;\text{ or }\; \pi x_0^{-1} \tag{7.84}$$

This solution is periodic for $K = 0$ and antiperiodic for $K = \pi x_0^{-1}$. Eastham derives this in a different way [Eas73].

We now consider two examples of second-order equations with periodic coefficients: one illustrates a possible peculiar behavior of the solutions in both real and Fourier spaces; and the second one is key to understanding multilayered optical structures and photonic crystals, as discussed in Sections 7.5, 7.8.

Example 25. Equation

$$u''(x) \;+\; \exp(\mathrm{i}\kappa_0x)\,u(x) \;=\; 0; \qquad \kappa_0 \;=\; 2\pi x_0^{-1} \tag{7.85}$$

is an interesting illustrative case. Although the periodic coefficient is complex, much of the analysis above is still applicable.

Let us first assume that solution $u(x)$ has a valid Fourier transform $U(k)$ at least in the sense of distributions (a discrete spectrum is viewed as a particular case of a continuous spectrum – a set of Dirac delta-functions at some frequencies; see Appendix 6.15, p. 343). Since multiplication by $\exp(\mathrm{i}\kappa_0 x)$ amounts simply to a spatial frequency shift in the Fourier domain, and the second derivative translates into multiplication by $-k^2$, equation (7.85) becomes

$$-k^2 U(k) \;+\; U(k-\kappa_0) \;=\; 0 \tag{7.86}$$

Viewing this as a recursion relation

$$U(k-\kappa_0) \;=\; k^2 U(k) \tag{7.87}$$

one observes that the sequence of values $U(k-\kappa_0)$, $U(k-2\kappa_0)$, $U(k-3\kappa_0)$, $\ldots$, will generally be unbounded, with rapidly growing magnitudes. There is only one exception: this backward recursion gets terminated if $k = n\kappa_0$ for some positive integer n. Then $U(-\kappa_0) = U(-2\kappa_0) = \ldots = 0$ due to (7.87).

In this exceptional case, the spectrum is discrete, with some values U_n at spatial frequencies $k_n = n\kappa_0$ ($n = 0,1, \ldots$). Normalizing U_0 to unity and reversing recursion (7.87) to get

$$U_{n+1} \;=\; \frac{U_n}{\kappa_0^2\,(n+1)^2} \tag{7.88}$$

one obtains

$$U_n \;=\; \frac{1}{(n!\,\kappa_0^n)^2} \tag{7.89}$$

Hence one solution is expressed via the Fourier series

$$u(x) \;=\; \sum_{n=0}^{\infty} \frac{1}{(n!\,\kappa_0^n)^2}\,\exp(\mathrm{i}n\kappa_0 x) \tag{7.90}$$

Indeed, due to the presence of factorials in the denominators in (7.90), the Fourier series and its derivatives are uniformly convergent, so it is legal to differentiate the series and verify that its sum satisfies the original equation (7.85).

This Fourier series solution is obviously periodic with the period x_0. What about a second linearly independent solution? From the Fourier analysis above, it is clear that the second solution cannot have a valid Fourier transform. More specifically, it has to have the form (7.84).

The following numerical results for $x_0 = 1$ ($\kappa_0 = 2\pi$) illustrate the behavior of the solutions. The fundamental system $\psi_{1,2}$ was computed by high-order Runge–Kutta methods (see Section 2.4.1 on p. 20) for equation (7.85). (Matlab function `ode45` was used, with the relative and absolute tolerances of 10^{-10}.)

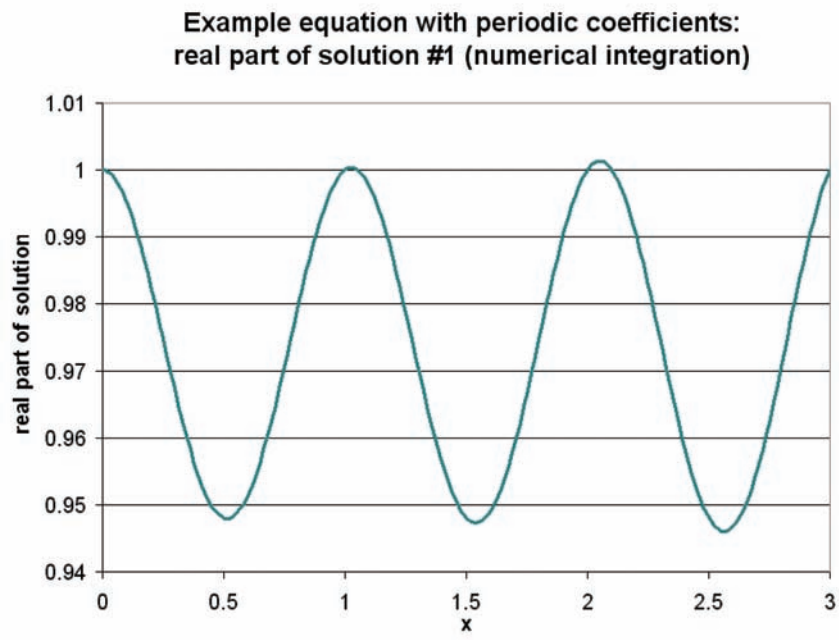


Fig. 7.1. The real part of the first fundamental solution for the example equation with a complex periodic coefficient.

For ψ_1, the initial conditions are $\psi_1(0) = 1$, $d_t\psi_1(0) = 0$; for ψ_2, $\psi_2(0) = 0$ and $d_t\psi_2(0) = 1$. The real and imaginary parts of these functions are plotted in Figs. 7.1–7.4 for reference. The governing matrix $\Psi(x_0)$ comprising the values of these solutions at $x_0 = 1$, is, with six digits of accuracy,

$$\Psi(x_0) \approx \begin{pmatrix} 1 - 0.165288\mathrm{i} & 1.051632 \\ 0.0259787 & 1 + 0.165288\mathrm{i} \end{pmatrix} \tag{7.91}$$

Matrix $\Psi(x_0)$ has a double eigenvalue of one, which numerically also holds with six digits of accuracy.

The Fourier series solution (7.90) of the original equation (7.85) is a linear combination of $\psi_{1,2}$ with the coefficients 1.025491 and 0.161179i. One way of finding this coefficient vector is to solve the linear system with matrix $\Psi(x_0)$ and the right hand side vector containing the values of the Fourier series solution and its derivative at $x = x_0$ (=1). This right hand side is $(1.025491, 0.161179\mathrm{i})^T$ – not coincidentally, identical with the coefficient vector above, as both of them are nothing other than the eigenvector of $\Psi(x_0)$ corresponding to the unit eigenvalue.

Example 26. We now turn to a case that is directly applicable to 1D-periodic multilayered structures in photonics. Consider a layered structure with

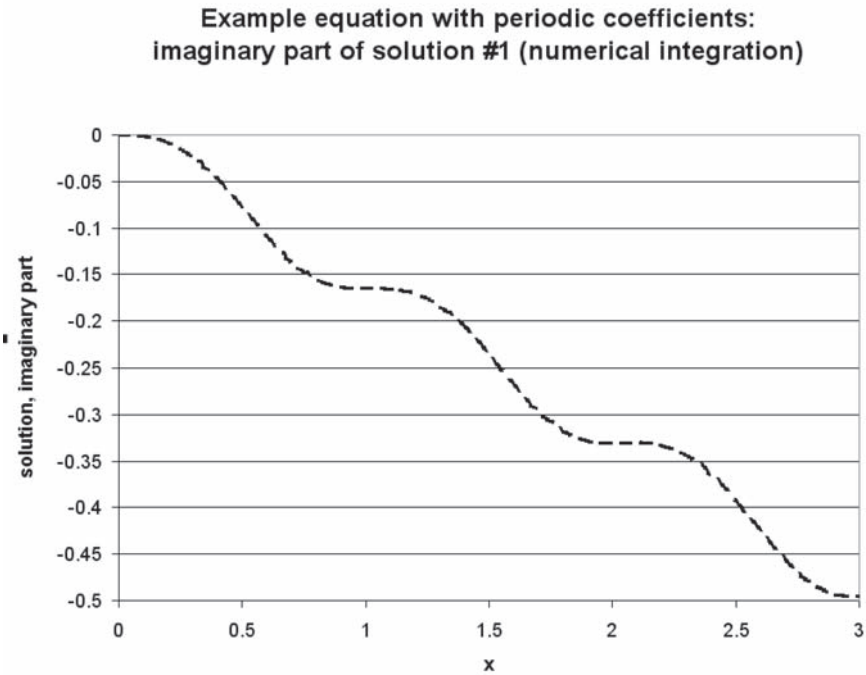


Fig. 7.2. The imaginary part of the first fundamental solution for the example equation with a complex periodic coefficient.

alternating electromagnetic material parameters ϵ_1, μ_1 and ϵ_2, μ_2 (Fig. 7.5). Let us focus on normal incidence (direction of propagation $\mathbf{k}$ perpendicular to the slabs); oblique incidence does not create any substantial difficulties. As theory prescribes, we first find the fundamental solutions and compute the transfer matrix $\Psi(x_0)$. However, since the coefficients of the underlying differential equation are now discontinuous, the equation should be treated in the weak form or, equivalently, the proper boundary conditions at the material interfaces should be imposed:

$$E_1(d_1) \;=\; E_2(d_1); \quad \mu_1^{-1}E_1'(d_1) \;=\; \mu_2^{-1}E_2'(d_1) \;\text{ at the interface}\, x = d_1 \tag{7.92}$$

where the origin ($x = 0$) is assumed to be at the left edge of the layer of thickness d_1. Similar conditions hold at $x = d_1 + d_2$ and all other interfaces.

The general solution of the differential equation within layer 1 is

$$E_1(x) \;=\; E_0\cos(k_1x) \;+\; k_1^{-1}E_0'\sin(k_1x), \quad k_1 = \omega(\mu_1\epsilon_1)^{\frac{1}{2}} \tag{7.93}$$

where the prime denotes x-derivatives and the coefficients E_0 and E_0' are equal to the values of E_1 and its derivative, respectively, at $x = 0$.

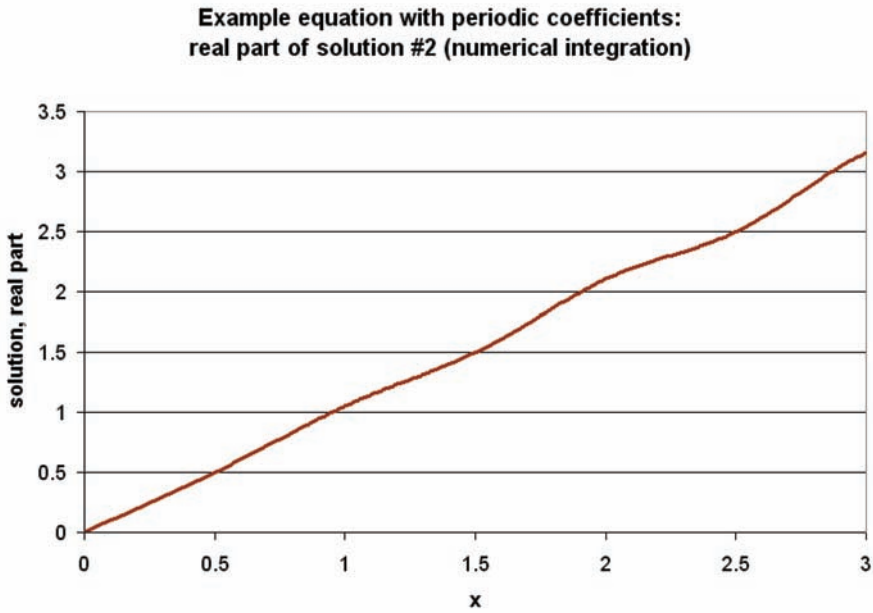


Fig. 7.3. The real part of the second fundamental solution for the example equation with a complex periodic coefficient.

We shall now "propagate" this solution through layers 1 and 2, with the final goal of obtaining the transfer matrix once the solution is evaluated over the whole period $x_0 = d_1 + d_2$.

First, we "follow" the solution to the interface between the layers, where it becomes

$$E_1(d_1) \;=\; E_0 \cos(k_1 d_1) \;+\; k_1^{-1} E_0' \sin(k_1 d_1) \tag{7.94}$$

and its derivative, one the side of layer 1, is

$$E_1'(d_1) \;=\; -\,k_1 E_0 \sin(k_1 d_1) \;+\; E_0' \cos(k_1 d_1) \tag{7.95}$$

Due to the interface boundary condition, the electric field and its derivative at $x = d_1$ *in the second layer* are

$$E_2(d_1) \;=\; E_1(d_1) \;=\; E_0 \cos(k_1 d_1) \;+\; k_1^{-1} E_0' \sin(k_1 d_1) \tag{7.96}$$

$$E_2'(d_1) \;=\; \frac{\mu_2}{\mu_1}\, E_1'(d_1) \;=\; -\,k_1 \frac{\mu_2}{\mu_1}\, [E_0 \sin(k_1 d_1) \;+\; E_0' \cos(k_1 d_1)] \tag{7.97}$$

Repeating this calculation for the second layer, with the "starting" values of the field and its derivative defined by (7.96), (7.97), one obtains the general solution *just beyond the second layer* at $x = (d_1 + d_2)_{+0}$. (Subscript "+0" indicates the limiting value from the right.)

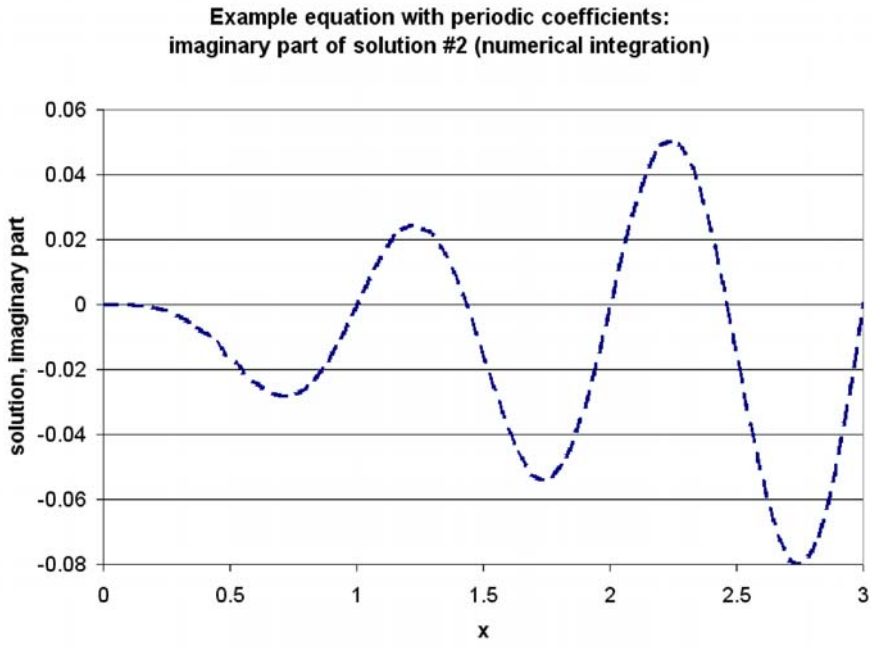


Fig. 7.4. The imaginary part of the second fundamental solution for the example equation with a complex periodic coefficient.

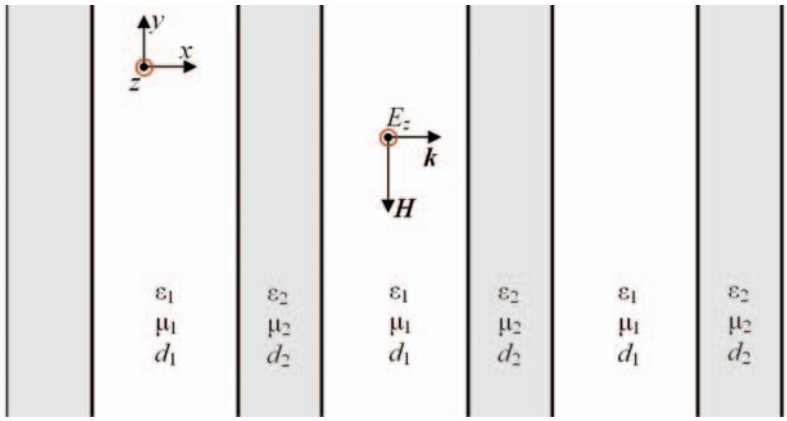


Fig. 7.5. An electromagnetic wave traveling through a multilayered 1D structure with normal incidence.

The first fundamental solution is obtained by setting $E_0 = 1$, $E_0' = 0$ and the second one by setting $E_0 = 0$, $E_0' = 1$. The transfer matrix $\Psi(d_1 + d_2)$ has these two solutions as its columns and is calculated to be

$$\Psi_{11}(d_1 + d_2)_{+0} = \cos(k_1 d_1)\cos(k_2 d_2) - \frac{k_1\mu_2}{k_2\mu_1}\sin(k_1 d_1)\sin(k_2 d_2) \quad (7.98)$$

$$\Psi_{12}(d_1 + d_2)_{+0} = \frac{\sin(k_1 d_1)\cos(k_2 d_2)}{k_1} + \frac{\mu_2\cos(k_1 d_1)\sin(k_2 d_2)}{\mu_1 k_2} \quad (7.99)$$

$$\Psi_{21}(d_1 + d_2)_{+0} = -\frac{\mu_1 k_2}{\mu_2}\cos(k_1 d_1)\sin(k_2 d_2) - k_1\sin(k_1 d_1)\cos(k_2 d_2) \quad (7.100)$$

$$\Psi_{22}(d_1 + d_2)_{+0} = -\frac{\mu_1 k_2}{\mu_2 k_1}\sin(k_1 d_1)\sin(k_2 d_2) + \cos(k_1 d_1)\cos(k_2 d_2) \quad (7.101)$$

The theoretical analysis in this section has shown that the nature of "scaled-periodic" solutions depends on the trace of $\Psi(d_1 + d_2)$:

$$\mathrm{Tr}\,\Psi(d_1 + d_2) = 2\cos(k_1 d_1)\cos(k_2 d_2) - \left(\frac{k_1\mu_2}{k_2\mu_1} + \frac{k_2\mu_1}{k_1\mu_2}\right)\sin(k_1 d_1)\sin(k_2 d_2) \quad (7.102)$$

This result is well known in optics – see e.g. J. Li *et al.* [LZCS03], I.V. Shadrivov *et al.* [SSK05], P. Yeh [Yeh05]. In the literature, equation (7.102) is derived in a somewhat different, but ultimately equivalent, way.

Numerical illustration. In the periodic structure of Fig. 7.5, assume that the widths of the layers are equal and normalized to unity, $d_1 = d_2 = 1$; materials are nonmagnetic (relative permeabilities $\mu_1 = \mu_2 = 1$); the relative dielectric constants are chosen as $\epsilon_1 = 1$, $\epsilon_2 = 5$.

For any given frequency ω, we can then calculate the trace of the transfer matrix by (7.102), with $k_{1,2} = \omega c^{-1}(\mu_{1,2}\epsilon_{1,2})^{1/2}$. This trace is plotted in Fig. 7.6. (The speed of light in free space is for simplicity normalized to one by a suitable choice of units.) As we have seen earlier in this section, propagating waves cannot exist in the infinite structure if the absolute value of the matrix trace exceeds two; the corresponding frequency gaps are shaded in Fig. 7.6.

The eigenvalues $\lambda_{1,2}$ of the Floquet problem are related to the matrix trace via (7.69). The absolute values of these roots are shown in Fig. 7.7. The bandgaps correspond to the real values of the roots (one of which is greater than one and the other one is less than one).

Within the pass bands, the roots lie on the unit circle: $\lambda_{1,2} = \exp(-\mathrm{i}K_{1,2}x_0)$, with the Bloch wavenumber K purely real (and defined modulo 2π). It is the relationship between this wavenumber and frequency that characterizes the bandgap structure. The plot of K vs. ω for our numerical example is shown in Fig. 7.8. It is customary, however, to rotate this plot: the wavenumber is displayed on the horizontal axis and frequency on the vertical one (Fig. 7.9).

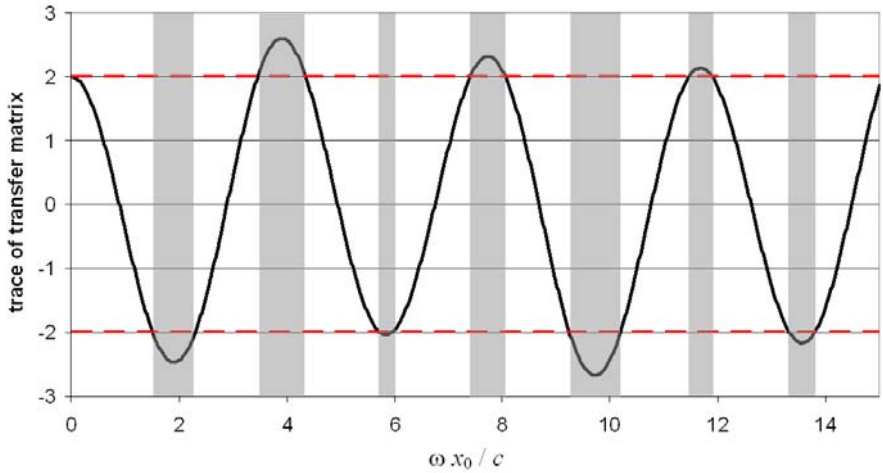


Fig. 7.6. The trace of the transfer matrix as a function of frequency. Periodic structure with $d_1 = d_2 = 1$; $\mu_1 = \mu_2$; $\epsilon_1 = 1$, $\epsilon_2 = 5$. Shaded areas indicate photonic bandgaps.

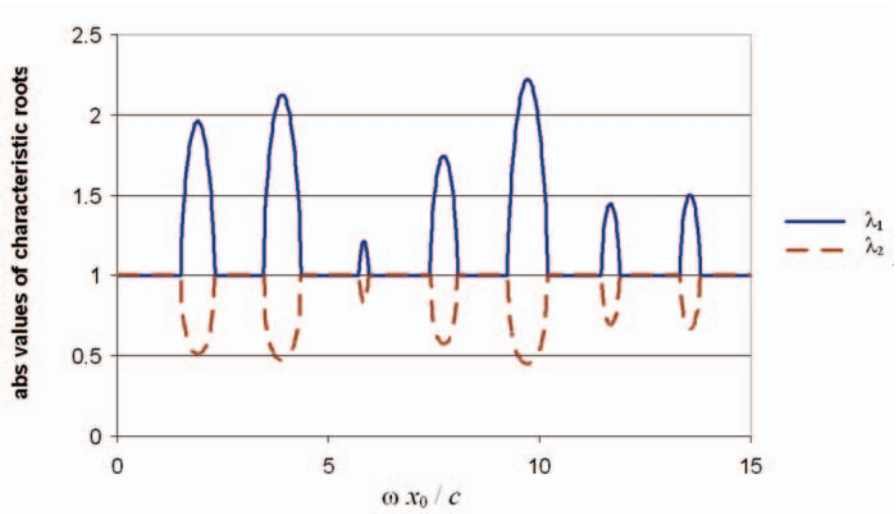


Fig. 7.7. Absolute values of the characteristic Floquet roots as a function of frequency. Periodic structure with $d_1 = d_2 = 1$; $\mu_1 = \mu_2 = 1$; $\epsilon_1 = 1$, $\epsilon_2 = 5$. Ranges with $|\lambda_{1,2}| \neq 1$ are photonic bandgaps.

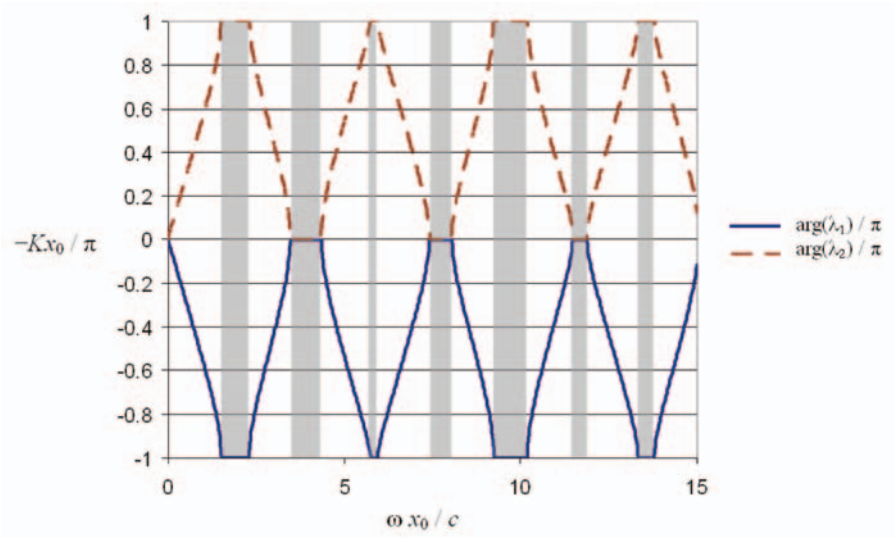


Fig. 7.8. The Bloch wavenumber as a function of frequency. Periodic structure with $d_1 = d_2 = 1$; $\mu_1 = \mu_2 = 1$; $\epsilon_1 = 1$, $\epsilon_2 = 5$. Shaded areas indicate photonic bandgaps.

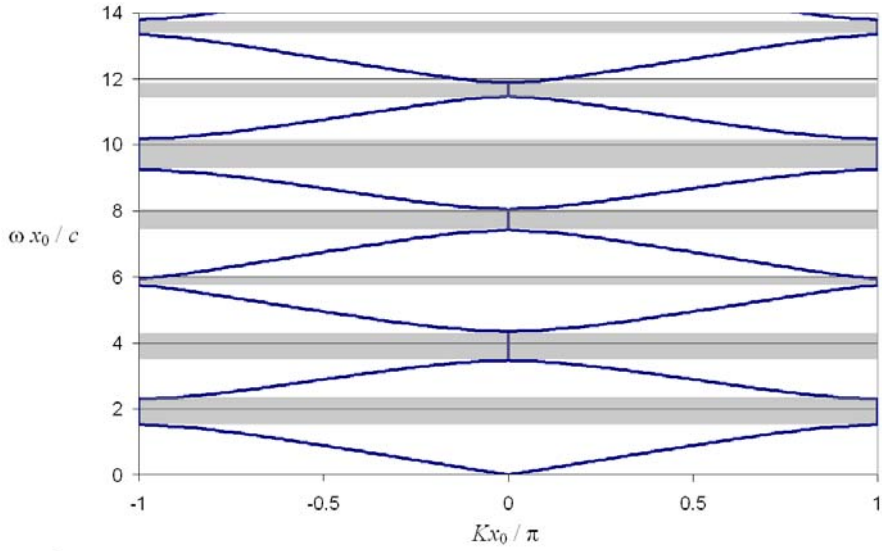


Fig. 7.9. The bandgap structure: frequency vs. Bloch wavenumber. Periodic structure with $d_1 = d_2 = 1$; $\mu_1 = \mu_2 = 1$; $\epsilon_1 = 1$, $\epsilon_2 = 5$. Shaded areas indicate photonic bandgaps.

7.5 Band Structure by Fourier Analysis (Plane Wave Expansion) in 1D

The fundamental matrix that played a central role in Section 7.4 is more important for theoretical analysis than for practical computation, as it contains analytical solutions that may be complicated or unavailable. In particular, the approach cannot be extended to two and three dimensions, where infinitely many independent solutions exist and are usually not available analytically.

Fourier analysis (Plane Wave Expansion, PWE) is the most common practical alternative for analyzing and computing the band structure in any number of dimensions. The 1D case is considered in this section, and 2D–3D computation is taken up later in this chapter.

For simplicity of exposition, let us assume a lossless nonmagnetic periodic medium, where the electric field $E = E_y(x)$ is governed by the wave equation

$$E''(x) \;+\; \omega^2 \mu_0 \epsilon(x) E(x) \;=\; 0 \tag{7.103}$$

Here ϵ is assumed to be a x_0-periodic function. We are looking for a solution in the form of the Bloch–Floquet wave

$$E(x) \;=\; E_{\mathrm{PER}}(x) \exp(-\mathrm{i} K_B x) \tag{7.104}$$

where $E_{\mathrm{PER}}(x)$ is a x_0-periodic function and K_B is the Bloch wavenumber. Both $E_{\mathrm{PER}}(x)$ and K_B are *a priori* unknown and need to be determined.

In Fourier space, $E_{\mathrm{PER}}(x)$ is given by its Fourier series with coefficients e_m $(m = 0, \pm 1, \pm 2, \ldots)$

$$E(x) \;=\; \sum_{m=-\infty}^{\infty} e_m \exp(\mathrm{i} m \kappa_0 x) \exp(-\mathrm{i} K_B x), \qquad \kappa_0 = \frac{2\pi}{x_0} \tag{7.105}$$

Similarly, ϵ is expressed via a Fourier series with coefficients ϵ_m:

$$\epsilon(x) \;=\; \sum_{m=-\infty}^{\infty} \epsilon_m \exp(\mathrm{i} m \kappa_0 x) \tag{7.106}$$

The Fourier coefficients e_m are given by the usual integral expressions

$$e_m \;=\; x_0^{-1} \int_{x_0} E_{\mathrm{PER}}(x) \exp(-\mathrm{i} m \kappa_0 x)\, dx \tag{7.107}$$

where the integration is over any period of length x_0.

Now we are in a position to Fourier-transform the wave equation (7.103). In Fourier space, multiplication $\epsilon(x)E(x)$ (i.e. multiplication of the Fourier series (7.105) and (7.106)) turns into convolution and the problem becomes

$$\mathcal{K}^2 \underline{e} \;=\; \omega^2 \mu_0\, \Xi \underline{e} \tag{7.108}$$

Here $\underline{e} = (\ldots, e_{-2}, e_{-1}, e_0, e_1, e_2, \ldots)^T$ is the (infinite) column vector of Fourier coefficients of the field; $\mathcal{K}$ is an infinite diagonal matrix with the entries $k_m = K_B - \kappa_0 m$, or equivalently

$$\mathcal{K} \;=\; K_B I \;-\; \kappa_0 N, \tag{7.109}$$

where I is the identity matrix and

$$N \;=\; \begin{pmatrix} \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & -2 & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & -1 & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & 0 & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & 1 & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & 2 & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \end{pmatrix} \tag{7.110}$$

Finally, matrix Ξ in (7.108) is composed of the Fourier coefficients of ϵ:

$$\Xi_{ml} = \epsilon_{m-l} \tag{7.111}$$

for any row m and column l $(-\infty < m, l < \infty)$.

The infinite-dimensional eigenproblem (7.108) must in practice be truncated to a finite number of harmonics. The computational trade-off is clear: as the number of harmonics grows, both computational complexity and accuracy increase.

Example 27. Volume grating. This problem is briefly stated in L.I. Mandelshtam's paper [Man45] and will be of even greater interest to us in the context of backward waves and negative refraction (Section 7.13). Consider a volume grating characterized by a sinusoidally changing permittivity of the form $\epsilon(x) = \epsilon_1 + \epsilon_2 \cos(2\pi x/x_0)$, with some parameters $\epsilon_1 > \epsilon_2 > 0$, $x_0 > 0$.

As a numerical example, let $\epsilon_1 = 2$, $\epsilon_2 = 1$, $x_0 = 1$, so that the permittivity and its Fourier decomposition are

$$\epsilon(x) \;=\; 2 + \cos 2\pi x \;=\; 2 + \frac{1}{2}\,\exp(\mathrm{i}2\pi x) + \frac{1}{2}\,\exp(-\mathrm{i}2\pi x)$$

Thus ϵ has only three nonzero Fourier coefficients: $\epsilon_{\pm 1} = 1/2$, $\epsilon_0 = 2$. (The permittivity of free space is not used in this example, so there should be no confusion with the Fourier coefficient ϵ_0.)

The eigenvalue problem (7.108), with the magnetic permeability normalized to unity for simplicity, is

$$\mathcal{K}^2 \underline{e} \;=\; \omega^2 \Xi\, \underline{e} \tag{7.112}$$

The diagonal matrix $\mathcal{K}^2$ has entries

$$\mathcal{K}^2_m \;=\; (K_B - 2\pi m)^2, \qquad m = 0, \pm 1, \pm 2, \ldots$$

and matrix Ξ is tridiagonal, with the entries in the m-th row equal to

$$\Xi_{m,m} = \epsilon_0 = 2; \quad \Xi_{m\pm1,m} = \epsilon_{\pm1} = \frac{1}{2}$$

For any given value of the Bloch parameter K_B, numerical solution can be obtained by truncating the infinite system to the algebraic eigenvalue problem with $2M+1$ equations ($m = -M, -M+1, \ldots M-1, M$).

The first four dispersion curves $\omega(K_B)$ are shown in Fig. 7.10; there are two frequency bandgaps in the figure, approximately [1.98, 2.55] and [4.40, 4.68], and infinitely many more gaps beyond the range of the chart. The numerical results are plotted for 41 equally spaced values of the normalized Bloch number $K_B x_0/\pi$ in $[-1, 1]$. There is no appreciable difference between the numerical results for $M = 5$ (11 equations) and $M = 20$ (41 equations). The high accuracy of the eigenfrequencies for a small number of plane waves in the expansion is due to the smooth variation of the permittivity. Discontinuities in ϵ would require a much higher number of harmonics (Section 7.9.3).

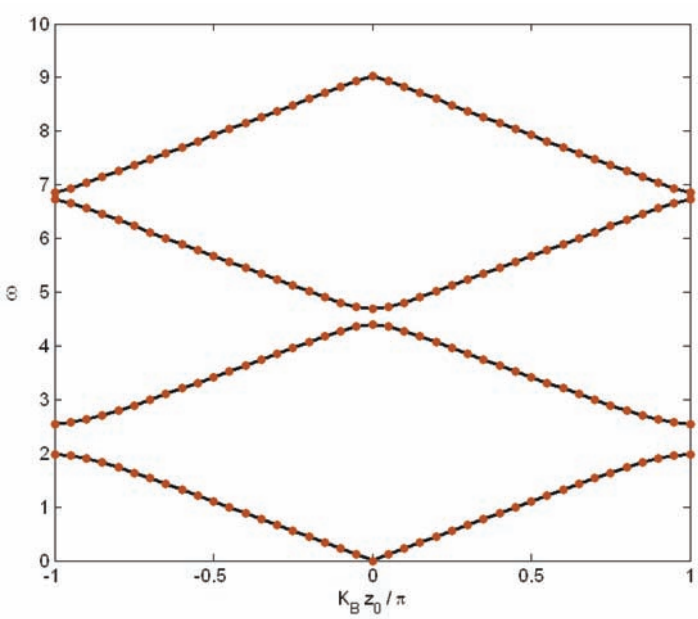


Fig. 7.10. The bandgap structure for the volume grating with $\epsilon(x) = 2 + \cos 2\pi x$. Solid line – $M = 5$ ($2 \times 5 + 1 = 11$ plane waves); circles – $M = 20$ ($2 \times 20 + 1 = 41$ plane waves).

In addition to the eigenvalues ω^2 of (7.112), the eigenvectors $\underline{e}$ are also of interest. As an example, let us set $K_B x_0 = \pi/10$. Stem plots of the four

eigenvectors corresponding to the four smallest eigenvalues $\omega^2 \approx 0.049$, 18.29, 23.12 and 77.83, are shown in Fig. 7.11. The first Bloch wave in Fig. 7.11(a) is almost a plane wave; the amplitudes of all harmonics other than e_0 are very small (but not zero, as it might appear from the figure); for example, $e_{-1} \approx 0.00057$, $e_1 \approx 0.00069$.

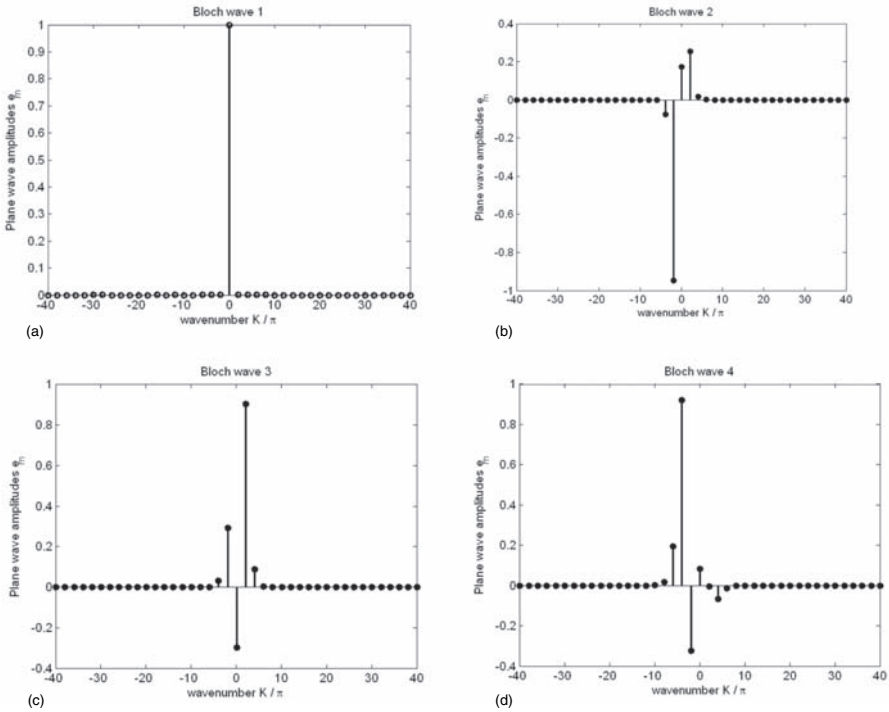


Fig. 7.11. The amplitudes of the plane wave components of the first four Bloch waves (a)–(d) for the volume grating with $\epsilon(x) = 2 + \cos 2\pi x$. Solution with 41 plane waves. $K_B x_0 = \pi/10$.

It is interesting to note that dispersion curves with positive and negative slopes $\partial\omega/\partial K_B$ (i.e. positive and negative group velocity) alternate in the diagram. Group velocity is positive for the lowest-frequency curve $\omega_1(K_B)$, negative for $\omega_2(K_B)$, positive again for $\omega_3(K_B)$, etc. This interesting issue will be further discussed in the context of backward waves and negative refraction (p. 461).

7.6 Characteristics of Bloch Waves

7.6.1 Fourier Harmonics of Bloch Waves

For analysis and physical interpretation of the properties of Bloch waves (7.60) – in particular, energy flow and the meaning of phase velocity – it is convenient to view these waves as a suite of (spatial) Fourier harmonics. The ideas are most easily explained in the 1D case but will be extended to 2D and 3D in subsequent sections. A very helpful reference is the paper by B. Lombardet *et al.* [LDFH05].

Consider one more time the Bloch wave

$$E(x) \;=\; E_{\mathrm{PER}}(x)\exp(-\mathrm{i}K_B x) \tag{7.113}$$

As before, subscript "PER" indicates a spatially periodic function with a given period x_0. Expressing this periodic function via its Fourier series, one obtains

$$E(x) \;=\; \sum_{m=-\infty}^{\infty} e_m \exp(\mathrm{i}m\kappa_0 x)\exp(-\mathrm{i}K_B x), \quad \kappa_0 = 2\pi x_0^{-1} \tag{7.114}$$

The Fourier decomposition (7.114) of $E(x)$ has a clear physical interpretation as a superposition of plane waves E_m:

$$E(x) \;=\; \sum_{m=-\infty}^{\infty} E_m(x), \quad E_m(x) \equiv e_m \exp(-\mathrm{i}k_m x), \quad k_m \equiv K_B - m\kappa_0 \tag{7.115}$$

Let us assume $\mu = \text{const}$, as the analysis in this important practical case simplifies. At optical frequencies, one may assume $\mu = \mu_0$ (L.D. Landau and E.M. Lifshitz [LL84], §60).[8] Then the above expression for $E(x)$ leads, via the Maxwell $\nabla \times \mathbf{E}$ equation, to a similar decomposition of the magnetic field $H \equiv H_z$:

$$\begin{aligned} H(x) \;&=\; -\frac{1}{\mathrm{i}\omega\mu}\frac{\partial E}{\partial x} \;=\; -\frac{1}{\mathrm{i}\omega\mu}\sum_{m=-\infty}^{\infty}(-\mathrm{i}k_m)\, e_m \exp(-\mathrm{i}k_m x) \\ &=\; \sum_{m=-\infty}^{\infty}\frac{k_m}{\omega\mu}\, e_m \exp(-\mathrm{i}k_m x) \end{aligned} \tag{7.116}$$

It is important to note from the outset, as Lombardet *et al.* do in [LDFH05], that the individual plane-wave components of the electromagnetic Bloch wave do not satisfy Maxwell's equations in the periodic medium and therefore do not represent physical fields. Only taken together do these Fourier harmonics form a valid electromagnetic field.

[8] Artificial magnetism can be created in periodic dielectric structures at optical frequencies (Section 7.13, W. Cai *et al.* [CCY+07], S. Linden *et al.* [LED+06]). The equivalent "mesoscopic" permeability may then be different from μ_0, but the intrinsic *microscopic* permeability of the materials involved is still μ_0.

7.6.2 Fourier Harmonics and the Poynting Vector

Consider now the Fourier decomposition of the time-averaged Poynting vector (power flow) $\mathbf{P} = \mathrm{Re}\{\mathbf{E} \times \mathbf{H}^*\}/2$. In the 1D case this vector has only one component $P = P_x$

$$P(x) \;=\; \frac{1}{2}\,\mathrm{Re}\{E(x)H^*(x)\} \tag{7.117}$$

In Fourier space, the product EH^* turns into convolution-like summation. The expression simplifies for lossless materials (ϵ real) because then the Poynting vector must be constant and pointwise values $P(x)$ are obviously equal to the spatial average $\langle P \rangle$. This average value over one period of the structure is easy to find due to the orthogonality of Bloch harmonics $\psi_m = \exp(-\mathrm{i}k_m x)$ ($k_m = K_B - m\kappa_0$):

$$\begin{aligned}(\psi_m,\, \psi_l) \equiv \int_{x_0} \psi_m \psi_l^*\, dx \;&=\; \int_{x_0} \exp(-\mathrm{i}k_m x)\exp(\mathrm{i}k_l x)\, dx \\ &=\; \int_{x_0} \exp[\mathrm{i}(l-m)\kappa_0 x]\, dx \;=\; 0\end{aligned}$$

The last equality represents orthogonality of the standard Fourier harmonics over one period. The Bloch harmonics have the same property because the $\exp(-\mathrm{i}K_B x)$ factor in one term of the integrand is canceled by the $\exp(+\mathrm{i}K_B x)$ factor in the other, complex conjugate, term. (This is true for lossless media when the Bloch wavenumber K_B is purely real.)

Parseval's theorem then allows us to rewrite the Poynting vector of the Bloch wave (7.117), in the lossless case, as the sum of the the Poynting vectors of the individual plane waves:

$$P \;=\; \sum_{m=-\infty}^{\infty} P_m; \qquad P_m \;=\; \frac{k_m}{2\omega\mu}\,|e_m|^2, \qquad m = 0, \pm 1, \pm 2, \ldots \tag{7.118}$$

In 2D and 3D, an analogous identity holds true for the *time-space averaged* Poynting vector (B. Lombardet *et al.* [LDFH05]) – again, due to the orthogonality of the Fourier harmonics. In 1D, the Poynting vector is constant and hence the spatial averaging is redundant.

7.6.3 Bloch Waves and Group Velocity

For the same reason as in homogeneous media (Section 7.3.3, p. 358), one may anticipate a connection between the Poynting vector, group and energy velocities of Bloch waves. The Poynting vector and group velocity are associated with energy flow and signal (information) transfer, respectively.

One can define group velocity in essentially the same way as for waves in homogeneous media:

$$v_g \;=\; \frac{\partial \omega}{\partial K_B} \tag{7.119}$$

K_B being the Bloch wavenumber. Recall that K_B generates a whole "comb" of wavenumbers $K_B - m\kappa_0$, where m is an arbitrary integer and $\kappa_0 = 2\pi/x_0$. Since any two numbers in the comb differ by a constant independent of K_B, differentiation in (7.119) can in fact be performed with respect to any of the comb values $K_B - m\kappa_0$. Loosely speaking, the group velocities of all plane wave components of the Bloch wave are the same. ("Loosely" – because these components do not exist separately as valid physical waves in the periodic medium, and therefore their group velocities are mathematical but arguably not physical quantities.)

To see that this definition of group velocity bears more than superficial similarity to the same notion for homogeneous media, we need to demonstrate that v_g in (7.119) is in fact related to signal velocity. To this end, let us follow the analysis in Section 7.3.2 on p. 355. We shall again consider, as a characteristic case, a pointwise source that produces amplitude modulation with a low-frequency waveform $\mathcal{E}(0,t)$ at $x = 0$ (7.31):

$$E(0,t) \;=\; \mathcal{E}(0,t)\exp(\mathrm{i}\omega_0 t) \tag{7.120}$$

In a homogeneous medium, each frequency component of this source gives rise to a plane wave, which leads to expression (7.33) (p. 356) for the field at an arbitrary location $x > 0$. In the periodic medium, plane waves are replaced with Bloch waves, so that in lieu of (7.33) one has

$$E(x,t) \;=\; \int_{-\infty}^{\infty} \hat{\mathcal{E}}(0,\omega-\omega_0)\, E_{\mathrm{PER}}(x,\omega)\exp[-\mathrm{i}K_B(\omega)x]\exp(\mathrm{i}\omega t)\, d\omega \tag{7.121}$$

where $E_{\mathrm{PER}}(x,\omega)$ is the space-periodic factor in the Bloch wave normalized for convenience to unity at $x = 0$. Of the two possible Bloch waves, equation (7.121) contains the one with the Poynting vector (energy flow) in the $+x$-direction. The respective low-frequency "signal" $\mathcal{E}(x,t)$ is

$$\mathcal{E}(x,t) = E(x,t)\exp(-\mathrm{i}\omega_0 t) = \int_{-\infty}^{\infty} \hat{\mathcal{E}}(0,\omega')\, E_{\mathrm{PER}}(x,\omega')\, \exp[-\mathrm{i}K_B(\omega')x]\, d\omega' \tag{7.122}$$

with

$$\omega' \;\equiv\; \omega - \omega_0$$

The velocity of this signal can again be found by setting the differential $d\mathcal{E}(x,t)$ to zero. This velocity is the ratio of partial differentials of $\mathcal{E}(x,t)$ with respect to t and x. For homogeneous media, these partial derivatives are given by expressions (7.35) and (7.36) on p. 357. For Bloch waves, due to the dependence of E_{PER} on x, the x-derivative acquires an additional (and unwanted) term

$$\int_{-\infty}^{\infty} \hat{\mathcal{E}}(0,\omega')\, \frac{\partial E_{\mathrm{PER}}(x,\omega')}{\partial x}\, \exp[-\mathrm{i}K_B(\omega')x]\, d\omega'$$

This field contains rapidly oscillating spatial components:

$$\frac{\partial E_{\mathrm{PER}}(x,\omega')}{\partial x} \;=\; \mathrm{i}\kappa_0 \sum_{m=-\infty}^{\infty} e_m m \exp(\mathrm{i}m\kappa_0 x)$$

A useful "macroscale" signal can be defined in a natural way as the average of this field over the lattice cell. For the m-th spatial harmonic this average is

$$x_0^{-1} \int_{x_0} \frac{\partial}{\partial x} \left[e_m \exp(\mathrm{i}m\kappa_0 x)\right] \exp(-\mathrm{i}K_B x)\, dx \;=\; e_m \kappa_0 \frac{\exp(-\mathrm{i}K_B x_0) - 1}{2\pi - K_B x_0/m}$$

This term is small *under the additional constraint* $K_B x_0 \ll 1$ – that is, if the Bloch wavelength $2\pi/K_B$ is much greater than the lattice size x_0. In that case, the analysis on p. 357 remains essentially unchanged and leads to the familiar expression for group velocity (7.119). Other reservations discussed on p. 357 in connection with signal velocity (7.37) must also be borne in mind.

7.6.4 Energy Velocity for Bloch Waves

This section shows that group velocity, as defined in (7.119), is equal to energy velocity for lossless nonmagnetic periodic media without dispersion. An alternative proof, but with a heavy dose of vector calculus, can be found in P. Yeh's paper [Yeh79] (1979).

This section builds up on the material of Section 7.5 (p. 375). The familiar equation for the electric field $E = E_y(x)$ in 1D is reproduced here for convenience:

$$E''(x) \;+\; \omega^2 \mu_0 \epsilon(x) E(x) \;=\; 0 \tag{7.123}$$

where ϵ is an x_0-periodic function. If $E(x)$ is a Bloch–Floquet wave, i.e. it satisfies the scaled-periodic boundary conditions with the Bloch factor $\exp(-\mathrm{i}K_B x_0)$ over the spatial period, an essential energy identity can be obtained from (7.123) by inner-multiplication with E and integration by parts:

$$\frac{1}{\omega^2 \mu_0} (E',\, E') \;=\; (\epsilon E,\, E)$$

The boundary terms in the integration by parts have canceled due to the boundary conditions. Now, from Maxwell's equations, $H = E'/(-\mathrm{i}\omega\mu_0)$, and therefore

$$(\mu_0 H,\, H) \;=\; (\epsilon E,\, E) \tag{7.124}$$

That is, the spatial averages of quasi-static magnetic and electric energies of the Bloch wave are equal. Note, however, that for dispersive media these quasi-static values constitute only part of the full electromagnetic energy; see equation (7.44) on p. 359.

In Fourier space, the eigenproblem given by (7.108)

$$\mathcal{K}^2 \underline{e} \;=\; \omega^2 \mu_0 \Xi \underline{e}$$

forms a basis for the plane wave method. For notation and details, see Section 7.5.

It will be convenient to rewrite the eigenvalue problem in the Galerkin form by inner-multiplying the equation with an arbitrary vector[9] $\underline{e}'$:

$$(\mathcal{K}^2\underline{e},\, \underline{e}') \;=\; \omega^2\mu_0\,(\Xi\underline{e},\, \underline{e}') \tag{7.125}$$

To find the group velocity, we write the variation of this Galerkin equation for a small change δK_B and the respective variation $\delta\omega^2$. The eigenvector $\underline{e}$ also depends on K_B and ω and is also subject to the variation. However, *the variation of $\underline{e}$ is irrelevant for the analysis.*

Indeed, in the eigenvalue problem one may scale the eigenvector arbitrarily. A convenient normalization is (for $K_B \neq 0$)

$$(\mathcal{K}^2\underline{e},\, \underline{e}) \;=\; 1$$

and concomitantly

$$(\Xi\underline{e},\, \underline{e}) \;=\; \frac{1}{\omega^2\mu_0} = \text{const}$$

This implies that the variation $\delta\underline{e}$ is $\mathcal{K}^2$- and Ξ-orthogonal to $\underline{e}$:

$$(\mathcal{K}^2\underline{e},\, \delta\underline{e}) \;=\; (\Xi\underline{e},\, \delta\underline{e}) \;=\; 0$$

This generalized orthogonality eliminates all (first-order) terms with $\delta\underline{e}$ in the variation of the Galerkin equation (7.125). This variation, then, for $\underline{e}' = \underline{e}$ is

$$2\kappa_0\,\delta K_B\,(N\underline{e},\, \underline{e}) \;=\; \delta\omega^2\,\mu_0\,(\Xi\underline{e},\, \underline{e}) \tag{7.126}$$

Now we can examine the expression for the group velocity:

$$v_g \;=\; \frac{\partial\omega}{\partial K_B} \;=\; \frac{\partial\omega^2}{2\omega\,\partial K_B} \;=\; \frac{\kappa_0\,(N\underline{e},\, \underline{e})}{\omega\mu_0\,(\Xi\underline{e},\, \underline{e})} \tag{7.127}$$

What remains to be done is to link the numerator of this expression to the Poynting vector and the denominator to the energy of the field. For the spatial average of the Poynting vector we have

$$\begin{aligned}\langle P\rangle \;&=\; \frac{1}{2}\,\mathrm{Re}\,x_0^{-1}\int_{x_0} EH^*dx \;=\; \frac{1}{2}\,\mathrm{Re}\,x_0^{-1}\int_{x_0} E\,\frac{1}{\mathrm{i}\omega\mu_0}E^{*\prime}dx \\ &=\; \frac{1}{2}\,\mathrm{Re}\,\frac{1}{\mathrm{i}\omega\mu_0 x_0}\int_{x_0} EE^{*\prime}dx = \;\frac{1}{2}\,\mathrm{Re}\,\frac{1}{\omega\mu_0}\,(\kappa_0 N\underline{e},\, \underline{e})\end{aligned}$$

The last equality follows from Plancherel's theorem; we also used the fact that differentiation of the m-th harmonic translates into multiplication with

[9] All vectors are infinite-dimensional, and it is tacitly assumed that their components decay rapidly enough, so that all infinite algebraic sums make mathematical sense.

$\mathrm{i}\kappa_0 m$ in Fourier space. The Bloch exponentials have again canceled out in the products of complex variables with their conjugates.

This connects the time-space averaged Poynting vector with the numerator of (7.127). For the denominator, Plancherel's Theorem gives

$$(\Xi \underline{e}, \underline{e}) \;=\; x_0^{-1} \int_{x_0} \epsilon |E|^2 \, dx$$

which is proportional to the (quasi-static) energy of the electric field.

Putting the numerator and denominator together and noting that the electric and magnetic energies in the non-dispersive case are equal due to (7.124), one obtains the final result similar to the one for a dispersive but homogeneous medium (7.43), p. 359:

$$v_g \;=\; \frac{\langle P \rangle}{\langle W \rangle} \;\equiv\; v_E \tag{7.128}$$

where $\langle W \rangle$ is the average electromagnetic energy of the Bloch wave in a lossless medium without dispersion. The physical interpretation of this identity is that energy is transferred through the periodic medium with group velocity.

7.7 Two-Dimensional Problems of Wave Propagation

Time-harmonic Maxwell's equations simplify significantly if the fields do not depend on one of the Cartesian coordinates – say, on z – and if there is no coupling in the material parameters between that coordinate and the other two (i.e. $\epsilon_{xz} = 0$, etc.) Upon writing out field equations (7.15) and (7.16) in Cartesian coordinates, one observes that they break up into two decoupled systems. The first system involves E_z, H_x and H_y and for isotropic materials (scalar $\epsilon = \epsilon(x, y)$, $\mu = \mu(x, y)$) has the form

$$\partial_y E_z \;=\; -\mathrm{i}\omega\mu H_x \tag{7.129}$$

$$-\partial_x E_z \;=\; -\mathrm{i}\omega\mu H_y \tag{7.130}$$

$$\partial_x H_y \;-\; \partial_y H_x \;=\; \mathrm{i}\omega\epsilon E_z \tag{7.131}$$

It is well known that the magnetic field can be eliminated from this set of equations, with the Helmholtz equation resulting for E_z. Indeed, multiplying the first two equations by μ^{-1} and differentiating, we get

$$\partial_y(\mu^{-1}\partial_y E_z) \;=\; -\mathrm{i}\omega\partial_y H_x \tag{7.132}$$

$$-\partial_x(\mu^{-1}\partial_x E_z) \;=\; -\mathrm{i}\omega\partial_x H_y \tag{7.133}$$

The difference of these two equations, with (7.131) in mind, leads to

$$\nabla \cdot (\mu^{-1}\nabla E_z) \;+\; \omega^2\epsilon E_z \;=\; 0 \tag{7.134}$$

In the special but important case of constant μ, this becomes

$$\nabla^2 E_z + k^2 E_z = 0, \quad \text{with } k^2 = \omega^2 \mu\epsilon, \quad \mu = \text{const} \tag{7.135}$$

The complementary equation for the triple H_z, E_x and E_y is, quite analogously,

$$\nabla \cdot (\epsilon^{-1} \nabla H_z) + \omega^2 \mu H_z = 0 \tag{7.136}$$

which for constant ϵ simplifies to

$$\nabla^2 H_z + k^2 H_z = 0, \quad \text{with } k^2 = \omega^2 \mu\epsilon, \quad \epsilon = \text{const} \tag{7.137}$$

The two decoupled solutions (E_z, H_x, H_y) and (H_z, E_x, E_y) are called TE and TM modes, respectively. Or rather, TM and TE modes, respectively.

There is regrettable ambiguity in the terminology used by different engineering and research communities. The "T" in "TE" and "TM" stands for "transverse," meaning, according to the dictionary definition, "in a crosswise direction; at right angles to the long axis". So, the electric field in a TE mode and the magnetic field in a TM mode are transverse... to what? In waveguide applications, they are transverse to the longitudinal axis of the guide; a TM mode in the guide thus *lacks* the H_z component of the magnetic field and is described by equation (7.135) for the E-field.[10] However, for 2D-periodic structures in photonics applications (photonic crystals), the same equation (7.135) describes the electric field that is "transverse" to the cross-section of the crystal and therefore some authors call it a *TE* mode. Others refer to the same field as a TM mode by analogy with waveguides.

Thus the E-field equation may wind up identifying either a TE or TM mode, depending on the application and one's point of view. Table 7.1 illustrates the terminological differences.

Only one E-component present	One E-component *absent*	Only one H-component present
I.V. Shadrivov *et al.* [SSK05]; T. Fujisawa & M. Koshiba [FK04]; A. Ishimaru *et al.* [ITJ05]	J.A. Stratton [Str41]; R.S. Elliott [Ell93]; R.F. Harrington [Har01]; A.F. Peterson, S.L. Ray & R. Mittra [PRM98]	G. Shvets & Y.A. Urzhumov [SU04]; S.G. Johnson & J.D. Joannopoulos [JJ01]; S. Yamada *et al.* [YWK+02]; R. Meisels *et al.* [MGKH06]

Table 7.1. Definitions of the TE mode may differ.

[10] In waveguides, even though some field components may be zero, the fields in general depend on all three coordinates, and hence the Laplacian operator in field equations should be interpreted as $\nabla^2 = \partial_x^2 + \partial_y^2 + \partial_z^2$. If the field does not depend on z, as in many 2D problems in photonics, the z-derivative in the Laplacian disappears.

Furthermore, in optics the waves with only one component of the electric field (perpendicular to the plane of incidence) are referred to as s-waves (or s-polarized); waves with only one H-component are p-waves.

From the computational (as well as analytical) perspective, fields with only one Cartesian component are of particular interest, as equations for these fields are scalar and thus much easier to deal with than the more general vector equations. With this in mind, in the remainder of this chapter I shall simply call waves with one E component E-waves (or E-modes); H-waves have a similar definition. It is hoped that the reader will find this convention straightforward and unambiguous.

7.8 Photonic Bandgap in Two Dimensions

In 2D and especially in 3D periodic structures, the bandgap phenomenon is much richer, and more difficult to analyze, than in 1D (Section 7.4). The Bloch wavenumber, scalar in 1D, becomes a wave *vector* in 2D and 3D, as the Bloch–Floquet wave can travel in different directions. Moreover, electromagnetic wave propagation in general depends on *polarization* – i.e. on the direction of the $\mathbf{E}$ vector in the wave; this adds one more degree of freedom to the analysis.

For each direction of propagation and for each polarization, there may exist a forbidden frequency range – a bandgap – where the corresponding Bloch wavenumber K_B is imaginary and hence no propagating modes exist. If these bandgaps happen to overlap for all directions of propagation and for both polarizations, so that no Bloch waves can travel in any direction, a *complete* bandgap is said to exist.

Let us consider a photonic crystal example that is general enough to contain many essential features of the two-dimensional problem. A square cell of the crystal, of size $a \times a$, contains a dielectric rod with radius r_{rod} and the relative dielectric permittivity ϵ_{rod} (Fig. 7.12). The medium outside the rod has permittivity ϵ_{out}. All media are nonmagnetic. The crystal lattice is obtained by periodically replicating the cell infinitely many times in both coordinate directions.

In the Fourier space of Bloch vectors $\mathbf{K}$, the corresponding "master" cell – called the first *Brillouin zone*[11] – is $[-\pi/a, \pi/a] \times [-\pi/a, \pi/a]$ (Fig. 7.13). This zone can also be periodically replicated infinitely many times in both K_x and K_y directions to produce a reciprocal (i.e. Fourier space) lattice. However, all possible Bloch waves $E_{\mathrm{PER}} \exp(-\mathrm{i}\mathbf{K} \cdot \mathbf{r})$ are already accounted for in the first Brillouin zone. Indeed, adding $2\pi/a$ to, say, K_x introduces just a periodic factor $\exp(-\mathrm{i}2\pi x/a)$, with period a, that can as well be "absorbed" into the periodic Bloch component $E_{\mathrm{PER}}(x, y)$.

Standard notation for some special points in the first Brillouin zone is shown in Fig. 7.13. The Γ point is $\mathbf{K} = 0$; the X point is $\mathbf{K} = [\pi/a, 0]$; the M

[11] Léon N. Brillouin, 1889–1969, an outstanding French and American physicist.

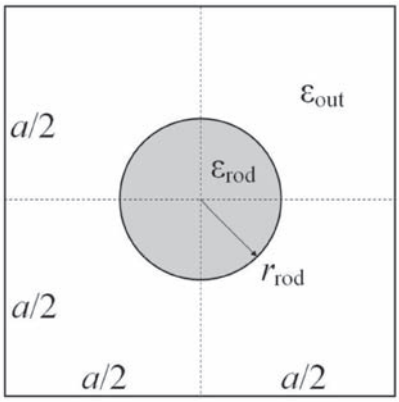


Fig. 7.12. A square cell of a photonic crystal lattice. The (infinite) crystal is an array of dielectric rods obtained by periodic replication of the cell in both coordinate directions.

point is $\mathbf{K} = [\pi/a, \pi/a]$; Δ is a generic point on ΓX (i.e. with $K_y = 0$); and Σ is a generic point on ΓM.

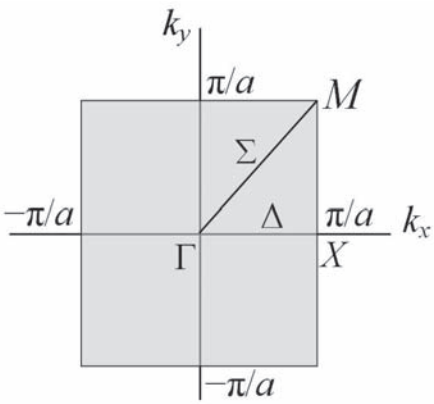


Fig. 7.13. The first Brillouin zone for the square photonic crystal lattice.

The problem can now be formulated as follows. First, the E-mode (one component of the electric field $E = E_z$) is described by equation (7.135), repeated here for easy reference:

$$\nabla^2 E + \omega^2 \mu \epsilon E = 0, \quad \text{for } \mu = \text{const} \tag{7.138}$$

where the E-field is sought as a Bloch wave with a (yet undetermined) wave vector $\mathbf{K}$:

$$E(\mathbf{r}) \;=\; E_{\mathrm{PER}}(\mathbf{r})\,\exp(-\mathrm{i}\mathbf{K}\cdot\mathbf{r}); \qquad \mathbf{r}\equiv(x,y), \qquad \mathbf{K}=(K_x,K_y) \tag{7.139}$$

There are two general options: solving for the full E-field of (7.138) or, alternatively, for the periodic factor $E_{\mathrm{PER}}(x,y)$. In the first case, the governing equation is fairly simple (Helmholtz) but the boundary conditions are non-standard due to the Bloch exponential $\exp(-\mathrm{i}\mathbf{K}\cdot\mathbf{r})$ (details below). In the second case, with E_{PER} as the unknown, standard periodic boundary conditions apply, but the differential operator is more complicated.

More precisely, the problem for the full E-field includes the Helmholtz equation (7.138) in the square $[-a/2,a/2]\times[-a/2,a/2]$ and the "scaled-periodic" boundary condition

$$E\left(\frac{a}{2},\,y\right) \;=\; \exp(-\mathrm{i}K_x a)\,E\left(-\frac{a}{2},\,y\right); \quad -\frac{a}{2}\le y\le\frac{a}{2} \tag{7.140}$$

$$E\left(x,\,\frac{a}{2}\right) \;=\; \exp(-\mathrm{i}K_y a)\,E\left(x,\,-\frac{a}{2}\right); \quad -\frac{a}{2}\le x\le\frac{a}{2} \tag{7.141}$$

In the alternative formulation, with E_{PER} as the main unknown, the Helmholtz equation takes on a different form because

$$\nabla\left(E_{\mathrm{PER}}(\mathbf{r})\,\exp(-\mathrm{i}\mathbf{K}\cdot\mathbf{r})\right) \;=\; \left(\nabla E_{\mathrm{PER}} - \mathrm{i}\mathbf{K}E_{\mathrm{PER}}\right)\exp(-\mathrm{i}\mathbf{K}\cdot\mathbf{r}) \tag{7.142}$$

Formally, the ∇ operator acting on E is replaced with the $\nabla-\mathrm{i}\mathbf{K}$ operator acting on E_{PER}. Similarly, applying the divergence operator to (7.142), one obtains the Laplacian

$$\begin{aligned}\nabla^2 E \;&=\; [(\nabla-\mathrm{i}\mathbf{K})\cdot(\nabla-\mathrm{i}\mathbf{K})E_{\mathrm{PER}}]\,\exp(-\mathrm{i}\mathbf{K}\cdot\mathbf{r})\\ &=\; [\nabla^2 E_{\mathrm{PER}} - 2\mathrm{i}\mathbf{K}\cdot\nabla E_{\mathrm{PER}} - K^2 E_{\mathrm{PER}}]\,\exp(-\mathrm{i}\mathbf{K}\cdot\mathbf{r})\end{aligned} \tag{7.143}$$

The Bloch–Floquet eigenvalue problem for E_{PER} thus becomes (after canceling the common complex exponential in all terms)

$$-\nabla^2 E_{\mathrm{PER}} + 2\mathrm{i}\mathbf{K}\cdot\nabla E_{\mathrm{PER}} + K^2 E_{\mathrm{PER}} \;=\; \omega^2\mu\epsilon E_{\mathrm{PER}} \tag{7.144}$$

with the periodic boundary conditions

$$E_{\mathrm{PER}}\left(\frac{a}{2},\,y\right) \;=\; E_{\mathrm{PER}}\left(-\frac{a}{2},\,y\right); \quad -\frac{a}{2}\le y\le\frac{a}{2} \tag{7.145}$$

$$E_{\mathrm{PER}}\left(x,\,\frac{a}{2}\right) \;=\; E_{\mathrm{PER}}\left(x,\,-\frac{a}{2}\right); \quad -\frac{a}{2}\le x\le\frac{a}{2} \tag{7.146}$$

The dielectric permittivity in (7.144) is a function of position. In principle, the magnetic permeability may also depend on coordinates, but this is not the case in our present example or at optical frequencies in general.

Both eigenvalue problems (in terms of E or, alternatively, E_{PER}) are unusual, as they have three (and in the 3D case four) scalar eigenparameters: frequency ω and the components K_x, K_y of the Bloch vector. Solving for

all three parameters, and the respective eigenmodes, simultaneously is impractical. The usual approach is to fix the $\mathbf{K}$ vector and solve the resultant eigenvalue problem for ω only; then repeat the computation for a set of values of $\mathbf{K}$.[12] Of most interest are the values on the symmetry lines in the Brillouin zone (Fig. 7.13) $\Gamma \to X \to M \to \Gamma$; eigenfrequencies ω corresponding to these values are typically plotted in a single chart. For the lattice of cylindrical rods, this bandgap structure is computed below.

It is quite interesting to analyze the behavior of Bloch waves in the limiting case of a quasi-homogeneous material, when the lattice cell size tends to zero relative to the wavelength in a vacuum. This will be discussed in Section 7.13.6, in connection with backward waves and negative refraction in metamaterials.

In addition to the two ways of formulating the photonic bandgap problem, there are several approaches to solving it. We shall consider two of them: Finite Element analysis and plane wave expansion (i.e. Fourier transform).

7.9 Band Structure Computation: PWE, FEM and FLAME

7.9.1 Solution by Plane Wave Expansion

As a periodic function of coordinates, factor E_{PER} (7.145), (7.146) can be expanded into a Fourier series with some (yet unknown) coefficients $\tilde{E}_{\mathrm{PER}}(\mathbf{k_m})$,

$$E_{\mathrm{PER}} \;=\; \sum_{\mathbf{m}\in\mathbb{Z}^2} \tilde{E}_{\mathrm{PER}}(\mathbf{k_m}) \exp(\mathrm{i}\mathbf{k_m}\cdot\mathbf{r}), \quad \mathbf{k_m} \;=\; \frac{2\pi}{a}\,\mathbf{m} \;\equiv\; \frac{2\pi}{a}\,(m_x, m_y) \tag{7.147}$$

with integers m_x, m_y. The full field E is obtained by multiplying E_{PER} with the Bloch exponential:

$$E \;=\; E_{\mathrm{PER}} \exp(-\mathrm{i}\mathbf{K}\cdot\mathbf{r}) \;=\; \sum_{\mathbf{m}\in\mathbb{Z}^2} \tilde{E}_{\mathrm{PER}}(\mathbf{m}) \exp\left(\mathrm{i}(\mathbf{k_m}-\mathbf{K})\cdot\mathbf{r}\right) \tag{7.148}$$

The dielectric permittivity $\epsilon(x,y)$ is also a periodic function of coordinates and can be expanded into a similar Fourier series. However, it is often advantageous to deal with the *inverse* of ϵ, $\gamma = \epsilon^{-1}$. The reason is that, after multiplying the governing equation (7.138) through by γ, one arrives at an eigenvalue problem without any coordinate-dependent coefficients in the right hand side:

$$-\gamma(x,y)\,\nabla^2 E \;=\; \omega^2 \mu E \qquad (\mu = \mathrm{const}) \tag{7.149}$$

[12] However, in Flexible Local Approximation MEthods (FLAME, Section 7.9.6) it is ω that acts as an "independent variable" because the basis functions in FLAME depend on it. The Bloch wave vector is computed as a function of frequency. Also, for lossy materials $\mathbf{K}$ is complex, and it may make sense to fix ω and solve for $\mathbf{K}$.

This ultimately leads to a standard eigenvalue problem of the form $Ax = \lambda x$ rather than a more complicated *generalized* problem $Ax = \lambda Bx$. (See also the previous section on FEM, where a generalized eigenproblem arises due to the presence of the FE "mass matrix".) As before, E satisfies the scaled-periodic boundary conditions with the complex exponential Bloch factor.

The downside of the multiplication by γ is that the operator in the left hand side of the eigenvalue problem (7.149) is not self-adjoint. (The coordinate-dependent factor $\gamma(x, y)$ outside the divergence operator gets in the way of the usual integration-by-parts argument for self-adjointness.) The original formulation, $-\nabla^2 E = \omega^2 \mu\, \epsilon(x, y)\, E$, has self-adjoint operators on both sides if the medium is lossless (real ϵ). The choice thus is between a Hermitian but *generalized* eigenvalue problem and a regular but non-Hermitian one.

For the Bloch–Floquet E-field (7.148), the negative of the Laplace operator turns, in the Fourier domain, into multiplication by $|\mathbf{k_m} - \mathbf{K}|^2$. Further, the product $-\gamma \nabla^2 E$ in the left hand side of (7.149) turns into convolution; the $\mathbf{m}$-th Fourier harmonic of this product is

$$\mathcal{F}_{\mathbf{m}}\{-\gamma\, \nabla^2 E\} \;=\; \sum_{\mathbf{s}\in\mathbb{Z}^2} |\mathbf{k_m} - \mathbf{K}|^2\, \tilde{\gamma}(\mathbf{m} - \mathbf{s})\, \tilde{E}_{\mathbf{s}}, \qquad \mathbf{k_m} = \frac{2\pi}{a}\, \mathbf{m} \tag{7.150}$$

where $\tilde{\gamma}$ are the Fourier coefficients for $\gamma = \epsilon^{-1}$:

$$\gamma \;=\; \sum_{\mathbf{m}\in\mathbb{Z}^2} \tilde{\gamma}(\mathbf{m}) \exp(\mathrm{i}\mathbf{k_m} \cdot \mathbf{r}), \tag{7.151}$$

Putting together the left and right hand sides of equation (7.149) in the Fourier domain, we obtain an eigenvalue problem for the Fourier coefficients:

$$\sum_{\mathbf{s}\in\mathbb{Z}^2} |\mathbf{k_m} - \mathbf{K}|^2 \tilde{\gamma}(\mathbf{m} - \mathbf{s})\, \tilde{E}(\mathbf{s}) \;=\; \omega^2 \mu \tilde{E}(\mathbf{m}); \tag{7.152}$$

$$\mathbf{m} = (m_x, m_y); \quad m_x, m_y = 0, \pm 1, \pm 2, \ldots$$

This is an infinite set of equations for the eigenfrequencies and eigenmodes. For computational purposes, the system needs to be truncated to a finite size; this size is an adjustable parameter in the computation.

Numerical results for a cylindrical rod lattice are presented in Sections 7.9.4 and 7.13.5 (p. 468).

7.9.2 The Role of Polarization

To avoid repetition, we have so far considered E-polarization only, with the corresponding equation (7.149) for the one-component E field. The problem for H-polarization is very similar:

$$-\nabla \cdot (\gamma(x, y) \nabla H) \;=\; \omega^2 \mu H \tag{7.153}$$

but its algebraic properties are better. Namely, the operator in the left hand side of (7.153), unlike the operator for the E-problem (7.149), is self-adjoint and nonnegative definite (which is easy to show using integration by parts and taking into account Remark 25 on boundary conditions, p. 394).

This unequal status of the E- and H-problems is due to the assumption that all materials are nonmagnetic. If this is not the case and μ depends on coordinates, the E- and H-problems are fully analogous.

7.9.3 Accuracy of the Fourier Expansion

The main factor limiting the accuracy of the plane wave solution is the Fourier approximation of the dielectric permittivity $\epsilon(x, y)$ or, alternatively, its inverse $\gamma(x, y)$. Abrupt changes in the dielectric constant lead in its Fourier representation to the ringing effect (the "Gibbs phenomenon," well known in Fourier analysis).

For illustration, let us use the cylindrical rod example (Fig. 7.12 on p. 387). The inverse dielectric constant in this case is

$$\gamma(x, y) \;=\; \begin{cases} \gamma_{\text{rod}}, & r \le r_{\text{rod}} \\ \gamma_{\text{out}}, & r > r_{\text{rod}} \end{cases}, \qquad r \equiv (x^2 + y^2)^{\frac{1}{2}}, \quad (x, y) \in \Omega \tag{7.154}$$

The Fourier coefficients $\tilde{\gamma}(\mathbf{m})$ (that is, the plane wave expansion coefficients) for this function of coordinates are found by integration:

$$\tilde{\gamma}(\mathbf{m}) \;=\; \int_{\Omega} \gamma(\mathbf{r}) \exp(-\mathrm{i}\mathbf{k}_{\mathbf{m}} \cdot \mathbf{r})\, dx\, dy \tag{7.155}$$

This integration can be carried out analytically by switching to the polar coordinate system and using the Bessel function expansion for the complex exponential; see e.g. K. Sakoda [Sak05]. The end result is

$$\tilde{\gamma}(\mathbf{m}) \;=\; \begin{cases} f\gamma_{\text{rod}} + (1-f)\gamma_{\text{out}}, & m = 0 \\ 2(\gamma_{\text{rod}} - \gamma_{\text{out}})(k_{\mathbf{m}} r_{\text{rod}})^{-1} f J_1(k_{\mathbf{m}} r_{\text{rod}}), & m \neq 0 \end{cases} \tag{7.156}$$

Fig. 7.14 is a plot of $\gamma(x, y) \equiv \epsilon^{-1}(x, y)$ along the straight line $x = y$, i.e. at 45° to the axes of the computational cell. Parameters are the same as in the FE example: cell size $a = 1$ in each direction; $\epsilon_{\text{rod}} = 9$; $r_{\text{rod}} = 0.38$. The true plot of γ is of course a rectangular pulse that changes abruptly from $\gamma_{\text{rod}} = 1/9$ to $\gamma_{\text{out}} = 1$ at $x = r_{\text{rod}}/\sqrt{2} \approx 0.2687$.

Summation of a finite number of harmonics in the Fourier series produces typical ringing around the points of abrupt changes of the material parameter. When the number of Fourier harmonics retained in the series is increased, this ringing becomes less pronounced but does not fully disappear – compare the plots corresponding to 20 and 50 harmonics per component of the wavevector, Fig. 7.14.

In practice, the number of plane waves in the expansion is limited by the computational cost of the procedure (see Appendix 7.15), which in turn

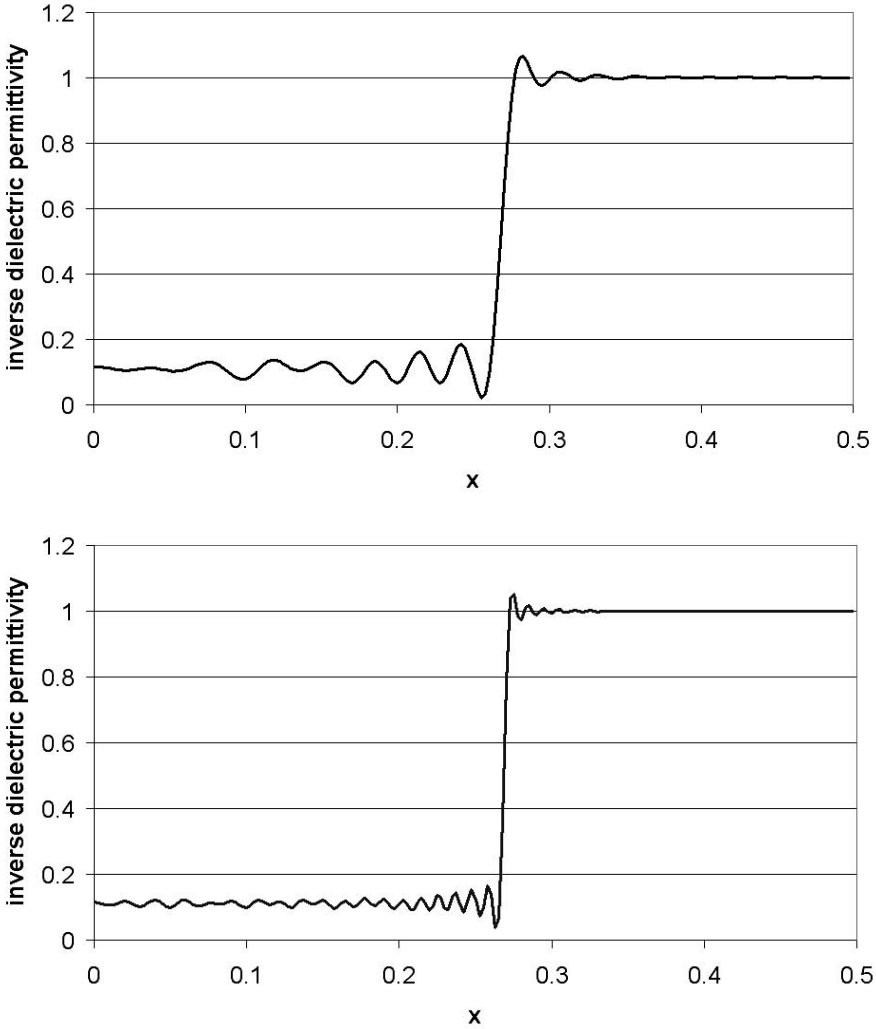


Fig. 7.14. An illustration of the Gibbs phenomenon for the Fourier series approximation of the inverse permittivity of a cylindrical rod in a square lattice cell. Cell size $a = 1$ in each direction; $\epsilon_{\rm rod} = 9$; $r_{\rm rod} = 0.38$. Top: 20 Fourier harmonics retained per coordinate direction; bottom: 50 harmonics.

limits the numerical accuracy of plane wave expansion. Because of that, in some cases the computational results initially reported in the literature had to be revised later. A. Moroz [Mor02] (p. 115109-3) gives one such example – the PBG of a diamond lattice of nonoverlapping dielectric spheres in air.

Remark 24. An alternative approach used by Moroz is the Korringa–Kohn–Rostoker[13] (KKR) method developed initially for the Schrödinger equation in the band theory of solids [KR54] and later adapted and adopted in photonics. KKR combines multipole expansions with transformations of lattice sums. This book deals with lattice sums for the static cases only, in the context of Ewald methods (Chapter 5). The wave case is substantially more involved, and the interested reader is referred to Chapter 2 of [Yas06] (by L.C. Botten *et al.*), to the work of R.C. McPhedran *et al.* [MNB05] and references therein.

To reduce the numerical errors associated with the Gibbs phenomenon in plane wave expansion, homogenization can be used to smooth out the dielectric permittivity at material interfaces; see R.D. Meade *et al.* [MRB+93] (with the erratum [MRB+97]). In particular, this approach is implemented in the MIT Photonic Bands eigenmode solver, a public-domain software package developed by the research groups of S.G. Johnson & J. Joannopoulos [JJ01].

7.9.4 FEM for Photonic Bandgap Problems in 2D

The Finite Element Method (FEM, Chapter 3) can be applied to either of the two formulations: for the full E field (7.138), (7.140), (7.141) or for the spatial-periodic factor E_{PER} (7.144), (7.145), (7.145). In 2D, both routes are analogous, but we focus on the first one to highlight the treatment of the special Bloch boundary conditions. (In 3D, FE analysis is more involved; see Section 7.10.)

The FE formulation starts with the definition of appropriate functional spaces (continuous and discrete) and with the weak form of the governing equations. This setup is needed not only as a mathematical technicality, but also for correct practical implementation of the algorithm – in particular, in the case under consideration, for the proper treatment of boundary conditions.

A natural functional space $\mathcal{B}(\Omega) \subset H^1(\Omega)$ ($\mathcal{B}$ for "Bloch") in our 2D example is the subspace of "scaled-periodic" functions in the Sobolev space $H^1(\Omega)$:

$$\mathcal{B}(\Omega) = \{E : \; E \in H^1(\Omega); \;\; E \text{ satisfies boundary conditions (7.140), (7.141)}\} \tag{7.157}$$

The weak formulation of the problem is

$$\text{Find } E \in \mathcal{B}(\Omega) : \; \left(\mu^{-1}\nabla E, \nabla E'\right) = \omega^2 \left(\epsilon E, E'\right), \quad \forall E' \in \mathcal{B}(\Omega) \tag{7.158}$$

or, for $\mu = \text{const}$,

$$\text{Find } E \in \mathcal{B}(\Omega) : \; (\nabla E, \nabla E') = \omega^2 \mu \left(\epsilon E, E'\right), \quad \forall E' \in \mathcal{B}(\Omega) \tag{7.159}$$

[13] Sometimes incorrectly spelled as "Rostocker".

Remark 25. The line integral (surface integral in 3D) that typically appears in the transition from the strong to the weak formulation and back (see Chapter 3) in this case vanishes:

$$\text{Find } E \in \mathcal{B}(\Omega): \quad \int_{\Gamma} \frac{\partial E}{\partial n} E'^{*} d\Gamma \;=\; 0; \qquad \forall E, E' \in \mathcal{B}(\Omega) \tag{7.160}$$

where Γ is the boundary of the computational cell Ω and n is the outward normal to this boundary. Indeed, the E field on the right edge of Ω has an additional Bloch factor $b = \exp(-\mathrm{i}K_x a)$ as compared to the left edge; similarly, the complex conjugate of the test function E' has an additional factor b^*. The integrals over the right and left edges then cancel out because $bb^* = 1$ (real $\mathbf{K}$ is assumed) and the directions of the outward normals on these edges are opposite. The integrals over the lower and upper edges cancel out for the same reason.

Next, assume that a finite element mesh (e.g. triangular or quadrilateral) has been generated. One special feature of the mesh is needed for the most natural implementation of the Bloch boundary conditions. The right and left edges of the computational domain Ω (a square in our example) need to be subdivided by the grid nodes in an identical fashion, so that the nodes on the right and left edges come in pairs with the same y-coordinate. A completely similar condition applies on the lower and upper edges.[14]

In each pair of boundary nodes, one node is designated as a "master" node (M) and the other one as a "slave" node (S).[15] The Bloch boundary condition directly relates the field values at the slave nodes to the respective values at their master nodes:

$$E(\mathbf{r}_S) \;=\; \exp\left(-\mathrm{i}\mathbf{K}\cdot(\mathbf{r}_S - \mathbf{r}_M)\right) E(\mathbf{r}_M) \tag{7.161}$$

where $\mathbf{r}_S$, $\mathbf{r}_M$ are the position vectors of any given slave–master pair of nodes.

Remark 26. For *edge elements* (see Chapter 3), one would consider pairs of master–slave *edges* rather than nodes.

We can now move on to the discrete FE formulation. Let $\mathcal{P}_h(\Omega)$ be one of the standard FE spaces of continuous piecewise-polynomial functions on the chosen mesh; see Chapter 3. The simplest such space is that of continuous piecewise-linear functions on a triangular grid. Any function $E_h \in \mathcal{P}_h$ can be represented as a linear combination of standard nodal FE basis functions $\psi_\alpha(x, y)$ (e.g. piecewise-linear "hat" functions, Chapter 3):

[14] For definiteness, let us attribute the corner nodes to the lower/upper edge pairs rather than to the left/right.

[15] For each pair of nodes, this assignment of M-S labels is in principle arbitrary; however, for consistency it is convenient to treat all nodes on, say, the left and lower edges as "masters" and the nodes on the right and upper edges as the respective "slaves".

$$E_h = \sum_{\alpha=1}^{n} E_\alpha \psi_\alpha, \qquad \alpha = 1, 2, \ldots, n \tag{7.162}$$

where n is (for nodal elements) the number of nodes of the mesh. The nodal values E_α of the field can be combined in one Euclidean vector $\underline{E} \in \mathbb{C}^n$.

The linear combination (7.162) establishes a one-to-one correspondence between each FE function E_h and the respective vector of nodal values $\underline{E}$. Bilinear forms in $\mathcal{P}_h \times \mathcal{P}_h$ and $\mathbb{C}^n \times \mathbb{C}^n$ are also related directly:

$$(\nabla E_h, \nabla E'_h) = (L\underline{E}, \underline{E}'), \quad \forall E_h \in \mathcal{P}_h(\Omega) \tag{7.163}$$

$$(\epsilon E_h, E'_h) = (M\underline{E}, \underline{E}'), \quad \forall E_h \in \mathcal{P}_h(\Omega) \tag{7.164}$$

In the left hand side of these two equations, the inner products are those of $(L_2(\Omega))^2$ and $L_2(\Omega)$, i.e.

$$(\nabla E_h, \nabla E'_h) \equiv \int_\Omega \nabla E_h \cdot \nabla E'^*_h \, d\Omega; \qquad (\epsilon E_h, E'_h) \equiv \int_\Omega \epsilon E_h E'^*_h \, d\Omega \tag{7.165}$$

In the right hand sides, the inner products are in $\mathbb{C}^n$:

$$(\underline{E}, \underline{E}') = \sum_{\alpha=1}^{n} E_\alpha E'^*_\alpha \tag{7.166}$$

Matrices L of (7.163) and M of (7.164) are, in the FE terminology, the "stiffness" matrix and the "mass" matrix, respectively (Chapter 3). Equations (7.163), (7.164) can be taken as definitions of these matrices. The entries of L and M can also be written out explicitly:

$$L_{\alpha\beta} = (\nabla\psi_\alpha, \nabla\psi_\beta) \quad 1 \le \alpha, \beta \le n \tag{7.167}$$

$$M_{\alpha\beta} = (\epsilon\, \psi_\alpha, \psi_\beta) \quad 1 \le \alpha, \beta \le n \tag{7.168}$$

where the inner products are again those of $L_2(\Omega)$ and the ψs are the FE basis functions.

To complete the FE formulation of the Bloch–Floquet problem, we need the subspace $\mathcal{B}_h \subset \mathcal{P}_h$ of piecewise-polynomial functions that satisfy the Bloch boundary condition (7.161) for each pair of master–slave nodes. (Practical implementation will be discussed shortly.) The FE-Galerkin formulation is nothing else but the weak form of the problem restricted to the FE space $\mathcal{B}_h$:

$$\text{Find } E \in \mathcal{B}_h(\Omega), \quad \omega \in \mathbb{C}: \quad (\nabla E, \nabla E') = \omega^2 \mu \, (\epsilon E, E'), \quad \forall E' \in \mathcal{B}_h(\Omega) \tag{7.169}$$

If there were no boundary constraints, this formulation in matrix-vector form would be

$$\text{Find } \underline{E} \in \mathbb{C}^n, \quad \omega \in \mathbb{C}: \quad \left(L\underline{E}, \underline{E}'\right) = \omega^2 \mu \, \left(M\underline{E}, \underline{E}'\right), \quad \forall \underline{E}' \in \mathbb{C}^n$$

where L and M are the stiffness and mass matrices previously defined.

However, the Bloch boundary conditions must be honored. To accomplish this algorithmically, let us separate out the slave nodes in the Euclidean vectors:

$$\underline{E} = \begin{pmatrix} \underline{E}_{\text{non-S}} \\ \underline{E}_S \end{pmatrix}; \quad \underline{E}_{\text{non-S}} \in \mathbb{C}^{n-n_S}; \quad \underline{E}_S \in \mathbb{C}^{n_S} \tag{7.170}$$

where n_S is the number of slave nodes. Vector $\underline{E}_S$ includes the field values associated with slave nodes; vector $\underline{E}_{\text{non-S}}$ is associated with "non-slaves," i.e. the non-boundary nodes and the master nodes.

Since the nodal values of slave nodes are completely defined by non-slaves, the full vector $\underline{E}$ can be obtained from its non-slave part by a linear operation:

$$\underline{E} = C\underline{E}_{\text{non-S}} \tag{7.171}$$

where C is a rectangular matrix

$$C = \begin{pmatrix} I \\ C_{\text{non-S}\to S} \end{pmatrix} \tag{7.172}$$

Each row of the matrix block $C_{\text{non-S}\to S}$ corresponds to a slave node and contains exactly one nonzero entry, the complex exponential Bloch factor of (7.161), in the column corresponding to the respective master node. The problem now takes on the following Galerkin matrix-vector form:

$$\text{Find } \underline{E}_{\text{non-S}} \in \mathbb{C}^{n-n_S}, \quad \omega \in \mathbb{C}:$$

$$\left(LC\underline{E}_{\text{non-S}}, C\underline{E}'_{\text{non-S}}\right) = \omega^2\mu \left(MC\underline{E}_{\text{non-S}}, C\underline{E}'_{\text{non-S}}\right), \quad \forall \underline{E}'_{\text{non-S}} \in \mathbb{C}^{n-n_S} \tag{7.173}$$

This immediately translates into the eigenvalue problem

$$\tilde{L}\underline{E}_{\text{non-S}} = \omega^2\mu\tilde{M}\underline{E}_{\text{non-S}} \tag{7.174}$$

where

$$\tilde{L} = C^*LC; \quad \tilde{M} = C^*MC \tag{7.175}$$

It is straightforward to show that both matrices $\tilde{L}$, $\tilde{M}$ are Hermitian; $\tilde{L}$ is positive definite if the Bloch wavenumber K is nonzero; $\tilde{M}$ is always positive definite.[16]

In practice, there is no need to multiply matrices in the formal way of (7.175). Instead, the following procedure can be applied. Consider a stage of the matrix assembly process where an entry (i, j) of the stiffness or mass matrix is being formed. If i happens to be a slave node with its master $M(i)$, the matrix entry gets multiplied by the Bloch exponential factor $b(i, M(i))$

[16] Indeed, by definition of the FE matrices, $(\tilde{L}\underline{E}_{\text{non-S}}, \underline{E}_{\text{non-S}}) = \int_\Omega |\nabla E_h|^2 d\Omega$, $\forall E_h \in \mathcal{B}_h$. Since E_h for $K \neq 0$ cannot be constant due to the Bloch boundary condition, this energy integral is strictly positive. Similar considerations apply to $\tilde{M}$.

(7.161) and attributed to row $M(i)$ rather than row i. Likewise, if j is a slave node with the corresponding master node $M(j)$, the matrix entry gets multiplied by $b^*(j, M(j)) = \exp(\mathrm{i}\mathbf{K}\cdot(\mathbf{r}_j - \mathbf{r}_{M(j)}))$ (note the complex conjugate) and the result gets attributed to column $M(j)$ instead of column j.

In this procedure, the rows and columns corresponding to slave nodes remain empty and in the end can be removed from the matrices. However, it may be algorithmically simpler *not* to change the dimension and structure of the matrices and simply fill the "slave" entries in the diagonals with some dummy numbers – say, ones for matrix M and some large number X for matrix L. This will produce extraneous modes "living" on the slave nodes only and corresponding to eigenvalues $\omega^2\mu = X$. These modes can be easily recognized and filtered out in postprocessing.

A disadvantage of FEM for the bandgap structure calculation is that it leads to a *generalized* eigenvalue problem, of the form $Lx = \lambda Mx$ rather than $Lx = \lambda x$. This increases the computational complexity of the solver. Note, however, that if the Cholesky decomposition[17] of M ($M = TT^*$, where T is a lower triangular matrix) is not too expensive, the generalized problem can be reduced to a regular one by substitution $y = T^*x$:

$$Lx = \lambda TT^*x \quad \Rightarrow \quad T^{-1}LT^{-*}y = \lambda y \tag{7.176}$$

If iterative eigensolvers are used, matrix inverses need not be computed directly; instead, systems of equations with upper or lower triangular matrices are solved to find $T^{-1}LT^{-*}y$ for an arbitrary vector y. However, in the numerical example below the matrices are of very moderate size and the Matlab QZ algorithm (a direct solver for generalized eigenvalue problems) is employed.

7.9.5 A Numerical Example: Band Structure Using FEM

The numerical data was chosen the same as in the computational example of K. Sakoda ([Sak05], pp. 28–29), where the bandgap structure was computed using Fourier analysis (plane wave expansion). Our finite element results can then be compared with those of [Sak05]. The general setup, with a cylindrical dielectric rod in a square lattice cell, was already shown in Fig. 7.12 (p. 387). The cell size is taken as $a = 1$ and the radius of the cylindrical rod is $r_{\mathrm{rod}} = 0.38$. The dielectric constant of the rod is $\epsilon_{\mathrm{rod}} = 9$; the medium outside the rod is air, with $\epsilon_{\mathrm{out}} = 1$.

The FE mesh is generated by FEMLAB™ (COMSOL Multiphysics™) and exported to the Matlab environment; an FE matrix assembly for the Bloch–Floquet problem is then performed in Matlab. As already noted, the Matlab QZ solver is used. Postprocessing is again done in FEMLAB (COMSOL Multiphysics™).

[17] André-Louis Cholesky (1875–1918), a French mathematician. It is customary to write the Cholesky decomposition as LL^T or LL^*, but in our case symbol L is already taken, so T is used instead.

The initial FE mesh is fairly coarse, with 404 nodes and 746 first-order triangular elements (Fig. 7.15, left). The matrix assembly time is about half a second and the eigenvalue solver time is ∼8.5 seconds on a 2.8 GHz PC.

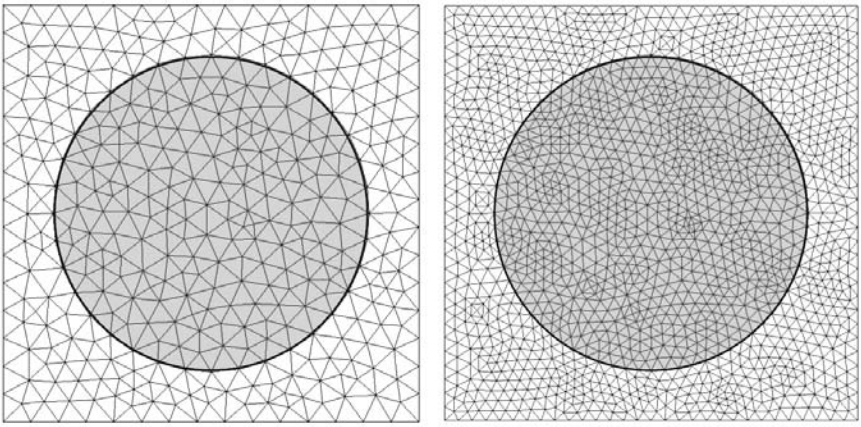


Fig. 7.15. Two finite element meshes for one cell of a photonic crystal lattice with cylindrical dielectric rods. The rod is shaded for visual clarity. Left: 404 nodes, 746 triangular elements. Right: 1553 nodes, 2984 triangular elements.

The main result of the FE simulation is the bandgap structure shown in Fig. 7.16 for the E-mode (s-polarization, one-component E-field). The first four normalized eigenfrequencies $\tilde{\omega} = \omega a/(2\pi c)$ (c being the speed of light in free space) are plotted vs. the normalized Bloch wavenumber Ka/π over the $M \to \Gamma \to X \to M$ loop in the Brillouin zone. The chart in Fig. 7.16 is almost exactly the same as the one in [Sak05].

The bandgaps, where no (real) eigenfrequencies exist for *any* $\mathbf{K}$, are shaded in the figure. The normalized frequency ranges for the first two gaps are, according to the FE calculation, [0.2462, 0.2688] and [0.4104, 0.4558].

To estimate the accuracy of this numerical result, the computation was repeated on a finer mesh, with 1553 nodes and 2984 first-order triangular elements (Fig. 7.15, right).[18] On the finer mesh, the first two bandgaps are calculated to be [0.2457, 0.2678] and [0.4081, 0.4527], which differs from the results on the coarser mesh by 0.2–0.7%.

For comparison, the first two bandgap frequency ranges reported for the same problem by K. Sakoda [SS97, Sak05] are [0.247, 0.277] and [0.415, 0.466]. This result was obtained by Fourier analysis, with expansion into 441 plane waves; the estimated accuracy is about 1% according to Sakoda.

[18] In modern FE analysis, much more elaborate *hp*-refinement procedures exist to estimate and improve the numerical accuracy. See Chapter 3.

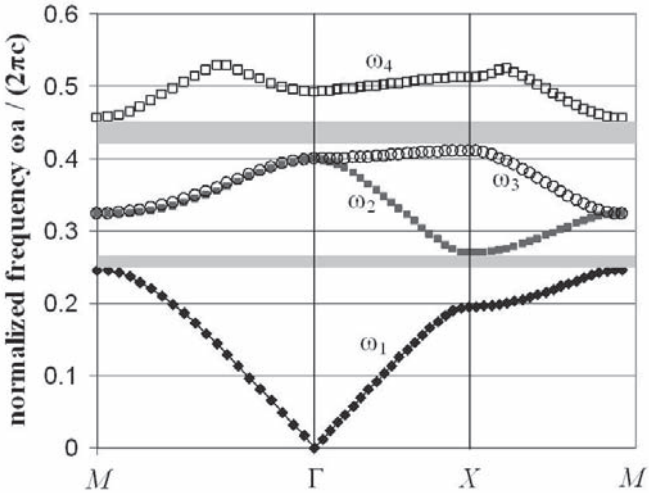


Fig. 7.16. The photonic band structure (plots correspond to the first four eigenfrequencies as a function of the wavevector) for a photonic crystal lattice; E-mode (one-component E-field). Dielectric cylindrical rods in air; cell size $a = 1$, radius of the cylinder $r_{\text{rod}} = 0.38$; the relative dielectric permittivity $\epsilon_{\text{rod}} = 9$.

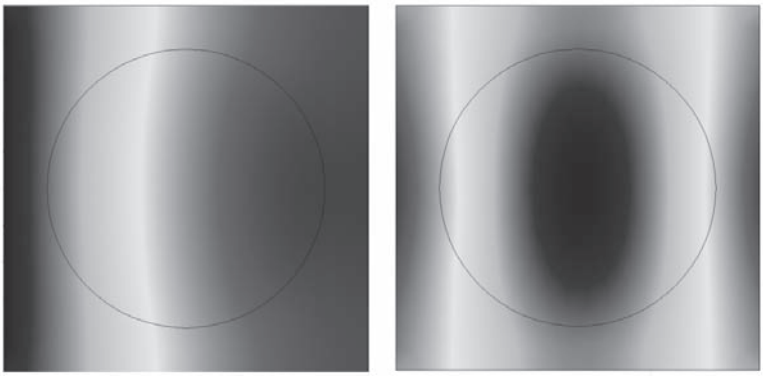


Fig. 7.17. The E-field distribution for the first (left) and the second (right) Bloch modes for $\mathbf{K} = (\frac{\pi}{2a}, 0)$. Same setup and parameters as in Fig. 7.16.

The field distribution of two low order Bloch modes is illustrated by Fig. 7.17 and Fig. 7.18. The first figure is for the Bloch vector $\mathbf{K} = (\frac{\pi}{2a}, 0)$ (a Δ-point exactly in the middle of ΓX), and the second one is for point $\mathbf{K} = (\frac{\pi}{2a}, \frac{\pi}{4a})$.

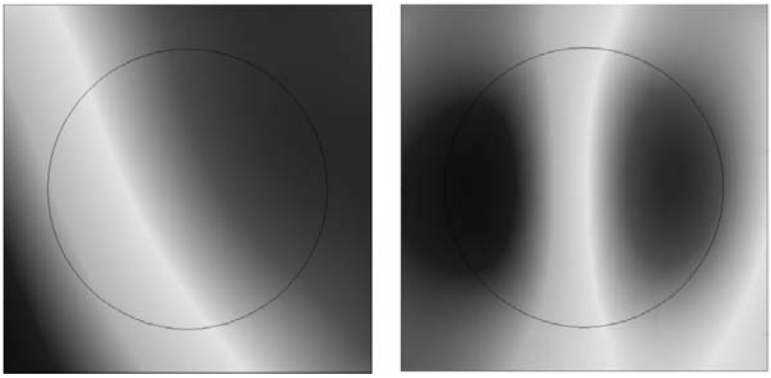


Fig. 7.18. The E-field distribution for the first (left) and the second (right) Bloch modes for $\mathbf{K} = (\frac{\pi}{2a}, \frac{\pi}{4a})$. Same setup and parameters as in Fig. 7.16.

This relatively simple comparison example of FEM vs. Fourier expansion is not a basis for far-reaching conclusions. Both methods have their strengths and weaknesses. A clear advantage of FEM is its effective and accurate treatment of geometrically complex structures, possibly with high dielectric contrasts. Another advantage is the sparsity of the system matrices. Unfortunately, FEM leads to a *generalized* eigenvalue problem, with the FE "mass" matrix in the right hand side.[19] A special FE technique known as "mass lumping" makes the mass matrix diagonal, with applications to both eigenvalue and time-dependent problems. Mass lumping is usually achieved by applying, in the FE context, numerical quadratures with the integration knots chosen to coincide with element nodes. For details, see papers by M.G. Armentano & R.G. Durán [AD03]; A. Elmkies & P. Joly [EJ97a, EJ97b]; G. Cohen & P. Monk [CM98], and references there. In addition, as already noted, the generalized problem can be converted to a regular one by Cholesky decomposition.

[19] The presence of the mass matrix is also a disadvantage in time-dependent problems, where this matrix is associated with the time derivative term and makes explicit time-stepping schemes difficult to apply.

7.9.6 Flexible Local Approximation Schemes for Waves in Photonic Crystals

As an alternative to plane wave expansion and to Finite Element analysis, the Flexible Local Approximation MEthod (FLAME, Chapter 4) can be used for wave simulation in photonic crystal devices.

FLAME incorporates accurate local approximations of the solution into a difference scheme. Applications of FLAME to photonic crystals are attractive because local analytical approximations for typical photonic crystal structures are indeed available and the corresponding FLAME basis functions can we worked out once and for all. In particular, for crystals with cylindrical rods the FLAME basis functions are obtained by matching, via the boundary conditions on the rod, cylindrical harmonics inside and outside the rod. These Bessel-based basis functions were already derived in Chapter 4 for the problem of electromagnetic scattering from a cylinder. In 3D, FLAME bases for electromagnetic fields near dielectric spheres could be constructed by matching the (vector) spherical harmonics inside and outside the sphere as in Mie theory (J.A. Stratton [Str41] or R.F. Harrington [Har01]).

When the dielectric structures are not cylindrical or spherical, the field can still be expanded into cylindrical/spherical harmonics, and the T- ("transition") matrix provides the relevant relationships between the coefficients of incoming and outgoing waves. A comprehensive treatment of T-matrix methods and related electromagnetic theory can be found in the books and articles by M.I. Mishchenko *et al.* [MTM96, MTL02, MTL06], with a large reference database [MVB$^+$04] and a public-domain FORTRAN code [MT98] being available. In contrast with methods that analytically combine multipole expansions and lattice sums (see Remark 24 on p. 393), the role of multipole expansions in FLAME is to generate a difference scheme.

As an illustrative example, we consider a photonic crystal analyzed by T. Fujisawa & M. Koshiba [FK04, Web07]. The waveguide with a bend is obtained by eliminating a few dielectric cylindrical rods from a 2D array (Fig. 7.19). Fujisawa & Koshiba used a Finite Element–Beam Propagation method in the time domain to study fields in such a waveguide, with nonlinear characteristics of the rods. The use of complex geometrically conforming finite element meshes may well be justified in this 2D case. However, regular Cartesian grids have the obvious advantage of simplicity, especially with extensions to 3D in mind. This is illustrated by numerical experiments below.

The problem is solved in the frequency domain and the material characteristic of the cylindrical rods is assumed linear, with the index of refraction $n = 3$. The radius of the cylinders and the wavenumber are normalized to unity; the air gap between the neighboring rods is equal to their radius. The field distribution is shown in Fig. 7.19.

For bandgap operation, the field is essentially confined to the guide, and the boundary conditions do not play a critical role. To get numerical approximation of these conditions out of the picture in this example, the field on

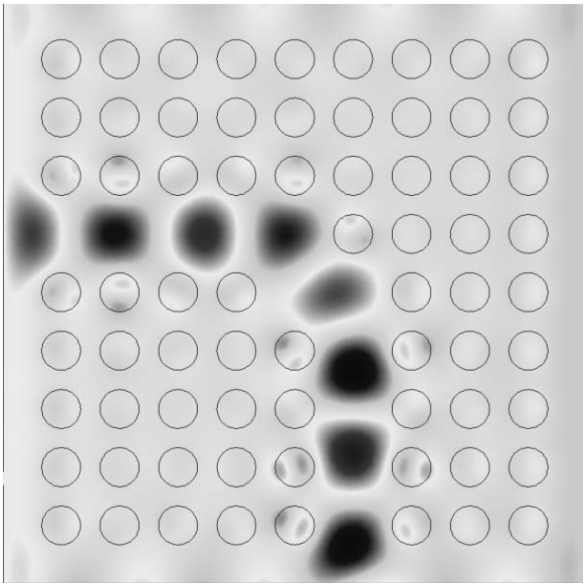


Fig. 7.19. The imaginary part of the electric field in the photonic crystal waveguide bend. The real part looks qualitatively similar. (Reprinted by permission from [Tsu05a] ©2005 IEEE.)

the surface of the crystal was simply set equal to an externally applied plane wave.

For comparison, FE simulations (FEMLAB – COMSOL Multiphysics^{TM}) with three meshes were run: the initial mesh with 9702 nodes, 19,276 elements, and 38,679 degrees of freedom (d.o.f.); a mesh obtained by global refinement of the initial one (38,679 nodes, 77,104 elements, 154,461 d.o.f.); and an adaptively refined mesh with 27,008 nodes, 53,589 elements, 107,604 d.o.f. The elements were second order triangles in all cases. The agreement between FLAME and FEM results is excellent. This is evidenced, for example, by Fig. 7.20, where almost indistinguishable FEM and FLAME plots of the field distribution along the central line of the crystal are shown.

Yet, a closer look at the central peak of the field distribution (Fig. 7.21) reveals that FLAME has essentially converged for the 50×50 grid, while FEM solutions approach the FLAME result as the FE mesh is refined. FEM needs well above 100,000 d.o.f. to achieve the level of accuracy comparable with the FLAME solution with 2500 d.o.f. [Tsu05a]. Fig. 7.22 gives a visual comparison of FEM and Trefftz–FLAME meshes that provide the same accuracy level.

Note that for the 50×50 grid there are about 10.5 points per wavelength (ppw) in the air but only 3.5 ppw in the rods, and yet the FLAME results are very accurate because of the special approximation used. Any alternative method, such as FE or FD, that employs a generic (piecewise-polynomial)

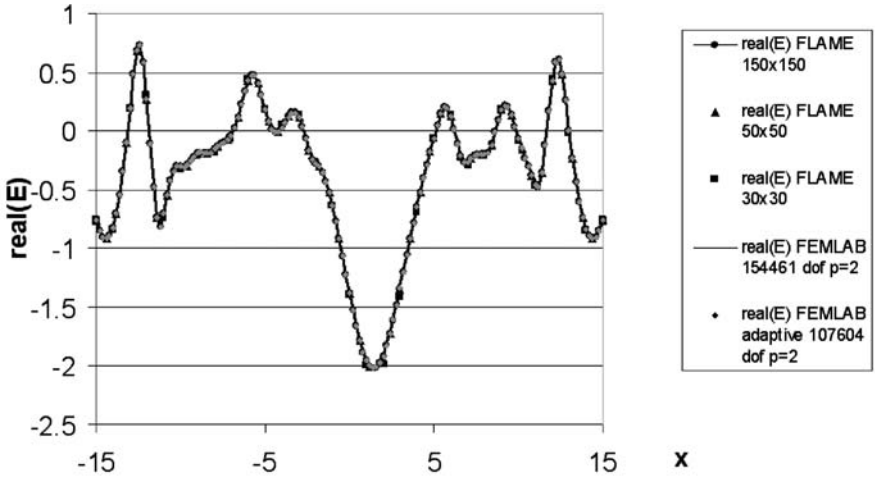


Fig. 7.20. Field distribution in the Fujisawa–Koshiba photonic crystal along the central line $y = 0$. FLAME vs. FE solutions. (Reprinted by permission from [Tsu05a] ©2005 IEEE.)

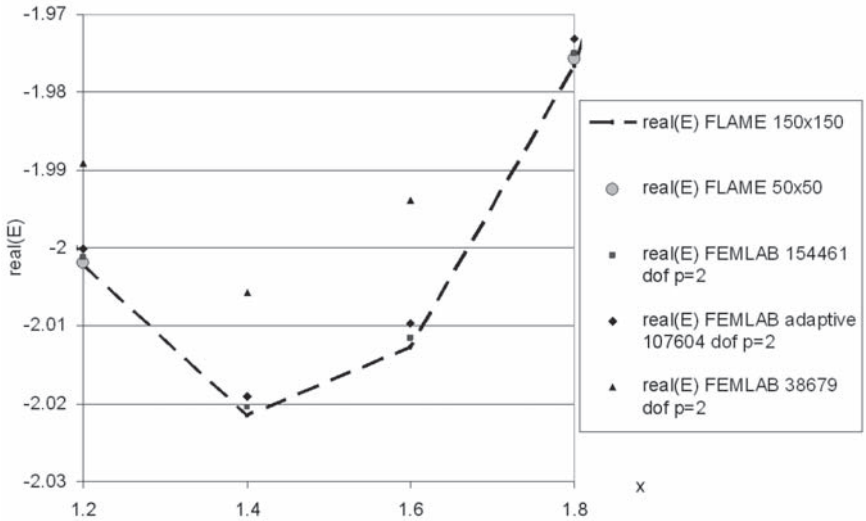


Fig. 7.21. Convergence of the field near the center of the bend. Trefftz–FLAME has essentially converged for the 50×50 grid (2500 d.o.f.); FEM results approach the FLAME values as the FE mesh is refined. FEM needs well over 100,000 d.o.f. for accuracy comparable with FLAME. (Reprinted by permission from [Tsu05a] ©2005 IEEE.)

approximation would require a substantially higher number of ppw to achieve the same accuracy.

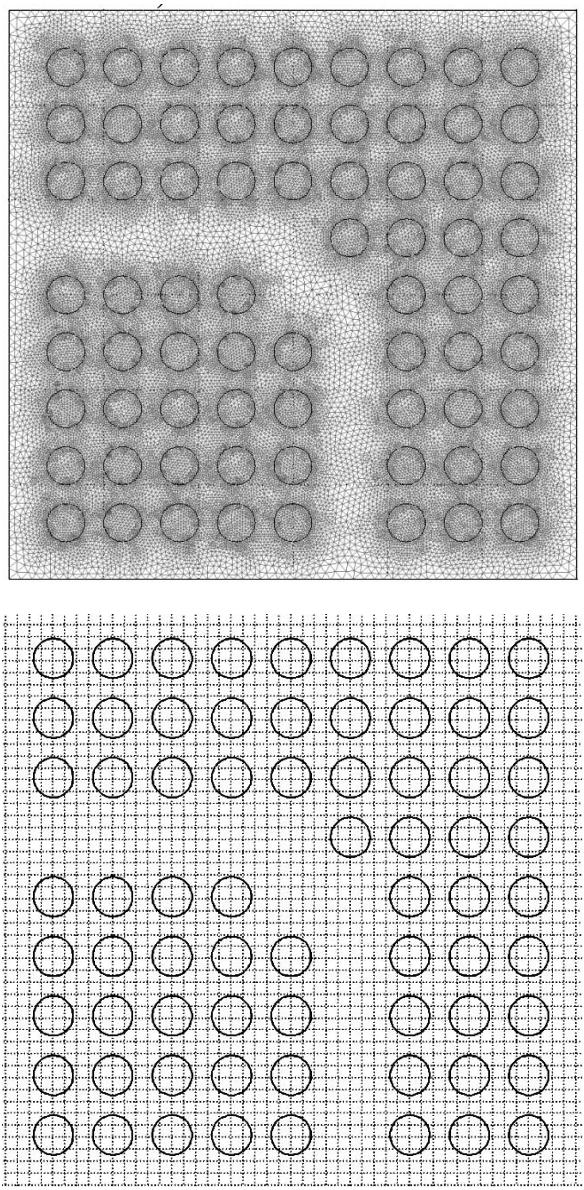


Fig. 7.22. The 50×50 FLAME grid (2500 d.o.f.) provides the same level of accuracy as the Finite Element mesh with 38,679 nodes, 77,104 elements and 154,461 d.o.f. (Reprinted by permission from [Tsu05a] ©2005 IEEE.)

Remark 27. As described in more detail in Section 7.9.7, the FLAME computation of Bloch–Floquet modes proceeds in a different manner than in the FE or plane wave methods. FLAME schemes rely on local analytical solutions that can be evaluated numerically only for a given (known) frequency. Hence ω becomes an "independent variable" in the simulation, and the Bloch–Floquet wave vector (say, along any given symmetry line in the Brillouin zone) is a parameter to be determined from a generalized eigenvalue problem.

FLAME eigenmode analysis has been performed by H. Pinheiro *et al.* [PWT07] in application to photonic crystal waveguides. The crystal is again formed by dielectric cylindrical rods. The waveguides "carved out" of the crystal lattice have ports that carry energy in and out of the device. What follows is a brief summary of the computational approach and results of [PWT07].

First, FLAME is used to compute Floquet-like modes that can propagate through the crystal in the direction of the waveguide (the energy of these modes is contained mostly within the guide). For this purpose, FLAME is applied to one layer of cylindrical rods, with the Bloch–Floquet boundary condition imposed on two of its sides and the FLAME PML (Perfectly Matched Layer) on the other two. This is a generalized eigenvalue problem that for moderate matrix sizes can be quickly solved using the QZ algorithm. There is normally no need to generate large matrices, as the convergence of FLAME is extremely rapid (see the following section).

Second, the boundary conditions for the field at the ports can be expressed via the dominant waveguide modes determined as described above. For the excited port(s), the excitation is assumed known; for other ports, zero Dirichlet conditions are used. FLAME is then applied again, this time for the whole crystal, with the proper boundary conditions at the ports and PML conditions on inactive surfaces.

The results of the first step of the analysis – computation of the propagation constant – show very good agreement with the plane wave expansion method when the FLAME grid has 6×6 nodes per lattice cell. Further, FLAME is applied to a 90° waveguide bend; the results obtained with 7744 degrees of freedom for FLAME agree well with those calculated by the FETD Beam Propagation Method using 158,607 d.o.f. (M. Koshiba *et al.* [KTH00]). Equally favorable is the comparison of FLAME with FETD-BPM for photonic crystals with Y- and T-branches. For a T-branch, FLAME results with 25,536 d.o.f. are the same as FDTD results with 5,742,225 d.o.f. FLAME solutions exhibit very fast convergence as the grid is refined. As an example, Fig. 7.23 shows transmission and reflection coefficients of a directional coupler (H. Pinheiro *et al.* [PWT07]).

7.9.7 Band Structure Computation Using FLAME

As an alternative to plane wave expansion (Section 7.9.1, p. 389) and FEM (Section 7.9, p. 389), let us now consider FLAME for band structure

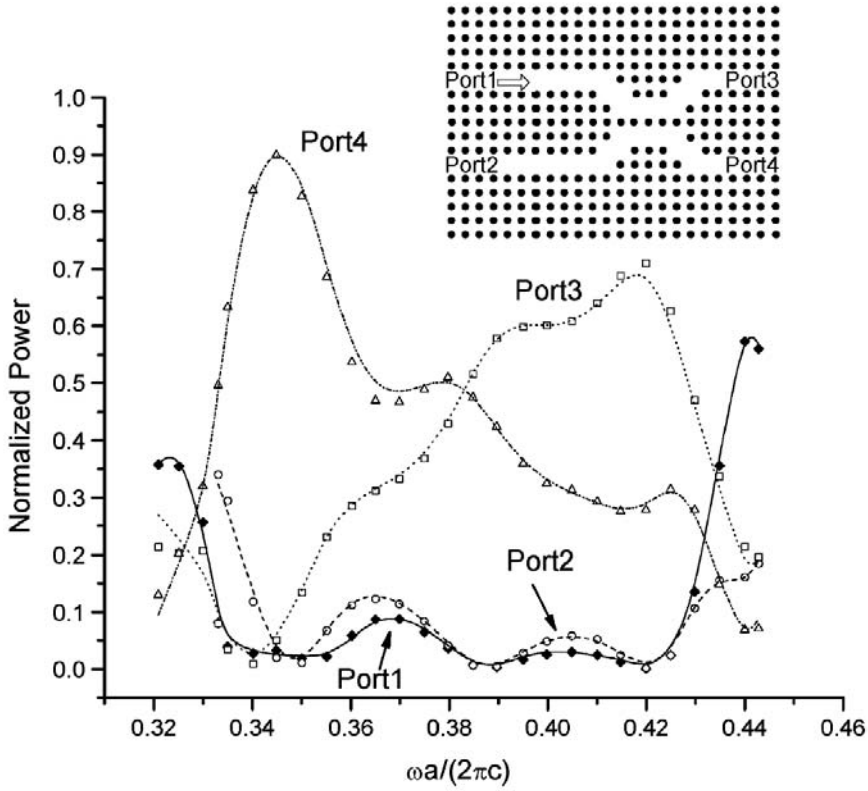


Fig. 7.23. (Credit: H. Pinheiro *et al.* Reprinted by permission from [PWT07] ©2007 IEEE.) Transmission and reflection coefficients of a directional coupler. Markers: FLAME results; lines: FETD-BPM results by M. Koshiba *et al.* [KTH00].

calculation.[20] The familiar case with a dielectric cylindrical rod of radius r_{rod} and dielectric permittivity ϵ_{rod} in a square lattice cell will again serve as a computational example.

In the vicinity of a cylindrical rod centered at the origin of a polar coordinate system (r, ϕ), the FLAME basis $\psi_\alpha^{(i)}$ contains Bessel/Hankel functions (see also Sections 4.4.11, 7.9.6, 7.11.5):

$$\psi_\alpha^{(i)} = a_n J_n(k_{\text{cyl}} r) \exp(in\phi), \quad r \le r_{\text{rod}}$$

$$\psi_\alpha^{(i)} = [c_n J_n(k_{\text{air}} r) + H_n^{(2)}(k_{\text{air}} r)] \exp(in\phi), \quad r > r_{\text{rod}}$$

where J_n is the Bessel function, $H_n^{(2)}$ is the Hankel function of the second kind [Har01], and the coefficients a_n, c_n are found by matching the values of $\psi_\alpha^{(i)}$ inside and outside the rod.

[20] The material of this section appears in [Tv07].

The 9-point (3×3) stencil with a grid size h is used and $1 \leq \alpha \leq 8$. The eight basis functions ψ are obtained by retaining the monopole harmonic ($n = 0$), two harmonics of orders $n = 1, 2, 3$ (i.e. dipole, quadrupole and octupole), and one of harmonics of order $n = 4$. This set of basis functions produces a 9-point scheme as the null vector of the respective matrix of nodal values (Sections 4.4.11, 7.9.6, 7.11.5).

The Bloch wave satisfying the second order differential equation calls for *two* boundary conditions – for the E field and for its derivative in the direction of wave propagation (or, equivalently, for the H field). Consequently, there are two discrete boundary conditions per Cartesian coordinate (compare this with a similar treatment in [PWT07] (p. 405) where, however, the algorithm is effectively one-dimensional). The implementation of these discrete conditions is illustrated by Fig. 7.24. As an example, the square lattice cell is covered with a 5×5 grid of "master" nodes (filled circles). In addition, there is a border layer of "slave" nodes (empty circles).

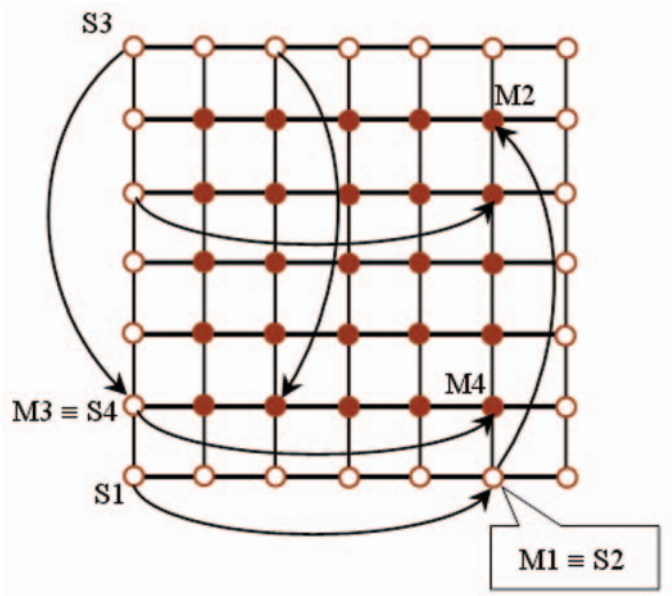


Fig. 7.24. Implementation of the Bloch–Floquet boundary conditions in FLAME. Empty circles – "slave" nodes, filled circles – "master" nodes. A few of the "slave–master" links are indicated with arrows. The corner nodes are the "slaves of slaves".

The FLAME scheme is generated for each of the *master nodes* ("M"). At slave nodes ("S"), the field is constrained by the Bloch–Floquet condition rather than by the difference scheme:

$$E(\mathbf{r}_S) \;=\; \exp\left(-\mathrm{i}\mathbf{K}_B \cdot (\mathbf{r}_S - \mathbf{r}_M)\right) E(\mathbf{r}_M) \tag{7.177}$$

Here $\mathbf{r}_S$, $\mathbf{r}_M$ are the position vectors of any given slave–master pair of nodes. Several such pairs are indicated in Fig. 7.24 by the arrows for illustration. Note that the corner nodes are the "slaves of slaves": for example, master node M1 for slave S1 is itself a slave S2 of node M2. This is algebraically equivalent to linking node S1 to M2; however, if the link $S1 \to M2$ were imposed directly rather than via $S1 \to M1 \to M2$, the corresponding factor would be the product of two Bloch exponentials in the x- and y-direction, leading to a complicated eigenvalue problem, bilinear with respect to the two exponentials.

Example equations for the Bloch boundary conditions, in reference to Fig. 7.24, are

$$E_{\mathrm{S1}} = b_x E_{\mathrm{M1}}; \quad b_y E_{\mathrm{S3}} = E_{\mathrm{M3}} \tag{7.178}$$

where b_x and b_y are the Bloch factors

$$b_x = \exp(\mathrm{i}K_x L_x); \quad b_y = \exp(\mathrm{i}K_y L_y) \tag{7.179}$$

In matrix-vector form, the FLAME eigenvalue problem is

$$L\underline{E} = (b_x B_x + b_y B_y)\underline{E} \tag{7.180}$$

where $\underline{E}$ is the Euclidean vector of nodal values of the field. The rows of matrix L corresponding to the master nodes contain the coefficients of the FLAME scheme, and the respective rows of matrices $B_{x,y}$ are zero. Each slave-node row of matrices L and B contains only one nonzero entry – either 1 or $b_{x,y}$, as exemplified by (7.178). Matrices L and (especially) B are sparse; typical sparsity patterns, for a 10×10 grid, are shown in Fig. 7.25.

Problem (7.180) contains three key parameters: ω, on which the FLAME scheme and hence the L matrix depend (for brevity, this dependence is not explicitly indicated), and the Bloch exponentials $b_{x,y}$. Finding three or even two independent eigenparameters simultaneously is not feasible. First, one chooses a value of ω and constructs the difference operator L for that value. In principle, for any given value of either of the b parameters (say, b_x) one could solve for the other parameter and scan the (b_x, b_y)-plane that way. Typically, however, the focus is only on the symmetry lines $\Gamma \to X \to M \to \Gamma$ of the first Brillouin zone. On ΓX, $b_y = 1$ and b_x is the only unknown; on XM, the only unknown is b_y; and on $M\Gamma$, the single unknown is $b = b_x = b_y$.

For comparison purposes, in the numerical example the numerical data was chosen the same as in the PWE computation of [Sak05], pp. 28–29. In the lattice of cylindrical rods, the size of the computational square cell is $a = 1$, and the radius of the cylindrical rod is $r_{\mathrm{rod}} = 0.38$. The dielectric constant of the rod is $\epsilon_{\mathrm{rod}} = 9$; the medium outside the rod is air, with $\epsilon_{\mathrm{out}} = 1$. In our FLAME simulation, due to very rapid convergence of the method, matrices L and M need only be of very moderate size, in which case the Matlab QZ algorithm (a direct solver for generalized eigenvalue problems) is very efficient.

Fig. 7.26 shows the same band diagram for the E-mode as Fig. 7.16, but the focus now is on the accuracy of FLAME and its comparison with other

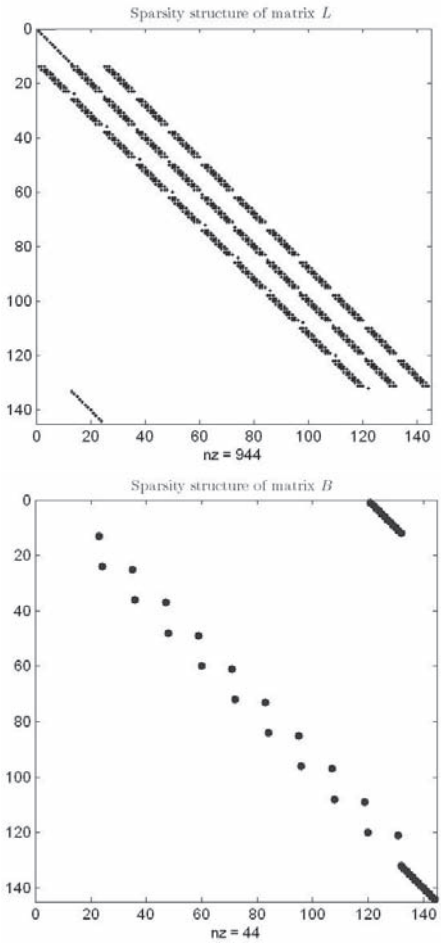


Fig. 7.25. Sparsity structure of the FLAME matrices for a 10×10 grid: L (top) and $B = B_x + B_y$ (bottom).

methods. Plotted in the figure are the first four normalized eigenfrequencies $\tilde{\omega} = \omega a/(2\pi c)$ (c being the speed of light in free space) vs. the normalized Bloch wavenumber $\tilde{K} = Ka/\pi$ over the $M \to \Gamma \to X \to M$ loop in the Brillouin zone. The bandgaps, where no (real) eigenfrequencies exist for *any* $\mathbf{K}_B$, are shaded in the figure. The excellent agreement between PWE, FEM and FLAME gives us full confidence in these results and allows us to proceed to a more detailed assessment of the numerical errors.[21]

[21] All numerical results were also checked for consistency on several meshes and for an increasing number of PWE terms.

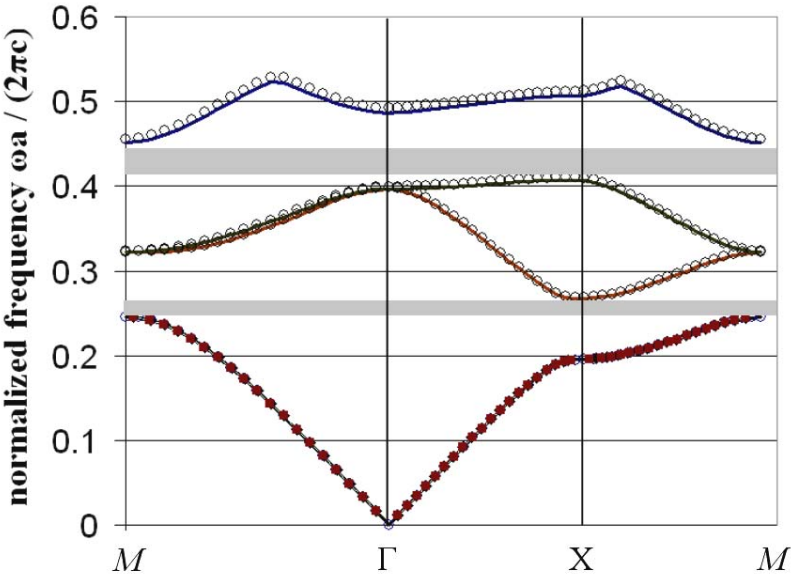


Fig. 7.26. The photonic band structure (first four eigenfrequencies as a function of the wavevector) for a photonic crystal lattice; E-mode. FEM (circles), PWE (solid lines), FLAME, grid 5×5 (diamonds), FLAME, grid 20×20 (squares). Dielectric cylindrical rods in air; cell size $a = 1$, radius of the cylinder $r_{\text{rod}} = 0.38$; the relative dielectric permittivities $\epsilon_{\text{rod}} = 9$; $\epsilon_{\text{out}} = 1$.

The accuracy of FLAME is much higher than that of PWE or FEM, with negligible errors achieved already for a 10×10 grid. Indeed, inspecting the computed Bloch–Floquet wavenumbers as the FLAME grid size decreases, we observe that 6–8 digits in the result stabilize once the grid exceeds 10×10 and 8–10 digits stabilize once the grid exceeds 20×20. This clearly establishes the 40×40 results as an "overkill" solution that can be taken as quasi-exact for the purpose of error analysis.

Errors in the Bloch wavenumber are plotted in Fig. 7.27. Very rapid convergence of FLAME with respect to the number of grid nodes is obvious from the figure. Further, the FLAME error for the Bloch number is about *six orders of magnitude lower* than the FEM error for approximately the same number of unknowns: 484 nodes (including "slaves") in FLAME and 404 nodes in FEM.

In the numerical example presented, FLAME provides 6–8 orders of magnitude higher accuracy in the photonic band diagram than PWE or FEM with the same number of degrees of freedom ($\sim$400).

To apply FLAME to more general shapes of dielectric structures, one needs accurate local approximations of the theoretical solution. This can be

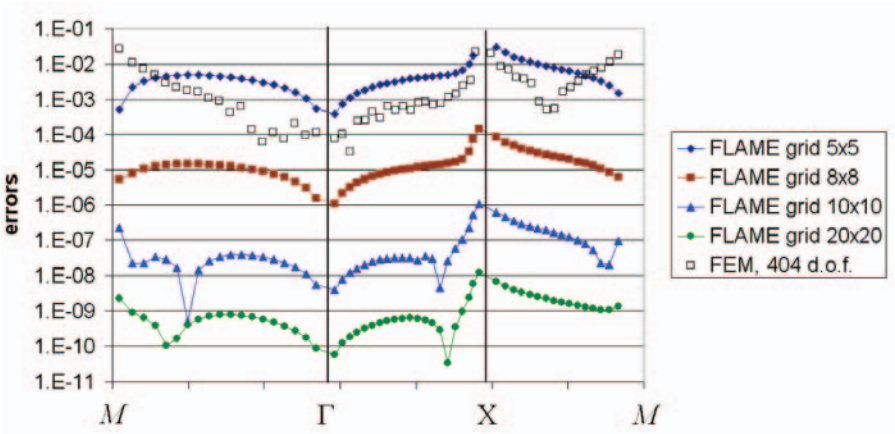


Fig. 7.27. Numerical errors in the Bloch wavenumber. Same parameters as in the previous figure. FLAME grids: 5×5 (diamonds), 8×8 (squares), 10×10 (triangles), 20×20 (circles). FEM, 404 d.o.f. (empty squares).

achieved, for example, by approximating the air-dielectric boundaries with arcs in a piecewise fashion and then using the Bessel-Hankel basis described in the paper. Alternatively, basis functions can be obtained as accurate finite element or boundary element solutions of local problems that are much smaller than the global one [DT06]. Extensions of the methodology to 3D appear to be possible, with FLAME basis functions derived either from Mie theory at (piecewise-)spherical boundaries or, alternatively, by solving small-size local problems with finite elements or boundary elements.

7.10 Photonic Bandgap Calculation in Three Dimensions: Comparison with the 2D Case

This section reviews the main ideas of PBG analysis in three dimensions, highlighting the most substantial differences with the 2D case and the complications that arise.

7.10.1 Formulation of the Vector Problem

One of the most salient new features of the 3D formulation, as compared to 2D, is that it is no longer a scalar problem. Maxwell's equations for time-harmonic fields, with no external currents ($\mathbf{J} = 0$), are

$$\nabla \times \mathbf{E} = -\mathrm{i}\omega \mathbf{B} \tag{7.181}$$

$$\nabla \times \mathbf{H} = \mathrm{i}\omega \mathbf{D} \tag{7.182}$$

See Section 7.2 (p. 353) for more details on Maxwell's equations, as well as the notational conventions on complex phasors and symbol i.

We shall assume simple material relationships $\mathbf{B} = \mu\mathbf{H}$ and $\mathbf{D} = \epsilon\mathbf{E}$, where μ and ϵ can depend on coordinates (in photonics, however, materials are usually nonmagnetic and then $\mu = \mu_0 = \text{const}$).

Taking the curl of either one of the Maxwell equations and substituting into the other one yields a single second-order equation for the field:

$$\nabla \times \mu^{-1} \nabla \times \mathbf{E} - \omega^2 \epsilon \mathbf{E} = 0 \tag{7.183}$$

or, alternatively,

$$\nabla \times \epsilon^{-1} \nabla \times \mathbf{H} - \omega^2 \mu \mathbf{H} = 0 \tag{7.184}$$

The two formulations are analogous but not computationally equivalent as we shall see.

For simplicity of exposition, let us assume a cubic primary cell $[-a/2, a/2]^3$ in real space; extensions to hexahedral and triclinic cells are straightforward both in plane wave methods and in FE analysis. (The plane wave method is currently used much more widely in PBG calculation than FEM.) As in 2D, the E-field in formulation (7.183) is sought as a Bloch wave with some wave vector $\mathbf{K}$:

$$\mathbf{E}(\mathbf{r}) = \mathbf{E}_{\text{PER}}(\mathbf{r}) \exp(-\mathrm{i}\mathbf{K} \cdot \mathbf{r}); \qquad \mathbf{r} \equiv (x, y, z) \tag{7.185}$$

One can solve for the full E-field of (7.183) or, alternatively, for factor $\mathbf{E}_{\text{PER}}(x, y, z)$ that satisfies periodic conditions on the boundary of the computational cell. As in 2D, the trade-off between these two formulations is in the relative complexity of the boundary conditions vs. that of the differential operator.

The "scaled-periodic" boundary condition for the full E-field is

$$\mathbf{E}\left(\frac{a}{2}, y, z\right) = \exp(-\mathrm{i}K_x a)\, \mathbf{E}\left(-\frac{a}{2}, y, z\right); \tag{7.186}$$

and analogous conditions for two other pairs of faces

In the formulation for $\mathbf{E}_{\text{PER}}$, the $\nabla\times$ operator applied to $\mathbf{E}$ can be formally replaced with $(\nabla - \mathrm{i}\mathbf{K})\times$ applied to $\mathbf{E}_{\text{PER}}$, and the boundary conditions are purely periodic. A detailed and mathematically rigorous exposition, with the finite element (more specifically, edge element) solution is given by D.C. Dobson & J.E. Pasciak [DP01]; they use the $\mathbf{E}_{\text{PER}}$ formulation.

We turn to the plane wave method first; the finite element solution will be considered later in this section. As in 2D, the periodic factor $\mathbf{E}_{\text{PER}}$ can be expanded into a Fourier series with some coefficients $\tilde{\mathbf{E}}_{\text{PER}}(\mathbf{k}_\mathbf{m})$ to be determined:

$$\mathbf{E}_{\text{PER}}(\mathbf{r}) = \sum_{\mathbf{m} \in \mathbb{Z}^3} \tilde{\mathbf{E}}_{\text{PER}}(\mathbf{k}_\mathbf{m}) \exp(\mathrm{i}\mathbf{k}_\mathbf{m} \cdot \mathbf{r}), \quad \mathbf{k}_\mathbf{m} = \frac{2\pi}{a}\mathbf{m} \equiv \frac{2\pi}{a}(m_x, m_y, m_z) \tag{7.187}$$

with integers m_x, m_y, m_z. The full field $\mathbf{E}$ is obtained by multiplying $\mathbf{E}_{\mathrm{PER}}$ with the Bloch exponential:

$$\mathbf{E}(\mathbf{r}) \;=\; \mathbf{E}_{\mathrm{PER}}(\mathbf{r})\exp(-\mathrm{i}\mathbf{K}\cdot\mathbf{r}) \;=\; \sum_{\mathbf{m}\in\mathbb{Z}^3} \tilde{\mathbf{E}}_{\mathrm{PER}}(\mathbf{m})\exp\left(\mathrm{i}(\mathbf{k_m}-\mathbf{K})\cdot\mathbf{r}\right) \tag{7.188}$$

The dielectric permittivity $\epsilon = \epsilon(x, y, z)$ or its inverse $\gamma = \epsilon^{-1}$ are also periodic functions of coordinates and can be expanded into similar Fourier series. For the E-problem (7.183), there is, as in 2D, a trade-off between a *generalized* Hermitian problem and a regular non-Hermitian one. The latter is obtained if the equation for the E-field is divided through by ϵ, so that the ω-term (= the right hand side) of the eigenvalue problem does not contain any coordinate-dependent functions:

$$\gamma\nabla\times\mu^{-1}\nabla\times\mathbf{E} \;=\; \omega^2\mathbf{E} \tag{7.189}$$

For the E-field in the Bloch–Floquet form (7.188), the curl operator translates in the Fourier domain into vector multiplication $\mathrm{i}(\mathbf{k_m}-\mathbf{K})\times$. Materials are assumed nonmagnetic, so the permeability μ is constant and equal to μ_0. Multiplication by γ turns into convolution. Overall, the Fourier transformation of the differential equation is similar to the 2D case. The eigenvalue problem for the Fourier coefficients is (see e.g. K. Sakoda [Sak05])

$$-\sum_{\mathbf{s}\in\mathbb{Z}^3}\tilde{\gamma}(\mathbf{m}-\mathbf{s})\,(\mathbf{k_s}-\mathbf{K})\times[(\mathbf{k_s}-\mathbf{K})\times\tilde{\mathbf{E}}(\mathbf{s})] \;=\; \omega^2\mu\tilde{\mathbf{E}}(\mathbf{m}); \tag{7.190}$$

$$\mathbf{m} = (m_x, m_y, m_z); \quad m_x, m_y, m_z = 0, \pm1, \pm2, \ldots$$

where the Fourier coefficients $\tilde{\gamma}$ are

$$\tilde{\gamma}(\mathbf{m}) \;=\; \int_\Omega \gamma(\mathbf{r})\,\exp(-\mathrm{i}\mathbf{k_m}\cdot\mathbf{r})\,dx\,dy\,dz \tag{7.191}$$

In practice, the infinite set of equations (7.190) is truncated and the resultant eigenvalue problem for a finite set of coefficients is solved by direct or iterative methods (Appendix 7.15). If M reciprocal (Fourier) vectors $\mathbf{k_m}$ are retained, the system comprises M vector equations or equivalently $3M$ scalar ones; consequently there are $3M$ Bloch–Floquet modes.

An undesirable feature of the E-formulation is the presence of static eigenmodes ($\omega = 0$) that for purposes of wave analysis in photonics can be considered spurious. These static modes are gradients of scalar potentials $\exp(\mathrm{i}(\mathbf{k_m}-\mathbf{K})\cdot\mathbf{r})$. Indeed, these gradients satisfy (in a trivial way) the curl–curl Maxwell equation (7.183) as well as the Bloch–Floquet boundary conditions on the cell. The number of these static modes is M, out of the $3M$ vector modes.

In the H-formulation (7.184), these electrostatic modes can be eliminated from the outset by employing only transverse waves as a basis:

$$\mathbf{H} \;=\; \sum_{\mathbf{m}\in\mathbb{Z}^3}\tilde{\mathbf{H}}(\mathbf{k_m})\exp(\mathrm{i}(\mathbf{k_m}-\mathbf{K})\cdot\mathbf{r}), \quad \tilde{\mathbf{H}}(\mathbf{k_m})\cdot(\mathbf{k_m}-\mathbf{K}) = 0, \tag{7.192}$$

$$\mathbf{k_m} = \frac{2\pi}{a}\,\mathbf{m} \equiv \frac{2\pi}{a}\,(m_x, m_y, m_z)$$

The transversality condition $\tilde{\mathbf{H}}(\mathbf{k_m}) \perp (\mathbf{k_m} - \mathbf{K})$ eliminates the electrostatic modes because those would be longitudinal (field in the direction of the wave vector):

$$\nabla \exp(\mathrm{i}(\mathbf{k_m} - \mathbf{K}) \cdot \mathbf{r}) = \mathrm{i}(\mathbf{k_m} - \mathbf{K}) \exp(\mathrm{i}(\mathbf{k_m} - \mathbf{K}) \cdot \mathbf{r})$$

No longitudinal H-modes exist because $\nabla \cdot \mathbf{H} = 0$. The absence of these spurious static modes makes the H-field expansion substantially different from that of the E-field. The dimension of the system is reduced from $3M$ to $2M$: each wave vector $\mathbf{k_m}$ has two associated plane waves, with two independent directions of the H-field perpendicular to $(\mathbf{k_m} - \mathbf{K})$.

Another important advantage of the H-formulation in the lossless case (real γ) is that its differential operator, $\nabla \times \gamma(x, y, z)\nabla\times$ is Hermitian,[22] unlike the operator $\gamma(x, y, z)\nabla \times \nabla\times$ of the E-formulation. This is completely analogous to the two-dimensional case and can be verified using integration by parts. In Fourier space, the corresponding problem is also Hermitian. Real-space operations in the differential equation are translated into reciprocal space in the usual manner ($\nabla\times \to \mathrm{i}(\mathbf{k_m} - \mathbf{K})\times$, multiplication $\to$ convolution), and the eigenvalue equations for the H-formulation become

$$-\sum_{\mathbf{s}\in\mathbb{Z}^3} \tilde{\gamma}(\mathbf{m} - \mathbf{s})\,(\mathbf{k_m} - \mathbf{K}) \times [(\mathbf{k_s} - \mathbf{K}) \times \tilde{\mathbf{H}}(\mathbf{s})] = \omega^2 \mu \tilde{\mathbf{H}}(\mathbf{m}); \qquad (7.193)$$

$$\mathbf{m} = (m_x, m_y, m_z); \quad m_x, m_y, m_z = 0, \pm 1, \pm 2, \ldots$$

A small but significant difference from the E-formulation is that the wave vector in the first cross-product now corresponds to the equation index $\mathbf{m}$ rather than the dummy summation index $\mathbf{s}$; this reflects the interchanged order of operations, $\nabla \times \gamma\times$ rather than $\gamma\nabla \times \times$ and makes the system matrix in the Fourier domain Hermitian.

Although the $\mathbf{E}$ and $\mathbf{H}$ fields appear in Maxwell's equations in a perfectly symmetric way (at least in the absence of given electric currents), the E- and H-formulations for the photonic bandgap problem are not equivalent as we have seen. The symmetry between the formulations is broken due to the different behavior of the dielectric permittivity and magnetic permeability: while μ at optical frequencies is essentially equal to μ_0, ϵ is a function of coordinates. This disparity works in favor of the formulation where ϵ appears in the differential operator and the term with the eigenfrequency ω does not contain coordinate-dependent factors.

[22] All operators are considered in the space of functions satisfying the Bloch–Floquet boundary conditions. The permittivity tensor is assumed to be symmetric.

7.10.2 FEM for Photonic Bandgap Problems in 3D

As in 2D (Section 7.9.4), the Finite Element Method can be applied either to the full $\mathbf{E}$ field (or, alternatively, $\mathbf{H}$-field) or to the spatial-periodic factor $\mathbf{E}_{\mathrm{PER}}$ (or $\mathbf{H}_{\mathrm{PER}}$). In the first case, one deals with the usual differential operator but somewhat unusual for FEM boundary conditions (Bloch–Floquet); the second case has standard periodic boundary conditions but an unusual operator. This second case is considered rigorously by D.C. Dobson & J.E. Pasciak in a terse but mathematically comprehensive paper [DP01]. As an alternative, and in parallel with Section 7.9.4, we now review the first formulation.

A natural functional space $\mathcal{B}(\Omega)$ for this problem is the subspace of "scaled-periodic" functions – not in the Sobolev space $H^1(\Omega)$ as in 2D but rather in $H(\mathrm{curl}, \Omega)$:

$$\mathcal{B}(\mathrm{curl}, \Omega) \;=\; \{\mathbf{E}: \; \mathbf{E} \in H(\mathrm{curl}, \Omega); \\ \mathbf{E} \times \mathbf{n} \text{ satisfies Bloch} - \text{Floquet boundary conditions with wave vector } \mathbf{K}\} \tag{7.194}$$

$H(\mathrm{curl}, \Omega)$ is the space of vector functions in $(L_2(\Omega))^3$ whose curl is also in $(L_2(\Omega))^3$; the tangential component $\mathbf{E} \times \mathbf{n}$ of vector fields in this space is mathematically well defined. The $\mathcal{B}$ space depends on the given value of $\mathbf{K}$, although for simplicity of notation this is not explicitly indicated. At this book's level of rigor, the technical details of this definition will not be required; for the interested reader, an excellent mathematical reference is the monograph by P. Monk [Mon03] that is also very useful in connection with edge element formulations.

The weak form of the H-field problem is

$$\text{Find } \mathbf{H} \in \mathcal{B}(\mathrm{curl}, \Omega): \; (\gamma \nabla \times \mathbf{H}, \nabla \times \mathbf{H}') \;=\; \omega^2 \mu \; (\mathbf{H}, \mathbf{H}') \,, \quad \forall \mathbf{H}' \in \mathcal{B}(\mathrm{curl}, \Omega) \tag{7.195}$$

The surface integral in the derivation of the weak formulation vanishes for the same reason as in 2D (Remark 25 on p. 394).

Since the early 1980's, thanks to the work by J.C Nédélec [Né80, Né86], A. Bossavit [BV82, BV83, Bos88b, Bos88a, Bos98], R. Kotiuga [Kot85], D. Boffi [BFea99, Bof01], P. Monk [MD01, Mon03], and many others, the mathematical and engineering research communities have come to realize that the "right" FE discretization of electromagentic vector fields is via "edge elements," where the degrees of freedom are associated with the element edges rather than nodes. For eigenvalue problems, the use of edge elements is particularly important, because they, in contrast with nodal elements, do not produce spurious (nonphysical) modes; see Section 3.12.1, p. 139.

Further details and references on the edge element formulation are given in Chapter 3. From the finite element perspective, the only nonstandard feature of the problem at hand is the Bloch boundary condition. It is dealt with in full analogy with the scalar case in 2D (Section 7.9.4), with "master–slave" edge pairs instead of node pairs.

7.10.3 Historical Notes on the Photonic Bandgap Problem

It is well known that the seminal papers by E. Yablonovitch [Yab87, YG89], S. John [Joh87] and K.M. Ho *et al.* [HCS90] led to an explosion of interest in photonic bandgap structures. An earlier body of work, dating back to at least 1972, is not, however, known nearly as widely. The 1972 and 1975 papers by V.P. Bykov [Byk72, Byk75] (see also [Byk93]), originally published in Russian behind the Iron Curtain, were perhaps ahead of their time. A. Moroz on his website gives a condensed but informative review of the early history of photonic bandgap research.[23] The following excerpts from the website and the original papers speak for themselves.

> A. Moroz: "A study of wave propagation in periodic structures has a long history, which stretches back to, at least, Lord Rayleigh classical article on the influence of obstacles arranged in rectangular order upon the properties of a medium.[24] ... Later on, wave propagation in periodic structures was a subject of the book [BP53] ... by Brillouin and Parodi. ... Some of early history of acoustic and photonic crystals can also be found in a review [Kor94] by Korringa.
> A detailed investigation of the effect of a photonic band gap on the spontaneous emission (SE) of embedded atoms and molecules has been performed by V.P. Bykov [Byk72, Byk75]. For a toy one-dimensional model, he obtained the energy and the decay law of the excited state with transition frequency in the photonic band gap, and calculated the spectrum which accompanies this decay. Bykov's detailed analytic investigation revealed that the SE can be strongly suppressed in volumes much greater than the wavelength."

> V.P. Bykov ([Byk75], "Discussion of Results," p. 871): "The most interesting qualitative conclusion is the possibility of influencing the spontaneous emission and, particularly, suppressing it in large volumes. ... in a large volume we can use a periodic structure and thus control the spontaneous emission.
> Control of the spontaneous emission and particularly its suppression may be important in lasers. For example, the active medium of a laser may have a three-dimensional periodic structure. Let us assume that this structure has such anisotropic properties that at the transition frequency of a molecule there is a narrow cone of directions in which the propagation of electromagnetic waves is allowed, whereas all the other directions are forbidden. Then, the laser threshold of this medium (in the allowed direction) should be much lower than that of a medium without a periodic structure ..."

[23] http://www.wave-scattering.com/pbgheadlines.html and .../pbgprehistory.html
[24] Lord Rayleigh, On the influence of obstacles arranged in rectangular order upon the properties of a medium, *Philos. Mag.* 34, 481-502 (1892).

7.11 Negative Permittivity and Plasmonic Effects

The linear model of constitutive relationships between the electric field $\mathbf{E}$, the polarization $\mathbf{P}$ (= dipole moment per unit volume) and the displacement vector $\mathbf{D}$. Namely,

$$\mathbf{P} = \epsilon_0 \chi \mathbf{E} \tag{7.196}$$

$$\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P} = \epsilon \mathbf{E}, \quad \epsilon = \epsilon_0 (1 + \chi) \tag{7.197}$$

Normally, the dielectric susceptibility χ is nonnegative and the permittivity $\epsilon \geq \epsilon_0$. This section, however, is concerned with a special but exceptionally interesting case where the complex dielectric constant can have a negative real part. How is that possible?

A well known phenomenological description of polarization is obtained by applying Newton's equation of motion to an individual electron in the medium:

$$m\ddot{\mathbf{r}} + m\Gamma\dot{\mathbf{r}} + m\omega_0^2 \dot{\mathbf{r}} = -e\mathbf{E}(t) \tag{7.198}$$

The mass of the electron is m and its charge is $-e$; $\mathbf{r}$ is the position vector; Γ is a phenomenological damping constant that can physically be interpreted as the rate of collisions – the reciprocal of the mean time between collisions. For electrons bound to atoms, the third term in the left hand side represents the restoring force with the "spring constant" $m\omega_0^2$; if the electrons are not bound (e.g. in metals), $\omega_0 = 0$.

For time-harmonic excitation $\mathbf{E}(t) = \mathbf{E}_0 \exp(\mathrm{i}\omega t)$, one solves Newton's equation (7.198) by switching to complex phasors:[25]

$$\mathbf{r} = -\frac{e\mathbf{E}_0}{m} \frac{1}{\omega_0^2 - \omega^2 + \mathrm{i}\omega\Gamma} \tag{7.199}$$

where the same symbols are used for complex phasors as for time functions, with little possibility of confusion.

By definition, polarization (dipole moment per unit volume) is $\mathbf{P} = -N_e e\mathbf{r}$, where N_e is the volume concentration of the electrons,[26] and hence

$$\mathbf{P} = \frac{N_e e^2 \mathbf{E}_0}{m} \frac{1}{\omega_0^2 - \omega^2 + \mathrm{i}\omega\Gamma} \tag{7.200}$$

The dielectric susceptibility is thus

$$\chi = \frac{\omega_p^2}{\omega_0^2 - \omega^2 + \mathrm{i}\omega\Gamma}, \quad \omega_p^2 = \frac{N_e e^2}{\epsilon_0 m} \tag{7.201}$$

Parameter ω_p is called the *plasma frequency.*

[25] The $\exp(+\mathrm{i}\omega t)$ pahsors are used. The $\exp(-\mathrm{i}\omega t)$ convention would lead to the opposite sign of the terms containing odd powers of ω. See also p. 352.

[26] Averaging over $\mathbf{r}$ for all electrons is implied and for simplicity omitted in the expressions.

This phenomenological description of polarization is known as the *Lorentz model*. Of most interest to us in this section is the *Drude model*, where $\omega_0 = 0$ (typical for metals) and the susceptibility becomes

$$\chi = \frac{\omega_p^2}{-\omega^2 + \mathrm{i}\omega\Gamma} \tag{7.202}$$

The relative dielectric constant is

$$\epsilon_r = 1 + \chi = \left(1 - \frac{\omega_p^2}{\Gamma^2 + \omega^2}\right) - \mathrm{i}\,\omega_p^2 \frac{\Gamma/\omega}{\Gamma^2 + \omega^2} \tag{7.203}$$

A peculiar feature of this result is the behavior of the real part of ϵ_r (expression in the large brackets). For frequencies ω below the plasma frequency (more precisely, for $\omega^2 < \omega_p^2 - \Gamma^2$) the real part of the dielectric constant is *negative* – in stark contrast with the normal values greater than one for simple dielectrics.

The negative permittivity is, in the Drude model, ultimately due to the fact that for $\omega_0 = 0$ (no restoring force on the electrons) and sufficiently small damping forces, Newton's law (7.199) puts acceleration – rather than displacement – in sync with the applied electrostatic force. Acceleration, being the second derivative of the displacement, is shifted by 180° relative to the displacement. Therefore displacement, and hence polarization, are shifted by approximately 180° with respect to the applied force, leading to negative susceptibility. For frequencies below the plasma frequency, the real part of susceptibility is even less than -1, which makes the real part of the dielectric constant negative.

Why would anyone care about negative permittivity? As we shall see shortly, it opens many interesting opportunities in *subwalength* optics, with far-reaching practical implications: strong resonances, with very high local enhancement of optical fields and signals; nano-focusing of light; propagation of surface plasmon polaritons (charge density waves on metal–dielectric interfaces), anomalous transmission of light through arrays of holes, and so on. This area of research and development – now one of the hottest in applied physics – is known as *plasmonics*; see U. Kreibig & M. Vollmer [KV95], S.A. Maier & H.A. Atwater [MA05], S.A. Maier [Mai07].[27]

Also associated with negative permittivity is the superlensing effect of metal nanolayers (J.B. Pendry [Pen00], N. Fang *et al.* [FLSZ05], D.O.S. Melville & R.J. Blaikie [MB05]). These subjects are discussed later in this chapter.

[27] Mark Brongersma from Stanford University discovered what almost certainly would be the first paper on the subject of plasmonics; it dates back to 1972. Unfortunately for the physicists, the article is in fact devoted to communication by fish (M.D. Moffler, Plasmonics: Communication by radio waves as found in Elasmobranchii and Teleostii fishes, *Hydrobiologia*, vol. 40 (1), pp. 131–143, 1972, http://www.springerlink.com/content/t103277051264440). Intriguingly, the author discovered "the phenomenon of fish communication, via hydronic radio waves" that are "neither sonic nor electrical".

7.11.1 Electrostatic Resonances for Spherical Particles

Exhibit #1 for electrostatic resonances[28] is the classic example of the electrostatic field distribution around a dielectric spherical particle immersed in a uniform external field. The electrostatic potential can easily be found via spherical harmonics. In fact, since the uniform field (say, in the z-direction) has only one dipole harmonic ($u = -E_0 z = -E_0 r\cos\theta$, in the usual notation), the solution also contains only the dipole harmonic. However, later on in this section higher order harmonics will also be needed, and so for the sake of generality let us recall the expansion of the potential into an infinite series of harmonics.

The potential inside the particle is (e.g. W.K.H. Panofsky & M. Phillips [PP62], R.F. Harrington [Har01] or W.B. Smythe [Smy89]) is

$$u_{\rm in}(r,\theta,\phi) \;=\; \sum_{n=0}^{\infty}\sum_{m=-n}^{n} a_{nm} r^n P_n^m(\cos\theta)\exp({\rm i}m\phi) \tag{7.204}$$

where the standard notation for the associated Legendre polynomials P_n^m and the spherical angles θ, ϕ is used; a_{nm} are some coefficients. The potential outside, in the presence of the applied field E_0 in the z-direction, is

$$u_{\rm out}(r,\theta,\phi) \;=\; -\,E_0 z \;+\; \sum_{n=0}^{\infty}\sum_{m=-n}^{n} b_{nm} r^{-n-1} P_n^m(\cos\theta)\exp({\rm i}m\phi) \tag{7.205}$$

The coefficients a_{nm} and b_{nm} for the field inside/outside are related via the boundary conditions on the surface of the particle:

$$u_{\rm in}(r_{\rm p},\theta,\phi) \;=\; u_{\rm out}(r_{\rm p},\theta,\phi); \quad \epsilon_{\rm in}\,\frac{\partial u_{\rm in}(r_{\rm p},\theta,\phi)}{\partial r} \;=\; \epsilon_{\rm out}\,\frac{\partial u_{\rm out}(r_{\rm p},\theta,\phi)}{\partial r} \tag{7.206}$$

Substitution of harmonic expansions (7.204), (7.205) into these boundary conditions yields a system of decoupled equations for each spherical harmonic. For the special case $n = 1$, $m = 0$ (the dipole term), noting the contribution of the applied field $-E_0 r\cos\theta \equiv -E_0 r P_1(\cos\theta)$, we obtain

$$a_{10} r_{\rm p} \;=\; b_{10} r_{\rm p}^{-2} \tag{7.207}$$

$$\epsilon_{\rm in} a_{10} \;=\; -2\epsilon_{\rm out}(b_{10} r_{\rm p}^{-3} \;+\; E_0) \tag{7.208}$$

where $\epsilon_{\rm in}$, $\epsilon_{\rm out}$ are the dielectric constants of the media inside and outside the particle, respectively. The Legendre polynomials have disappeared because they are the same in all terms. The coefficients a_{10}, b_{10} are easily found from this system:

$$a_{10} \;=\; -\,\frac{3\epsilon_{\rm out}}{2\epsilon_{\rm out}+\epsilon_{\rm in}}\,E_0, \quad b_{10} \;=\; -\,\frac{\epsilon_{\rm out}-\epsilon_{\rm in}}{2\epsilon_{\rm out}+\epsilon_{\rm in}}\,r_{\rm p}^3\,E_0 \tag{7.209}$$

[28] I thank Isaak Mayergoyz for introducing the term "electrostatic resonances" to me; I believe he coined this term.

This result is very well known [Har01, Smy89, PP62]. The dipole moment of the particle is $\mathbf{p} = -b_{10}\hat{z}$, where $\hat{z}$ is the unit vector in the z-direction, and the polarizability (dipole moment per unit applied field) is

$$\alpha = \frac{\epsilon_{\text{out}} - \epsilon_{\text{in}}}{2\epsilon_{\text{out}} + \epsilon_{\text{in}}} r_{\text{p}}^3 \tag{7.210}$$

For simple dielectrics with the dielectric constant greater or equal that of a vacuum, there is nothing unusual about this formula. However, if the permittivity can be negative, as in the quasi-static regime for metals at frequencies below the plasma frequency, the denominator of (7.210) can approach zero. The obvious special case – the *plasmon resonance* condition – for a spherical particle is

$$\epsilon_{\text{in}} = -2\epsilon_{\text{out}} \tag{7.211}$$

If the relative permittivity of the outside medium is unity (air or vacuum), then the resonance occurs for the relative permittivity of the particle equal to -2. Notably, this resonance condition does not depend on the size of the particle – as long as this size remains sufficiently small for the electrostatic approximation to be valid. This size independence turns out to be true for any shapes, not necessarily spherical.

It is worth repeating that although plasmon resonance phenomena usually manifest themselves at optical frequencies, they are to a large extent quasi-static effects – the limiting case for particles much smaller than the wavelength; see U. Kreibig & M. Vollmer [KV95] and D.R. Fredkin & I.D. Mayergoyz [FM03, MFZ05a].

However, while the electrostatic picture is relatively simple and qualitatively correct, full wave simulation is needed for higher accuracy (see Section 7.12.3). From the analytical viewpoint, the field can be expanded into an asymptotic series with respect to the small parameter – the size of the particle relative to the wavelength [MFZ05a], the zeroth term of this expansion being the electrostatic problem.

At the resonance, division by zero in the expression for polarizability (7.210) and in similar expressions for the dipole moment and field indicates a nonphysical situation. In reality, losses (represented in our model by the imaginary part of the permittivity), nonlinearities and dephasing/retardation will quench the singularity.

Under the electrostatic approximation, a *source-free* field can exist if losses are neglected. In the case of a spherical particle, the boundary conditions for any spherical harmonic n, m (not necessarily dipole) are

$$a_{nm} r_{\text{p}}^{n} = b_{nm} r_{\text{p}}^{-n-1} \tag{7.212}$$

$$n\,\epsilon_{\text{in}} a_{nm} r_{\text{p}}^{n-1} = -(n+1)\,\epsilon_{\text{out}} b_{nm} r_{\text{p}}^{-n-2} \tag{7.213}$$

It is straightforward to find that this system of two equations has a nontrivial solution a_{nm}, b_{nm} if the permittivity of the particle is

$$\epsilon_{\text{in}} \;=\; -\,\frac{n+1}{n}\,\epsilon_{\text{out}} \tag{7.214}$$

In particular, for $n = 1$ this is the already familiar condition $\epsilon_{\text{in}} = -2\,\epsilon_{\text{out}}$.

The resonance permittivity is different for particles of different shape; although no simple closed-form expression for this resonance value exists in general, theoretical and numerical considerations for finding it are presented in the following sections. Computing plasmon resonances is of great practical importance due to a variety of applications ranging from nano-optics to nanosensors to biolabels; see S.A. Maier & H.A. Atwater's review of plasmonics [MA05].

7.11.2 Plasmon Resonances: Electrostatic Approximation

If the characteristic dimension of the system under consideration (e.g. the size of a plasmonic particle) is small relative to the wavelength, analysis can be simplified dramatically by electrostatic approximation – the zero-order term in the asymptotic expansion of the solution with respect to the characteristic size (see the previous section).

The governing equation for the electrostatic potential u is

$$\nabla \cdot \epsilon \nabla u \;=\; 0; \qquad u(\infty) = 0 \tag{7.215}$$

An unusual feature here is the zero right hand side of the equation, along with the zero boundary condition. Normally this would yield only a trivial solution: the operator in the left hand side is self-adjoint and, if the dielectric constant has a positive lower bound, $\epsilon(x, y, z) \geq \epsilon_{\text{min}} > 0$, positive definite. More generally, however, the dielectric constant can be complex, so the operator is no longer positive definite and for a real negative permittivity can have a nontrivial null space. This is the plasmon resonance case that we have already observed for spherical particles.

To study plasmonic resonances, let us revisit the formulation of the problem in the electrostatic limit. Since the dielectric constant need not be smooth (it is often piecewise-constant, with jumps at material interfaces), the derivatives in the differential equation (7.215) are to be understood in the generalized sense. It is therefore helpful to write the equation in the weak form:

$$(\epsilon \nabla u,\, \nabla u')_{L^3_2(\mathbb{R}^3)} \;=\; 0; \qquad \forall u' \in H^1(\mathbb{R}^3) \tag{7.216}$$

In contrast with standard electrostatics, for complex ϵ this bilinear form is not in general elliptic. Importantly, ϵ can be (at least approximately) real and negative in some regions, and this equation can therefore admit nontrivial solutions.

To make further progress in the analysis, let us consider a specific case of great practical interest: region(s) Ω_{p} with one dielectric constant ϵ_{p} (particles, particle clusters, layers, etc.) embedded in some "background" medium with

another dielectric constant $\epsilon_{\rm bg} \neq \epsilon_{\rm p}$. It is assumed that $\epsilon_{\rm p}$ and $\epsilon_{\rm bg}$ do not depend on coordinates. The weak form of the governing equation can then be rewritten as

$$\epsilon_{\rm bg}(\nabla u,\, \nabla u')_{L_2^3(\mathbb{R}^3)} \;+\; (\epsilon_{\rm p} - \epsilon_{\rm bg})(\nabla u,\, \nabla u')_{L_2^3(\Omega_{\rm p})} \;=\; 0; \qquad \forall u' \in H^1(\mathbb{R}^3) \tag{7.217}$$

or equivalently

$$(\nabla u,\, \nabla u')_{L_2^3(\Omega_{\rm p})} \;=\; \lambda(\nabla u,\, \nabla u')_{L_2^3(\mathbb{R}^3)}; \;\; \forall u' \in H^1(\mathbb{R}^3), \tag{7.218}$$

$$\lambda \;=\; \frac{\epsilon_{\rm bg}}{\epsilon_{\rm bg} \;-\; \epsilon_{\rm p}} \tag{7.219}$$

This is a generalized eigenvalue problem. Setting $u' = u$ reveals that all eigenvalues λ must lie in the closed interval $[0,\, 1]$. Indeed, both inner products with $u' = u$ are always real and nonnegative; the inner product over $\Omega_{\rm p}$ obviously cannot exceed the one over the whole $\mathbb{R}^3$. Thus we have

$$0 \;\leq\; \frac{\epsilon_{\rm bg}}{\epsilon_{\rm bg} - \epsilon_{\rm p}} \;\leq\; 1$$

and consequently

$$\frac{\epsilon_{\rm p}}{\epsilon_{\rm bg}} \;<\; 0 \tag{7.220}$$

This result again highlights the key role of negative permittivity – without that the resonance, in the strict sense of the word (the presence of a source-free eigenmode), is not possible. If the dielectric constant of the particle is close but not exactly equal to its resonance value (e.g. $\epsilon_{\rm p}$ has a non-negligible imaginary part), one can expect strong local amplification of applied external fields in the vicinity of the particle,[29] giving rise to many practical applications.

To find the actual numerical values of the eigenparameter λ in (7.218) – and hence the corresponding value of the dielectric constant – one can discretize the problem using finite element analysis, finite differences (K. Li *et al.* [LSB03]), integral equation methods (D.R. Fredkin & I.D. Mayergoyz [FM03, MFZ05a], T-matrix methods and other techniques. It goes without saying that the plasmon modes and their spectrum do not depend on a specific formulation of the problem or on a specific method of solving it. In particular, regardless of the formulation, the problem with two media (e.g. host and particles) splits up into a purely "geometric" eigenproblem (7.218) with no material parameters and the relationship (7.219) between the eigenvalue λ and the permittivity ϵ.

[29] Unless the external field happens to be orthogonal to the respective resonance eigenmode.

7.11.3 Wave Analysis of Plasmonic Systems

Although the electrostatic approximation does provide a very useful insight into plasmon resonance phenomena, accurate evaluation of resonance conditions and field enhancement requires electromagnetic wave analysis. Effective material parameters ϵ and μ are needed for Maxwell's equations, but questions do arise about the applicability of bulk permittivity to nanoparticles.

Various physical mechanisms affecting the value of the effective dielectric constant in individual nanoparticles and in particle clusters are discussed in detail in the physics literature: U. Kreibig & C. Von Fragstein [Fra69], U. Kreibig & M. Vollmer [KV95], A. Liebsch [Lie93a, Lie93b], B. Palpant *et al.* [PPL+98], M. Quinten [Qui96, Qui99], L.B. Scaffardi & J.O. Tocho [ST06]. As an example of such complicated physical phenomena, at the surfaces of silver particles due to quantum effects the 5s electron density "spills out" into the vacuum, where 5s electronic oscillations are not screened by the 4d electrons [Lie93a, Lie93b]. Further, for small particles the damping constant Γ in the Drude model is increased due to additional collisions of free electrons with the boundary of the particle [Fra69, KV95]; Scaffardi & Tocho [ST06] and Quinten [Qui96] provide the following approximation;

$$\Gamma \;=\; \Gamma_{\text{bulk}} \;+\; C\,\frac{v_F}{r_{\text{p}}}$$

where v_F is the electron velocity at the Fermi surface and r_{p} is the radius of the particle ($v_F \sim 14.1 \cdot 10^{14}$ nm $\cdot$ s^{-1} for gold, C is on the order of 0.1–2 [ST06]).

Fortunately, the cumulative effect of the nanoscopic factors affecting the value of the permittivity may be relatively mild, as suggested by spectral measurements of plasmon resonances of extremely thin nanoshells by C.L. Nehl *et al.* [NGG+04]: "the resonance line widths fit Mie theory without the inclusion of a size-dependent surface scattering term". Moreover, the measurements by P. Stoller *et al.* [SJS06] show that bulk permittivity is applicable to gold particles as small as 10–15 nm in diameter.

There is a large body of literature on the optical behavior of small particles. In addition to the publications cited above, see M. Kerker *et al.* [KWC80] and K.L. Kelly *et al.* [KCZS03]. In the remainder of this section, our focus is on the computational tools rather than the physics of effective material parameters. Hence these parameters will be considered as given, with an implicit assumption that proper adjustments have been made for the difference between the parameters in the particles and in the bulk. However, it should be kept in mind that such adjustments may not be valid if nonlocal effects of electron charge distribution are appreciable.

7.11.4 Some Common Methods for Plasmon Simulation

This section is a brief summary of computational methods that are frequently used for simulations in plasmonics. In the following sections, two other

computational tools – the generalized finite-difference method with flexible local approximation and the Finite Element Method – are considered in greater detail.

Analytical Solutions

As an analytical problem, scattering of electromagnetic waves from dielectric objects is quite involved. Closed-form solutions are available only for a few cases (see e.g. M.I. Mishchenko *et al.* [MTL02]): an isotropic homogeneous sphere (the classic Lorenz–Mie–Debye case); concentric core-mantle spheres; concentric multilayered spheres; radially inhomogeneous spheres; a homogeneous infinite circular cylinder; an infinite elliptical cylinder; homogeneous and core–mantle spheroids. For objects other than homogeneous spheres or infinite cylinders, the complexity of analytical solutions (if they are available) is so high that the boundary between analytical and numerical methods becomes blurred. At present, further extensions of purely analytical techniques seem unlikely. On the other hand, with the available analytical cases in mind, *local* analytical approximations to the field are substantially easier to construct than global closed-form solutions. Such local analytical approximations can be incorporated into "Flexible Local Approximation Methods" (FLAME), Section 7.11.5 and Chapter 4.

T-matrix Methods

T-matrix methods (M.I. Mishchenko *et al.* [MTM96, MTL02]) are widely used in scattering problems. Mishchenko *et al.* [MVB$^+$04] collected a comprehensive database of references and have developed a T-matrix software package [MT98].

If a monochromatic wave impinges on a scattering dielectric object of arbitrary shape, both the incident and scattered waves can be expanded into spherical harmonics around the scatterer. If the electromagnetic properties of the scatterer (the permittivity and permeability) are linear, then the expansion coefficients of the scattered wave are linearly related to the coefficients of the incident wave. The matrix governing this linear relationship is called the T- ("transition") matrix. For a collection of scattering particles, the overall field can be sought as a superposition of the individual harmonic expansions around each scatterer. The transformation of vector spherical harmonics centered at one particle to harmonics around another one is accomplished via well-established translation and rotation rules (Theorems) (e.g. D.W. Mackowski [Mac91], M.I. Mishchenko *et al.* [MTL02], D.W. Mackowski & M.I. Mishchenko [MM96], Y.-l. Xu [lX95]).

Self-consistency of the multi-centered expansions then leads to a linear system of equations for the expansion coefficients. Since the system matrix is dense, the computational cost may become prohibitively high if the number of scatterers is large. For spherical, spheroidal and other particles that admit

a closed-form solution of the wave problem (see above), the T-matrix can be found analytically. For other shapes, the T-matrix is computed numerically. If the scatterer is homogeneous, the "Extended Boundary Condition Method" (EBCM) (e.g. P. Barber & C. Yeh [BY75], M.I. Mishchenko *et al.* [MTL02]) is usually the method of choice. EBCM is a combination of integral equations for equivalent surface currents and expansions into vector spherical harmonics (R.F. Harrington [Har01] or J.A. Stratton [Str41]). While the T-matrix method is quite suitable for a moderate number of isolated particles and is also very effective for random distributions and orientations of particles (e.g. in atmospheric problems), it is not designed to handle large continuous dielectric regions. It is possible, however, to adapt the method to particles on an infinite substrate at the expense of additional analytical, algorithmic and computational work: plane waves reflected off the substrate are added to the superposition of spherical harmonics scattered from the particles themselves (A. Doicu *et al.* [DEW99], T. Wriedt and A. Doicu [WD00]).

The Multiple Multipole Method

In the Multiple Multipole Method (MMP), the computational domain is decomposed into homogeneous subdomains, and an appropriate analytical expansion – often, a superposition of multipole expansions as the name suggests – is introduced within each of the subdomains. A system of equations for the expansion coefficients is obtained by collocation of the individual expansions at a set of points on subdomain boundaries. Applications of MMP in computational electromagnetics and optics include simulations of plasmon resonances (E. Moreno *et al.* [MEHV02]) and of plasmon-enhanced optical tips (R. Esteban *et al.* [EVK06]).

A shortcoming of MMP is that no general systematic procedure for choosing the centers of the multiple-multipole expansions is available. The choice of expansions remains partly a matter of art and experience, which makes it difficult to evaluate and systematically improve the accuracy and convergence. The MaX platform developed by C. Hafner [Haf99b, Haf99a] has apparently overcome some of the difficulties.

The Discrete-Dipole Method

The Discrete-Dipole Method belongs to the general category of integral equation methods but admits a very simple physical interpretation. Scattering bodies are approximated by a collection of dipoles, each of which is directly related to the local value of the polarization vector. Starting with the volume integral equation for the electric field, one can derive a self-consistent system of equations for the equivalent dipoles (B. Draine & P. Flatau [DF94, DF], P.J. Flatau [Fla97], A. Lakhtakia & G. Mulholland [LM03], J. Peltoniemi [Pel96]).

The method has gained popularity in the simulation of plasmonic particles, as well as other scattering problems, because of its conceptual simplicity, relative ease of use and the availability of public domain software DDSCAT [DF94, DF] by Draine & Flatau. For application examples, see papers by K.L. Kelly *et al.* [KCZS03], M.D. Malinsky *et al.* [MKSD01], K.-H. Su *et al.* [SWZ+03].

DDM has some disadvantages typical for integral-equation methods. First, the treatment of singularities in DDM is quite involved (Lakhtakia & Mulholland [LM03], Peltoniemi [Pel96]). Second, the system matrix for the coupled dipoles is dense, and therefore the computational time increases rapidly with the increasing number of dipoles. If the dipoles are arranged geometrically on a regular grid, the numerical efficiency can be improved by using Fast Fourier Transforms to speed up matrix-vector multiplications in the iterative system solver. However, for such a regular arrangement of the sources DDM shares one additional disadvantage not with integral-equation methods but rather with finite-difference algorithms: a "staircase" representation of curved or slanted material boundaries. In DDM simulations (e.g. N. Félidj *et al.* [FAL99], M.D. Malinsky *et al.* [MKSD01]), there are typically thousands of dipoles in each particle and tens of thousands of dipoles for problems with a few particles on a substrate. As an example, in [MKSD01] 11,218 dipoles are used in the particle and 93,911 dipoles in the particle and substrate together, so that the overall system of equations has a dense matrix of dimension 280,000.

7.11.5 Trefftz–FLAME Simulation of Plasmonic Particles

This section shows an application of generalized finite difference schemes with flexible local approximation (FLAME, Chapter 4) to the computation of electromagnetic waves and plasmon field enhancement around one or several cylindrical rods. The axes of all rods are aligned in the z direction and the field is assumed to be independent of z, so that the computational problem is effectively two-dimensional. Two polarizations can be considered: the E-mode with the E field in the z direction, and the H-mode. (The reason for using this terminology, rather than more common "TE/TM" modes or s/p modes, is explained on p. 385.)

Note that it is in the H-mode (one-component H-field perpendicular to the xy-plane and the electric field in the plane) that the electric field "goes through" the plasmon particles, thereby potentially giving rise to plasmon resonances. The governing equation for the H-mode is:

$$\nabla \cdot (\epsilon^{-1} \nabla H) \; + \; \omega^2 \mu H \; = \; 0 \tag{7.221}$$

In plasmonics, permeability μ can be assumed equal to μ_0 throughout the domain; the permittivity is ϵ_0 in air and has a complex and frequency-dependent value within plasmonic particles. Standard radiation boundary conditions for the scattered wave apply.

One specific problem that will be used here as an illustrative example was proposed by J.P. Kottmann & O.J.F. Martin [KM01] and involves two cylindrical plasmon particles with a small separation between them (Fig. 7.28).

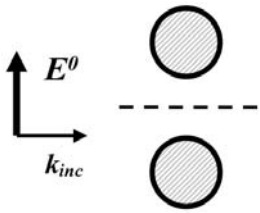


Fig. 7.28. Two cylindrical plasmonic particles. Setup due to Kottmann & Martin [KM01]. (This is one of the two cases they consider.)

Kottmann & Martin used integral equations in their simulation. In this section as an alternative, Trefftz–FLAME schemes of Chapter 4 on a 9-point (3×3) stencil are applied. It is natural to choose the basis functions as cylindrical harmonics in the vicinity of each particle and as plane waves away from the particles. "Vicinity" is defined by an adjustable threshold: $r \leq r_{\text{cutoff}}$, where r is the distance from the midpoint of the stencil to the center of the nearest particle, and the threshold r_{cutoff} is typically chosen as the radius of the particle plus a few grid layers.

Away from the particles, eight basis functions are taken as plane waves propagating toward the central node of the 9-point stencil from each of the other eight nodes

$$\psi_\alpha \;=\; \exp(\mathrm{i}k\,\hat{\mathbf{r}}_\alpha \cdot \mathbf{r}), \qquad \alpha = 1, 2, \ldots, 8, \quad k^2 \;=\; \omega^2 \mu_0 \epsilon_0 \tag{7.222}$$

(see Appendix 4.8 on p. 236).

The 9×8 nodal matrix (4.14) of FLAME comprises the values of the chosen basis functions at the stencil nodes, i.e.

$$N_{\beta\alpha} \;=\; \psi_\alpha(\mathbf{r}_\beta) \;=\; \exp(\mathrm{i}k\,\hat{\mathbf{r}}_\alpha \cdot \mathbf{r}_\beta) \quad \alpha = 1, 2, \ldots, 8; \;\; \beta = 1, 2, \ldots, 9 \tag{7.223}$$

The coefficient vector of the Trefftz–FLAME scheme (Chapter 4) is $\underline{s} = \operatorname{Null} N^T$. Straightforward symbolic algebra computation shows that this null space is indeed of dimension one, so that a single valid Trefftz–FLAME scheme exists (Appendix 4.8).

Substituting the nodal values of a "test" plane wave $\exp(-\mathrm{i}k\hat{\mathbf{r}} \cdot \mathbf{r})$, where $\hat{\mathbf{r}} = \hat{x}\cos\phi + \hat{y}\sin\phi$, into the difference scheme, one obtains, after some additional symbolic algebra manipulation, the consistency error

$$\epsilon_c \;=\; \frac{1}{12096}\,(hk)^6(\cos(\phi) - 1)\cos^2(\phi)(\cos(\phi) + 1)(2\cos^2(\phi) - 1)^2 \tag{7.224}$$

where for simplicity the mesh size h is assumed to be the same in both coordinate directions.

The ϕ-dependent factor has its maximum of $(2-2^{\frac{1}{2}})/8$ at $\cos 2\phi = (\frac{1}{2} + 2^{\frac{1}{2}}/4)^{\frac{1}{2}}$. Hence the consistency error $\epsilon_c \leq (hk)^6(2-2^{\frac{1}{2}})/96,768$ for any "test" plane wave. Since any solution of the Helmholtz equation in the air region can be locally represented as a superposition (Fourier integral) of plane waves, this result for the consistency error has general applicability. Note that by construction the scheme is exact for plane waves propagating in either of the eight special directions (at $\pm 45°$ to the axes if $h_x = h_y = h$). The domain boundary is treated using a FLAME-style PML (Perfectly Matched Layer), as mentioned on p. 218; see also [Tsu05a, Tsu06].

In the vicinity of each particle, the "Trefftz" basis functions satisfying the wave equation are chosen as cylindrical harmonics:

$$\psi_\alpha^{(i)} = \begin{cases} a_n J_n(k_{\text{cyl}} r)\exp(\mathrm{i}n\phi), & r \leq r_0 \\ \left(b_n J_n(k_{\text{air}} r) + H_n^{(2)}(k_{\text{air}} r)\right)\exp(\mathrm{i}n\phi), & r > r_0 \end{cases}$$

where J_n is the Bessel function, $H_n^{(2)}$ is the Hankel function of the second kind [Har01], and a_n, b_n are coefficients to be determined. These coefficients are found via the standard conditions on the particle boundary; the actual expressions for these coefficients are too lengthy to be worth reproducing here but are easily usable in computer codes.

Eight basis functions are obtained by retaining the monopole harmonic ($n = 0$), two harmonics of orders $n = 1, 2, 3$ (i.e. dipole, quadrupole and octupole), and one of harmonics of order $n = 4$. Numerical experiments for scattering from a *single* cylinder, where the analytical solution is available for comparison and verification, show convergence (not just consistency error!) of order six for this scheme [Tsu05a].

In Fig. 7.29, the electric field computed with Trefftz–FLAME is compared with the quasi-analytical solution via the multicenter-multipole expansion of the wave (V. Twersky [Twe52], M.I. Mishchenko *et al.* [MTL02]), for the following parameters.[30]

The radius of each silver nanoparticle is 50 nm. The wavelength of the incident wave varies as labeled in the figure; the complex permittivity of silver at each wavelength is obtained by spline interpolation of the Johnson & Christy values [JC72]. As evident from the figure, the results of FLAME simulation are in excellent agreement with the quasi-analytical computation.

Kottman & Martin applied volume integral equation methods where "the particles are typically discretized with 3000 triangular elements" [KM01]. For two particles, this gives about 6000 unknowns and a full system matrix with 36 million nonzero entries. For comparison, FLAME simulations were run on grids from 100×100 to 250×250 ($\sim$100–500 thousand nonzero entries in a very sparse matrix).

[30] The analytical expansion was implemented by Frantisek Čajko.

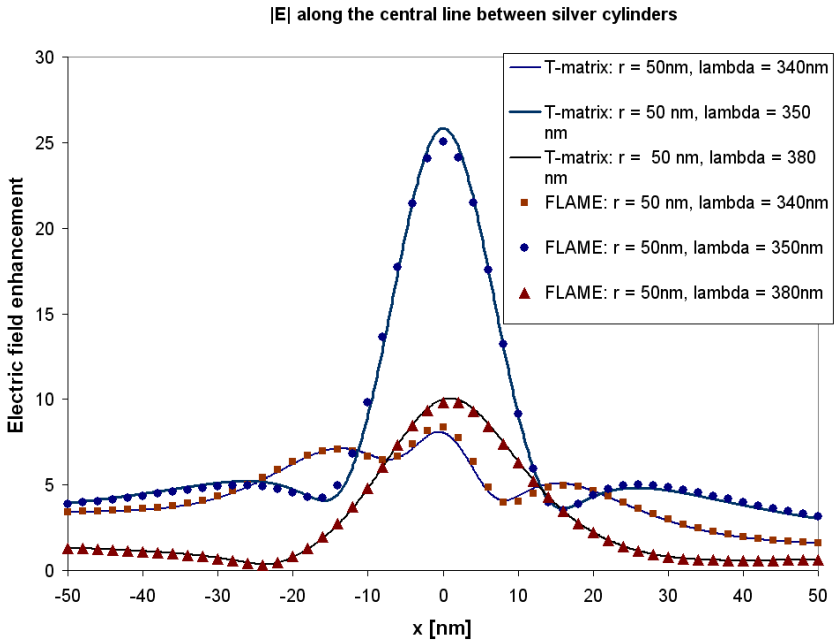


Fig. 7.29. (Credit: F. Čajko.) The magnitude of the electric field along the line connecting two silver plasmonic particles. Comparison of FLAME and multipole-multicenter results. Particle radii 50 nm; varying wavelength of incident light. (Reprinted by permission from [Tsu06] ©2006 Elsevier.)

7.11.6 Finite Element Simulation of Plasmonic Particles

As we have seen, plasmonic resonances of metal particles may lead to very high local enhancement of light. Cascade amplification may produce an even stronger effect.

As an illustration, an interesting self-similar cascade arrangement of particles in 3D, where an extremely high plasmon field enhancement can be achieved, was proposed by K. Li, M.I. Stockman and D. Bergman [LSB03] (Figs. 7.30, 7.31). Three spherical silver particles, with the radii 45, 15 and 5 nm as a characteristic example, are aligned on a straight line; the air gap is 9 nm between the 45 and 15 nm particles, and 3 nm between the 15 and 5 nm particles. Each of the smaller particles is in the field amplified by its bigger neighbor; hence cascade amplification of the field.

The quasi-static approximation of [LSB03] is helpful if the size of the system is much smaller that the wavelength. Electrodynamic effects were reported by another group of researchers (Z. Li *et al.* [LYX06]) to result in correction factors on the order of two for the maximum value of the electric field. However, as K. Li *et al.* argue in [LSB06], the grid size in the finite-difference

time-domain (FDTD) simulation of [LYX06] was too coarse to accurately represent the rapid variation of the field at the focus of the "lens". To analyze the impact of electrodynamic effects on the nano-focusing of the field more accurately, J. Dai *et al.* [DvTS] use adaptive finite element analysis in the frequency domain, which is more straightforward and reliable that reaching the sinusoidal steady state in FDTD.

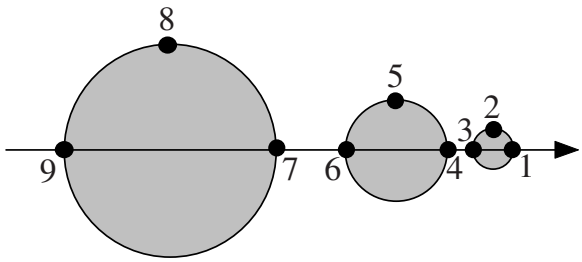


Fig. 7.30. A cascade of three particles and reference points for field enhancement.

Some of the results by J. Dai *et al.* are reported below. It is assumed (as was done in [LSB03]) that, to a reasonable degree of approximation, the permittivity of the particles is equal to its bulk value for silver. As already noted, the optical response of small particles is very difficult to model accurately due to nonlocality, surface roughness, "spillout" of electrons and other factors. Nevertheless the bulk value of the permittivity may still provide a meaningful approximation (p. 423).

Under the electrostatic approximation, the maximum field enhancement in the Li–Stockman–Bergman cascade is calculated to occur in the near ultraviolet at $\hbar\omega = 3.37$ eV, with the corresponding wavelength of ∼367.9 nm in a vacuum and the corresponding frequency ∼814.8 THz. The relative permittivity at this wavelength is, under the $\exp(+\mathrm{i}\omega t)$ phasor convention, $-2.74 - 0.232\,\mathrm{i}$ according to the Johnson & Christy data [JC72].

Electric field $\mathbf{E}$ is governed by the wave equation

$$\nabla \times \mu^{-1} \nabla \times \mathbf{E} \;-\; \omega^2 \epsilon \mathbf{E} \;=\; 0 \tag{7.225}$$

For analysis and simulation – particularly for imposing radiation boundary conditions – it is customary to decompose the total field into the sum of the incident field $\mathbf{E}_{\mathrm{inc}}$ and the scattered field $\mathbf{E}_{\mathrm{s}}$; by definition, $\mathbf{E}_{\mathrm{s}} = \mathbf{E} - \mathbf{E}_{\mathrm{inc}}$. In our simulations, the incident field is always a plane wave with the amplitude of the electric field normalized to unity.

The governing equation for the scattered field is

$$\nabla \times \nabla \times \mathbf{E}_{\mathrm{s}} - \omega^2 \mu_0 \epsilon \mathbf{E}_{\mathrm{s}} \;=\; -(\nabla \times \nabla \times \mathbf{E}_{\mathrm{inc}} - \omega^2 \mu_0 \epsilon \mathbf{E}_{\mathrm{inc}}) \tag{7.226}$$

(for $\mu = \mu_0$ at optical frequencies). The differential operators should be understood in the sense of generalized functions (distributions) that include surface

delta functions for charges and currents (Appendix 6.15 on p. 343). The right hand side of the equation is nonzero due to these surface terms and due to the volume term inside the particles, as the incident field is governed by the wave equation with the wavenumber of free space.

In the electrostatic limit, the governing equation is written for the total electrostatic potential ϕ:

$$\nabla \cdot \epsilon \nabla \phi \;=\; 0; \quad \phi(\mathbf{r}) \to \phi_{\text{ext}}(\mathbf{r}) \;\text{ as } \mathbf{r} \to \infty \tag{7.227}$$

where $\phi_{\text{ext}}(\mathbf{r})$ is the applied potential (typically a linear function of position $\mathbf{r}$, corresponding to a constant external field). The differential operators in (7.227) should again be understood in the generalized sense.

In FEM, (7.226) is rewritten in the weak (variational) form. Boundary conditions on the surfaces are natural – that is, the solution of the variational problem satisfies these conditions automatically. The mathematical and technical details of this approach are very well known (e.g. P. Monk [Mon03], J. Jin [Jin02]). J. Dai *et al.* [DvTS] used the commercial software package HFSS^{TM} by Ansoft Corp. for electrodynamic analysis[31] and FEMLAB^{TM} (COMSOL Multiphysics) in the electrostatic case. Both packages are FEM-based: second-order triangular nodal elements for the electrostatic problem and tetrahedral edge elements with 12 degrees of freedom for wave analysis. HFSS employs automatic adaptive mesh refinement for higher accuracy and either radiation boundary conditions or Perfectly Matched Layers to truncate the unbounded domain.

To assess the numerical accuracy, J. Dai *et al.* first considered a single particle. The average difference between Mie theory [Har01] and HFSS field values is ∼2.3% for a dielectric particle with $\epsilon = 10$ and ∼4.9% for a silver particle with $\epsilon_s = -2.74 - 0.232\,\mathrm{i}$. At the surface of the particle, the computed normal component of the displacement vector, in addition to smooth variation, was affected by some numerical noise. The noise was obvious in the plots and was easily filtered out. The HFSS mesh had 20,746 elements in all simulations.

Let us now turn to the simulations of particle cascades. A sample distribution of the field enhancement factor (i.e. the ratio of the amplitude of the total electric field to the incident field) in the cross-section of the cascade is shown in Fig. 7.31 for illustration; the incident wave is polarized along the axis of the cascade and propagates in the downward direction.

Four independent combinations of the directions of wave propagation and polarization can be considered (left–right and up–down directions are in reference to Fig. 7.30):

1. The incident wave propagates from right to left. Electric and magnetic fields are both perpendicular to the axis of the cascade. (Mnemonic label: $\Leftarrow\perp$.)

[31] Caution should be exercised when representing the measured Johnson & Christy data [JC72], with its $\exp(-\mathrm{i}\omega t)$ convention for phasors, as the HFSS input, with its $\exp(+\mathrm{i}\omega t)$ default.

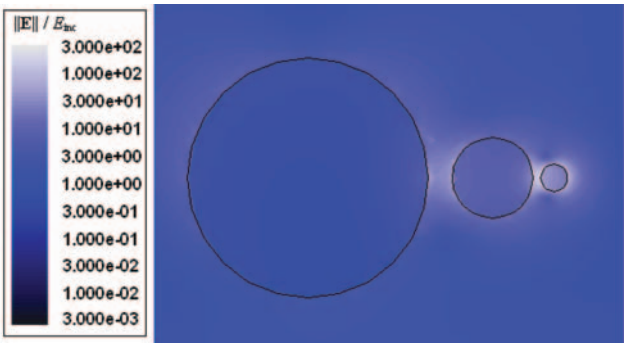


Fig. 7.31. Electric field enhancement factor around the cascade of three plasmonic spheres. (Simulation by J. Dai & F. Čajko.)

2. Same as above, but the wave impinges from the left. ($\Rightarrow \perp$)
3. The direction of propagation and electric field are both perpendicular to the axis of the cascade. ($\Uparrow \perp$)
4. The direction of propagation is perpendicular to the cascade axis and the electric field is parallel to it. ($\Uparrow \|$)

Table 7.2 shows the field enhancement factors at the reference points for cases (i)–(iv) [DvTS]. The "hottest spot," i.e. the point of maximum enhancement, is indicated in bold and is different in different cases. When the electric field is perpendicular to the axis of the cascade, the local field is amplified by a very modest factor $g < 40$. Not surprisingly, enhancement is much greater ($g \approx 205$) in case (iv), when the field and the dipole moments that it induces are aligned along the axis.

Table 7.2. Field enhancement for different directions of propagation and polarization of the incident wave. P1–P9 are the reference points shown in Fig. 7.30. (Simulation by J. Dai & F. Čajko.)

Case	P1	P2	P3	P4	P5	P6	P7	P8	P9
$\Leftarrow\perp$	5.45	17.3	10.2	9.43	**34.4**	10.7	5.53	10.4	3.21
$\Rightarrow\perp$	6.37	6.49	2.41	1.43	4.17	3.39	3.91	**11.2**	2.00
$\Uparrow\perp$	2.44	8.48	6.65	7.60	**23.3**	8.31	4.69	10.1	2.61
$\Uparrow\|$	90.8	35.9	**250**	146	10.3	70.9	51.9	2.72	6.47

To gauge the influence of electrodynamic effects, field enhancement is analyzed as a function of scaling of the system size. Scaling is applied across the board to all dimensions: all the radii of the particles and the air gaps between them are multiplied by the same factor. The radius of the smallest particle,

with its original value [LSB03] of 5 nm as reference, is used as the independent variable for plotting and tabulating the results (Fig. 7.32).

The enhancement factor decreases rapidly as the size of the system increases. This can be easily explained by dephasing effects. Conversely, as the system size is reduced, the local field increases significantly. It is, however, somewhat counterintuitive that the electrostatic limit does *not* produce the highest enhancement factor (Fig. 7.32). Further, the point of maximum enhancement does not necessarily lie on the axis of the cascade. As noted by F. Čajko, some clues can be gleaned by approximating each particle as an equivalent dipole in free space and neglecting higher-order spherical harmonics. The electric field of a Hertzian dipole is given by the textbook formula

$$\begin{aligned}\mathbf{E}_{\text{dip}} &= -\eta_0 \frac{k\omega p \exp(-\mathrm{i}kr)}{4\pi r} \left\{ \hat{\mathbf{r}} \left[\frac{1}{\mathrm{i}kr} + \left(\frac{1}{\mathrm{i}kr} \right)^2 \right] 2\cos\theta \right. \\ &\quad \left. + \hat{\theta} \left[1 + \frac{1}{\mathrm{i}kr} + \left(\frac{1}{\mathrm{i}kr} \right)^2 \right] \sin\theta \right\}, \quad \eta_0 = \left(\frac{\mu_0}{\epsilon_0} \right)^{\frac{1}{2}} \end{aligned} \tag{7.228}$$

where the dipole with moment p is directed along the z-axis of the spherical system (r, θ, ϕ). In the case under consideration, kr is on the order of unity, and no near/far field simplification is made in the formula. Since all dipole moments approximately scale as the cube of a characteristic system size l, the magnitude of the field, say, on the axis $\theta = 0$ behaves as $\propto c_1 + c_2 l^2$ with some positive coefficients $c_{1,2}$. This explains the mild local minimum of the field in the electrostatic limit in Fig. 7.32. Furthermore, since (7.228) includes both $\sin\theta$ and $\cos\theta$ variations, it is clear that the maximum magnitude of the field cannot in general be expected to occur on the axis $\theta = 0$.

To summarize, while electrostatic analysis provides a useful insight into plasmonic field enhancement, electrodynamic effects lead to appreciable corrections. Field enhancement factors on the order of a few hundred by self-similar chains of plasmonic particles may be realizable. Maximum enhancement does not necessarily correspond to polarization along the axis of the cascade and to the electrostatic limit; hence the size of the system is a non-trivial variable in the optimization of optical nano-lenses.

7.12 Plasmonic Enhancement in Scanning Near-Field Optical Microscopy

This section reflects some results of collaborative work with A. P. Sokolov and his group at the Department of Polymer Science, the University of Akron, and with F. Keilmann & R. Hillenbrand's group at the Max-Planck-Institut für Biochemie in Martinsried, Germany. The simulations in this section were performed by F. Čajko.

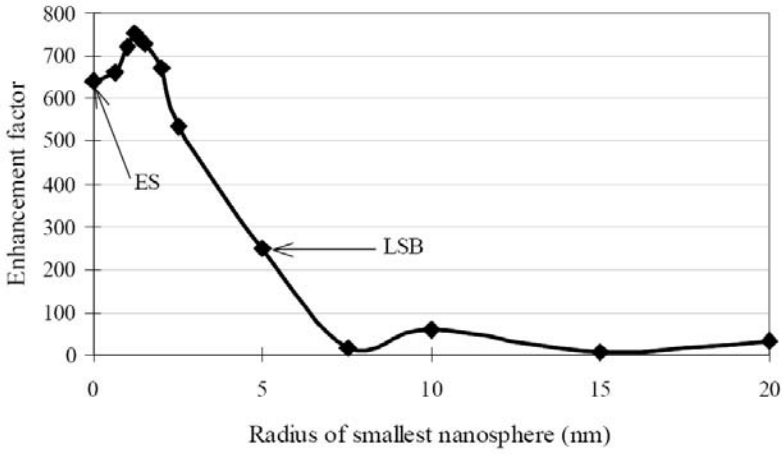


Fig. 7.32. Maximum field enhancement vs. radius of the smallest particle. All dimensions of the system are scaled proportionately. LSB: the specific example by K. Li *et al.* [LSB03], where the radius of the smallest particle is 5 nm. ES: the electrostatic limit. Credit: J. Dai & F. Čajko.

7.12.1 Breaking the Diffraction Limit

As a rule, diffraction constrains the focusing of light and the resolution in optical systems to about one half of the wavelength. While in geometric optics an ideal lens can focus a beam of light to a single point, in reality the focus is smeared to an area on the order of the wavelength in size. The previous section showed, however, that plasmon resonances, especially in particle cascades and clusters, can produce very strong fields in highly localized areas; this can be interpreted as *nano-focusing* or *nano-lensing*.

The diffraction limit is often viewed as a manifestation of the Heisenberg uncertainty principle

$$\Delta y \, \Delta p_y \;\sim\; \frac{\hbar}{2} \tag{7.229}$$

where $\hbar$ is the reduced Planck constant ($\sim 1.05457 \times 10^{-34}$ m$^2 \cdot$ kg/s); Δy, Δp_y are the uncertainties in the position and momentum of a quantum particle (in our case, a photon) along a given direction labeled in the formula as y. A photon with frequency ω arriving at the focus of a lens (Fig. 7.33) has the magnitude of momentum $p = \hbar k = \hbar\, 2\pi/\lambda$, where λ is the wavelength in the medium around the lens. Since the photon can come from any angle θ between some $-\theta_{\max}$ and $+\theta_{\max}$, the uncertainty in the y-component of its momentum is

$$\Delta p_y \;=\; 2p \sin\theta_{\max} \;=\; 4\pi\hbar \sin\theta_{\max}/\lambda$$

and hence the uncertainly in its position is, by the Heisenberg principle,

$$\Delta y \sim \frac{\hbar}{2\Delta p_y} = \frac{\lambda}{8\pi \sin\theta_{\max}} \tag{7.230}$$

Thus the uncertainly principle prohibits ideal focusing of light by a conventional lens.

Fortunately, however, the connection between the uncertainly principle and the diffraction limit is not cut and dried. Contrary to what the lens example may lead us to believe, there appears to be no fundamental theoretical limit on the level of optical resolution – only practical limitations.

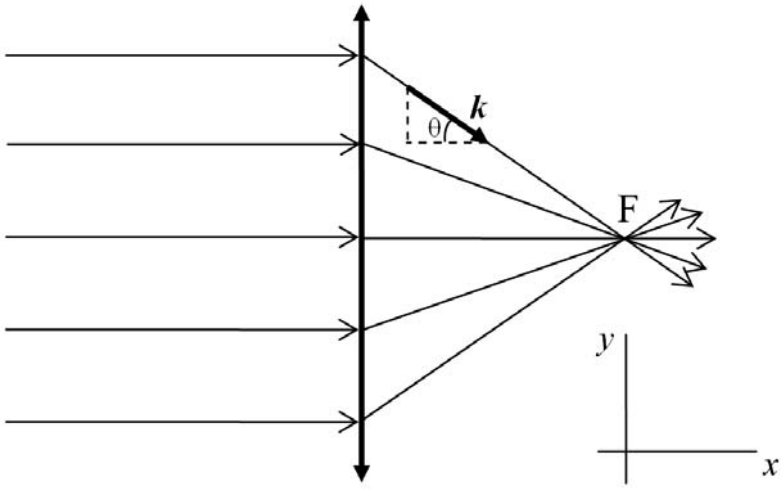


Fig. 7.33. In geometric optics, an ideal lens can focus light to a single point, but in reality the focusing is limited by diffraction. In this case, the diffraction limit can be linked to the Heisenberg uncertainty principle (see text).

A key case in point is the Veselago–Pendry "perfect lens" [Ves68, Pen00] (see p. 447) that is, in principle, capable of producing ideal (non-distorted) images.[32] This is possible because the evanescent waves with large wavenumbers k_x, k_y in the image plane xy, or equivalently with large components of momentum $p_x = \hbar k_x$, $p_y = \hbar k_y$ resolve [Pen01] the apparent contradiction

[32] The perfect lensing effect has been challenged by many researchers (N. Garcia & M. Nieto-Vesperinas [GNV02, NVG03], J.M. Williams [Wil01], A.L. Pokrovsky & A.L. Efros [PE03], P.M. Valanju *et al.* [VWV02, VWV03]) but for the most part has survived the challenge (see J.B. Pendry & D.R. Smith [PS04], J.R. Minkel [Min03]). Part of the difficulties and the controversy arise because the problem with the "perfect lens" parameters ($\epsilon = -1$, $\mu = -1$ for a slab) is ill-posed, and the analysis depends on regularization and on the way of passing to the small-loss and (in some cases) low-frequency limits.

[Wil01] between the diffraction limit and the uncertainty principle. Indeed, the dispersion relation for waves in free space (air) is

$$k_x^2 + k_y^2 + k_z^2 \;=\; \left(\frac{\omega}{c}\right)^2$$

In the evanescent field, k_x and k_y can be arbitrarily large, with the corresponding imaginary value of k_z and negative k_z^2. The uncertainty in the xy-components of the photon momentum is therefore infinite, and there is no uncertainty in the position in the ideal case.

The remainder of this section is devoted to a less exotic way of beating the diffraction limit: strong plasmon amplification of the field in SNOM (Scanning Near-Field Optical Microscopy). SNOM is a very significant enhancement of more traditional Scanning Probe Microscopy (SPM).

The first type of SPM, the Scanning Tunneling Microscope (STM), was developed by Gerd Binnig and Heinrich Rohrer at the IBM Zurich Research Laboratory in the early 1980's [BRGW82] (see also [BR99]). For this work, Binnig and Rohrer were awarded the 1986 Nobel Prize in Physics.[33] The main part of the STM is a sharp metallic tip in close proximity (~ 10 Å or less) to the surface of the sample; the tip is moved by a piezoelectric device. A small voltage, from millivolts to a few volts, is applied between the tip and the surface, and the system measures the quantum tunneling current (from pico- to nano-Amperes) that results. Since the probability of tunneling depends exponentially on the gap, the device is extremely sensitive. Binnig and Rohrer were able to map the surface with atomic resolution. STMs normally operate in a constant current mode while the tip is scanning the surface. The constant tunneling current is maintained by adjusting the elevation of the tip, which immediately identifies the topography of the surface.

The second type of Scanning Probe Microscopy is Atomic Force Microscopy (AFM). Instead of the tunneling current, AFM measures the interaction force between the tip and the surface (short-range repulsion or van der Waals attraction), which provides information about the surface structure and topography.

To achieve atomic-scale resolution in all types of SPM, the position of the tip has to be controlled with extremely high precision and the tip has to be very sharp, up to just one atom at its very apex. Modern SPM technology satisfies both requirements.

While the level of resolution in atomic force and tunneling microscopes is amazing, these devices are blind – they can only "feel" but not see the surface. Vision – a tremendous enhancement of the scanning probe technology – is acquired in Scanning Near-Field Optical Microscopy.

Two main approaches currently exist in SNOM. In the first one, light illuminates the sample after passing through a small (subwavelength) pinhole;

[33] Ernst Ruska received his share of that prize "for his fundamental work in electron optics, and for the design of the first electron microscope".

the size of the hole determines the level of resolution. The idea dates back to E.H. Synge's papers in 1928 and 1932 [Syn28, Syn32]. In modern realization, the "pinhole" is actually a metal-coated fiber (Fig. 7.34 and caption to it).

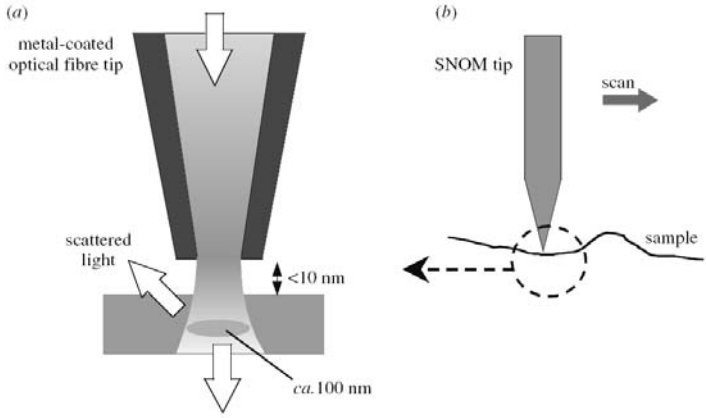


Fig. 7.34. A schematic of aperture-SNOM. An optical-fiber tip is scanned across a sample surface to form an image. The tip is coated with metal everywhere except for a narrow aperture at the apex.(Reprinted by permission from D. Richards [Ric03] ©2003 The Royal Society of London.)

An interesting timeline for the development of this aperture-*limited* type of SNOM is posted on the website of Nanonics Imaging Ltd.:[34]

> **1928/1932**
> E.H. Synge proposes the idea of using a small aperture to image a surface with subwavelength resolution using optical light. For the small opening, he suggests using either a pinhole in a metal plate or a quartz cone that is coated with a metal except for at the tip. He discusses his theories with A. Einstein, who helps him develop his ideas. ...
> **1956**
> J.A. O'Keefe, a mathematician, proposes the concept of Near-Field Microscopy without knowing about Synge's earlier papers. However, he recognizes the practical difficulties of near-field microscopy and writes the following about his proposal: "The realization of this proposal is rather remote, because of the difficulty providing for relative motion between the pinhole and the object, when the object must be

[34] http://www.nanonics.co.il/main/twolevels_item1.php?ln=en&item_id=34-&main_id=14 Nanonics Imaging Ltd. specializes in near-field optical microscopes combined with atomic force microscopes.

brought so close to the pinhole." [J.A. O'Keefe, "Resolving power of visible light," *J. of the Opt. Soc. of America*, 46, 359 (1956)].
In the same year, Baez performs an experiment that acoustically demonstrates the principle of near-field imaging. At a frequency of 2.4 kHz ($\lambda = 14$ cm), he shows that an object (his finger) smaller than the wavelength of the sound can be resolved.
1972
E.A. Ash and G. Nichols demonstrate $\lambda/60$ resolution in a scanning near-field microwave microscope using 3 cm radiation. [E.A. Ash and G. Nichols, "Super-resolution aperture scanning microscope," *Nature* 237, 510 (1972).]
1984
The first papers on the application of NSOM/SNOM appear. These papers ... show that NSOM/SNOM is a practical possibility, spurring the growth of this new scientific field. [A. Lewis, M. Isaacson, A. Harootunian and A. Murray, *Ultramicroscopy* 13, 227 (1984); D.W. Pohl, W. Denk and M. Lanz [PDL84]].

[End of quote from the Nanonics website.]

In aperture-limited SNOM, high resolution, unfortunately, comes at the expense of significant attenuation of the useful optical signal: the transmission coefficient through the narrow fiber is usually in the range of $\sim 10^{-3}$–10^{-5}, which limits the applications of this type of SPM only to samples with very strong optical response.

A very promising alternative is apertue*less* SNOM that takes advantage of local amplification of the field by plasmonic particles. This idea was put forward by J. Wessel in [Wes85]; his design is shown in Fig. 7.35 and is summarized in the caption to this figure.

A remarkably high optical resolution of ∼15–30 nm has already been demonstrated by several research groups (T. Ichimura *et al.* [IHH$^+$04], N. Anderson *et al.* [AHCN05]), albeit with rather weak useful optical signals. To realize the full potential of apertureless SNOM, the local field amplification by plasmonic particles needs to be maximized. However, this amplification is quite sensitive to the geometric and physical design of plasmon-enhanced tips. For a radical improvement in the strength of the useful optical signal, one needs to unify accurate simulation with effective measurements of the efficiency of the tips and with fabrication.

As an illustration, in A.P. Sokolov's laboratory at the University of Akron[35] a stable and reproducible enhancement of for the Raman signal on the order of $\sim 10^3$–10^4 was achieved for gold- and silver-coated Si_3N_4- and Si-tips in 2005–2006. As noted by Sokolov, this enhancement may be sufficient for the analysis of thin (a few nanometer) films. However, for thicker samples,

[35] A brief description of their experimental setup for Raman spectroscopy is given below.

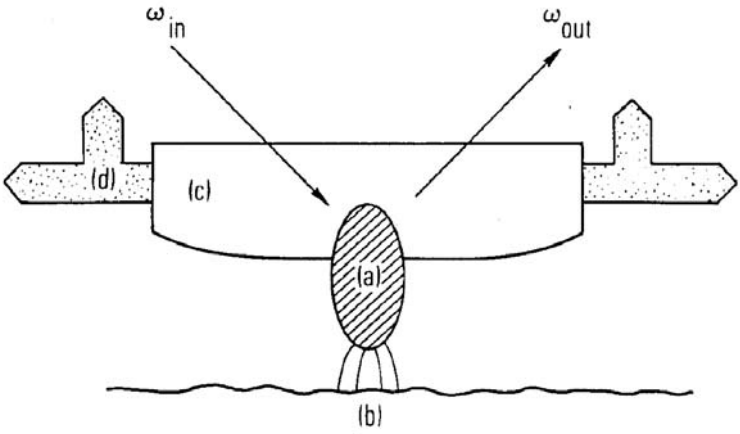


Fig. 7.35. The optical probe particle (a) intercepts an incident laser beam, of frequency ω_{in}, and concentrates the field in a region adjacent to the sample surface (b). The Raman signal from the sample surface is reradiated into the scattered field at frequency ω_{out}. The surface is scanned by moving the optically transparent probe-tip holder (c) by piezoelectric translators (d). (Reprinted by permission from J. Wessel [Wes85] ©1985 The Optical Society of America.)

due to the large volume contributing to the far-field signal relative to the volume contributing to the near-field signal, the Raman enhancement of $\sim 10^4$ does not produce a high enough ratio between near-field and far-field signals.

At the same time, a dramatically higher Raman enhancement, by a factor of $\sim 10^6$ or more, appears to be within practical reach if tip design is optimized. This would constitute an enormous qualitative improvement over the existing technology, as the useful Raman signal would exceed the background field. Since plasmon enhancement is a subtle and sensitive physical effect, and since human intuition with regard to its optimization is quite limited, computer simulation – the main subject of this book – becomes crucial.

The computational methods and simulation examples for plasmon-enhanced SNOM are described in Section 7.12.3, after an illustration of the experimental setup in Section 7.12.2. For general information on SNOM, the interested reader is referred to the books by M.A. Paesler & P. J. Moyer [PM96] and by P.N. Prasad [Pra03, Pra04].

7.12.2 Apertureless and Dark-Field Microscopy

This section briefly describes the experimental setup in A.P. Sokolov's laboratory at the University of Akron. The figures in this section are courtesy A.P. Sokolov. For further details, see D. Mehtani *et al.* [MLH+05, MLH+06].

A distinguishing feature of the setup is side-collecting optics (Fig. 7.36, top) that does not suffer from the shadowing effect of more common

illumination/collection optics above the tip. Another competing design, with illumination from below, works only for optically transparent substrates, whereas side illumination can be used for any substrates and samples. Finally, the polarization of the wave coming from the side can be favorable for plasmon enhancement. Indeed, it is easy to see that the electric field, being perpendicular to the direction of propagation of the incident wave, can have a large vertical component that will induce a plasmon-resonant field just below the apex of the tip, as desired. In contrast, for top or bottom illumination the direction of wave propagation is vertical, and hence the electric field has to be horizontal, which is not conducive to plasmon enhancement underneath the tip.

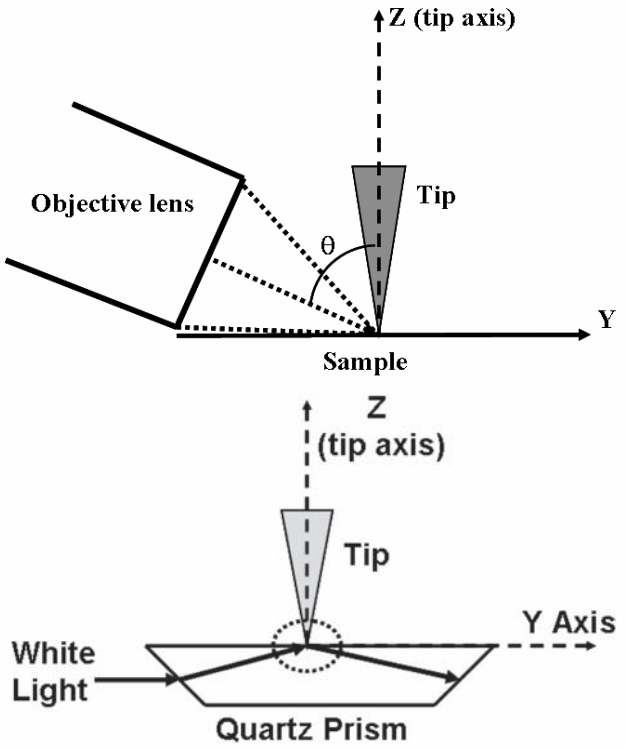


Fig. 7.36. Experimental setup. Top: schematics of side-illumination/collection optics. Bottom: dark-field microscopy for measuring plasmon field enhancement at the apex of the tip. (Figure courtesy A.P. Sokolov. Bottom part reprinted by permission from D. Mehtani *et al.* [MLH+06] ©2006 IOP Publishing Ltd.)

Before a plasmon-enhanced tip can be used, it is important to evaluate the level of field amplification at the apex. Direct measurements of the optical

response of the tip are not effective because the measured spectrum of the tip as a whole may differ significantly from the spectrum of the plasmon area at the apex.

An elegant solution is dark-field microscopy (C.C. Neacsu *et al.* [NSR04], D. Mehtani *et al.* [MLH+06]). The apex of the tip is placed in the evanescent field that exists above the surface of a glass prism due to total internal reflection (Fig. 7.36, bottom). Away from the glass surface, the evanescent field falls off exponentially and therefore is not seen by the collecting system. At the same time, the evanescent field does induce a plasmon resonance. Indeed, such resonance is, to a good degree of approximation, a quasi-static effect that will manifest itself once an external electric field is present and once the effective dielectric constant of the plasmonic structure is close to its resonance value. The exponential decay of the field matters only insofar as it can induce higher-order plasmon modes; this happens if the particle size is large enough for the variation of the field over the particle to be appreciable. The frequency of light affects the result indirectly, via frequency dependence of the dielectric permittivities.

The side-collecting optics is critical for dark-field measurements, as it allows virtually unobstructed collection of optical signals from the apex of the tip.

7.12.3 Simulation Examples for Apertureless SNOM

The dependence of plasmon-amplified fields on geometric and physical parameters, as well as the dependence of these parameters (dielectric permittivities) on frequency, is so complex that computer modeling is indispensable in tip design and optimization. A natural simulation protocol consists of two parts: electrostatics and wave analysis. Electrostatic simulations may give qualitative predictions and allow one to optimize the design but at the same time have substantial limitations, as discussed below.

Electrostatic Approximation in SNOM

The electrostatic approximation is useful because the dimensions of the apex of the tip, with its plasmon decoration, are typically much smaller than the wavelength of incident light. In addition, for axially symmetric designs, the electrostatic problem becomes effectively two-dimensional and hence is much faster to solve. One needs to be aware, however, of a major limitation of the electrostatic model: it cannot adequately represent dephasing, retardation, and antenna-like resonances along the length of the tip. Hence full electromagnetic wave analysis is in many instances indispensable and will be considered later in this section.

For the electrostatic simulations, the FEMLAB™ (COMSOL Multiphysics™) package was used. F. Čajko incorporated FEMLAB commands into Matlab scripts for postprocessing and multiparametric optimization

(D. Mehtani *et al.* [MLH+06]). In all simulations described below, the amplitude of the incident field is normalized to unity, so that the values in the plots represent the amplification of the electric field. To get a more realistic picture, it makes sense to deal with the *mean* value of the field rather than just the point-wise value at the very apex. To this end, the field is computed 1 nm below the tip (which represents a practically useful gap between the apex and the sample) and, since the resolution of the tip-enhanced spectroscopy is expected to approach ∼10–15 nm, the field is averaged over a horizontal disk with radius 10 nm located 1 nm below the apex.

In the simulations, the P.B. Johnson & R.W. Christy data [JC72] for the dielectric properties of silver and gold are used, and the M.A. Ordal *et al.* [OLB+83] and J.H. Weaver *et al.* [WOL75] data are adopted for tungsten tips.

One sample setup, due to Y.C. Martin *et al.* [MHW01], is useful for testing and verification and involves a semispherical gold or silver particle at the apex of the tungsten or silicon tip (Fig. 7.37, top). With the optimal dimensions of the particle, the field of coated Si tip is amplified by a factor of ∼47 for gold and ∼132 for silver.

F. Čajko's simulations have shown that the level of plasmon enhancement depends strongly not only on the dimensions and material of the plasmonic particle but also on other geometric parameters and on the material of the tip. For different materials (Au and Ag) the resonance wavelength is different as shown in Fig. 7.37, and the optimal aspect ratio of the semispheroid changes as well. For a slightly different design with a conical tip, Fig. 7.38 illustrates the effects of the varying permittivity of the tip and the angle of the cone. These two parameters have a lesser impact on the field enhancement than the aspect ratio of the particle.

Wave Simulations of Optical Tips

Full-wave simulations are performed using HFSS™ – the Finite-Element software from the Ansoft Corporation. Under the electrostatic approximation, the problem is axisymmetric; in wave analysis the distinctive direction of wave propagation breaks the axial symmetry.

The Martin *et al.* tip with a semispheroidal particle [MHW01] is again used as a test case. To limit the size of the computational domain, for simplicity of this model example the tip is truncated to a length of 100 nm. However, as discussed later in this section, one should be aware that such truncation may have undesirable side effects.

The simulation domain is cylindrical, with radius 800 nm and height 340 nm. Due to computational constraints, the radial distance between the scatterer and the domain boundary is about one wavelength, and second-order radiation boundary conditions are applied to reduce the error due to this finite domain size. Incident plane waves travel from the left and are polarized in the vertical direction (Fig. 7.37, top).

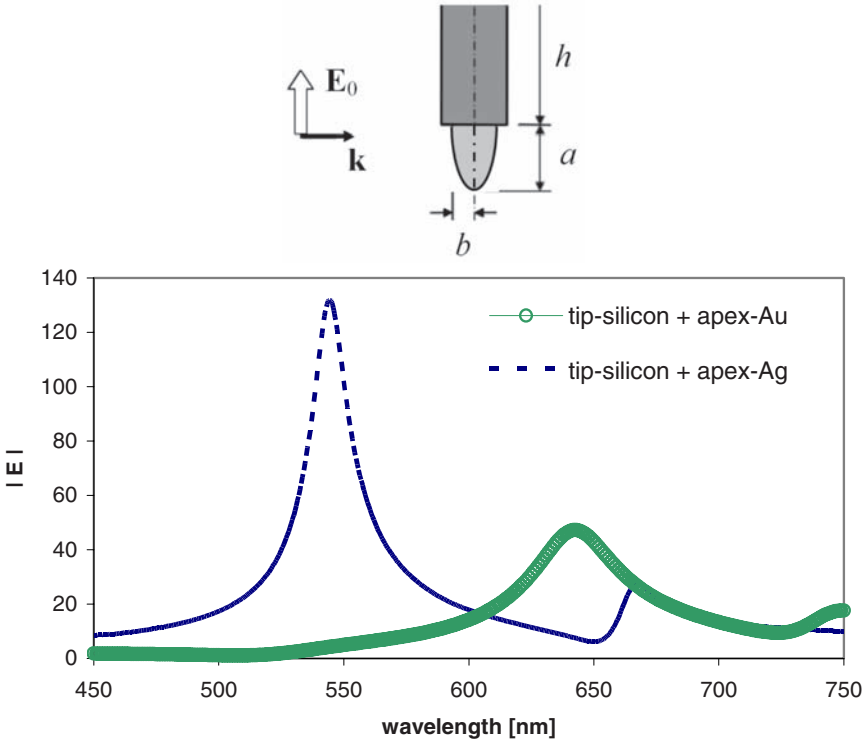


Fig. 7.37. (Credit: F. Čajko.) Electrostatic simulation. Top: Geometric setup [MHW01] – a semispherical plasmonic particle (with dimensions a, b as shown) attached to a tip with height h and radius r. Incident plane wave traveling in the $+x$-direction is linearly polarized in the vertical direction. Bottom: FEMLAB^{TM} simulation for Si tips with attached Au and Ag semispheroids. The mean total electric field 1 nm below the tip is shown as a function of wavelength in air for the geometric parameters $h = 100$ nm, $r = 17$ nm, $a = 40$ nm and $b = 8$ nm.

Wave analysis is used to relate local field amplification below the tip to the spatial distribution of the radiated fields. If a strong correlation between these fields exists, far fields measured by dark-field microscopy will indeed be a good indicator of the near-field enhancement at the tip. The simulations do confirm this correlation (D. Mehtani *et al.* [MLH$^+$06]), in agreement with experimental results published by other groups (C.C. Neacsu *et al.* [NSR04]). In Fig. 7.39, a measure of the far field is plotted against the magnitude of the near-field. Different curves in the figure correspond to different tip designs. Points on the curves correspond to different wavelengths. As the wavelength increases, the near- and far-fields increase simultaneously, and roughly in proportion to one another, until they reach their maximum values; then both fields simultaneously decrease as the wavelength keeps increasing.

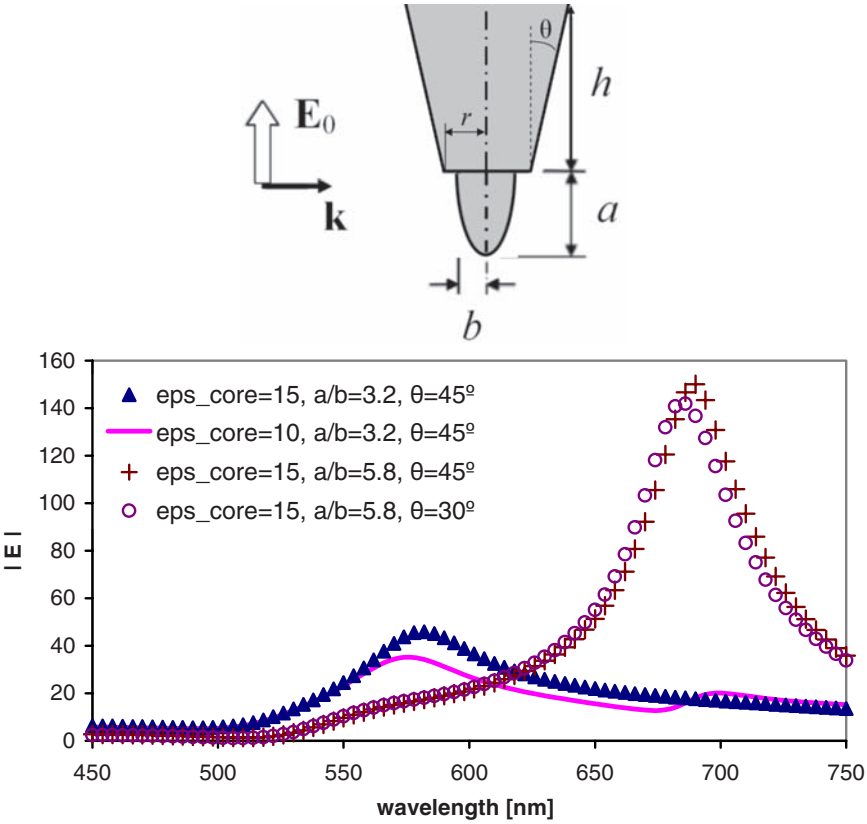


Fig. 7.38. (Credit: F. Čajko.) The electric field enhancement for a conical tip with a semispheroidal gold particle. Top: the geometric setup. The wave impinging from the left is linearly polarized in the vertical direction. Bottom: electrostatic simulations of the average electric field for several sets of parameters. The aspect ratio, the permittivity and the angle vary.

The main challenge in the full electrodynamic simulation of optical tips is the multiscale nature of the problem. The apex of the tip, with dimensions well below the wavelength (radius of ∼20–30 nm), has to be represented very accurately in the model, as it is the heart of the optical device. At the same time, the tip could be several wavelengths long, so that the difference between the height of the tip and the size of its apex could reach about three orders of magnitude. Although truncation of the tip for computational purposes may at first glance look like a reasonable idea, this truncation distorts the antenna-type resonances that can be induced along the length of the tip.

The problem can thus be viewed as multiscale not only in terms of its geometry but in terms of physics as well: antenna resonances along the length of the tip are coupled with plasmon resonances and scattering effects in the small area around the apex. This has been pointed out in the literature,

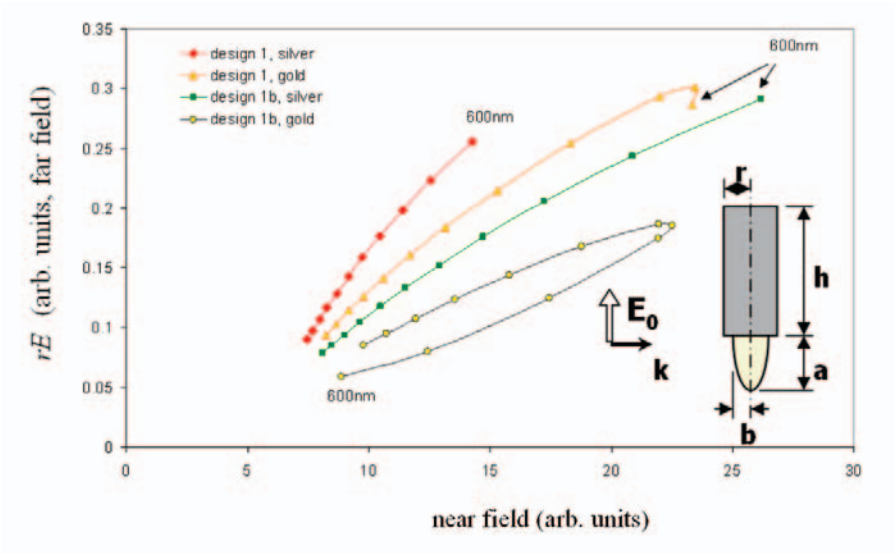


Fig. 7.39. (Credit: F. Čajko.) Correlation between the far field and the near-field for two tip designs and different wavelengths. $h = 100$ nm, $a = 40$ nm. Design 1: $r = 20$ nm, $b = 12$ nm; design 1b: $r = 17$ nm, $b = 8.6$ nm.

particularly by F. Keilmann's experimental group [KH04] and in the paper on tip simulation by R. Esteban *et al.* [EVK06].

The multiscale character of the modeling is a hurdle for any numerical method. Esteban *et al.* use the Multiple Multipole Method (see Section 7.11.4 on p. 425) and report a variety of interesting results, including the dependence of near-fields around the apex on the height of the tip, i.e. on its antenna-like behavior.

In FEM, there are several ways of dealing with multiscale challenges. One is adaptive mesh refinement described in Section 3.13 on p. 148. Another possibility – after solving the global problem on a relatively coarse mesh – is to "zoom in" on the apex area and solve a local problem there with high resolution. The boundary conditions for the local problem come from the global solution. This approach is not completely satisfactory, though, as the accuracy of the local boundary conditions is limited by the global mesh. A related systematic and rigorous procedure is known as *domain decomposition* and has been very extensively studied (A. Toselli & O. Widlund [TW05]).[36]

An example of adaptive mesh refinement and a distribution of the scattered field is given in Fig. 7.40. The simulation was performed by F. Čajko with the HFSS package, and the physical and experimental setup is due to F. Keilmann, R. Hillenbrand [HTK02, KH04] and others at the Max-Planck-Institut für

[36] See also http://www.ddm.org

Biochemie in Martinsried, Germany.[37] The useful signal is due to scattering from a sharp platinum tip; its apex is *not* plasmon-enhanced. This technique is known as *scattering-type* Scanning Near-field Optical Microscopy (s-SNOM), in contrast with plasmon-enhanced SNOM. The antenna-like behavior of the tip and its coupling with near-field at the apex are indeed very important in this setup. The near-field is strongly enhanced when a *polaritonic* sample (such as silicon carbide) with negative dielectric permittivity is probed.

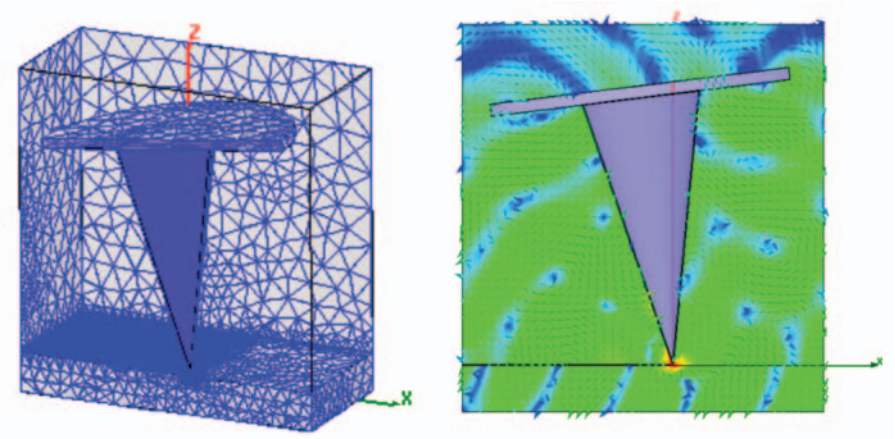


Fig. 7.40. (Credit: F. Čajko.) An example of a finite element mesh and the scattered field near the infrared tip. Experimental setup due to F. Keilmann's group.

F. Keilmann's device operates in the mid-infrared; the simulation example is for the wavelength $\lambda = 10.5\ \mu m$ in free space. The radius of the apex of the tip in the simulation is about 600 times smaller than the domain size, so that truly disparate scales are involved. Moreover, a thick SiC substrate with a thin (10–30 nm) gold layer is also included in the model. Very high mesh density in the gold layer is clearly visible in Fig. 7.40. Details of these simulations are left for more specialized publications and will not be described here.

7.13 Backward Waves, Negative Refraction and Superlensing

7.13.1 Introduction and Historical Notes

Since the beginning of the 21st century, negative refraction has become one of the most intriguing areas of research in nano-photonics, with quite a few books

[37] I am grateful to Fritz Keilmann for giving us an opportunity to work on this problem.

and review papers already written: P.W. Milonni [Mil04], G.V. Eleftheriades & K.G. Balmain (eds.) [EB05], S.A. Ramakrishna [Ram05]. Development of optical materials with negative refraction is examined by V.M. Shalaev in [Sha06].

In his 1967 paper [Ves68],[38] V.G. Veselago showed that materials with simultaneously negative dielectric permittivity ϵ and magnetic permeability μ would exhibit quite unusual behavior of wave propagation and refraction. More specifically:

- Vectors $\mathbf{E}$, $\mathbf{H}$ and $\mathbf{k}$, in that order, form a left-handed system.
- Consequently, the Poynting vector $\mathbf{E} \times \mathbf{H}$ and the wave vector $\mathbf{k}$ have opposite directions.
- The Doppler and Vavilov–Cerenkov effects are "reversed". The sign of the Doppler shift in frequency is opposite to what it would be in a regular material. The Poynting vector of the Cerenkov radiation forms an obtuse angle with the direction of motion of a superluminal particle in a medium, while the wave vector of the radiation is directed *toward* the trajectory of the particle.
- Light propagating from a regular medium into a double-negative material bends "the wrong way" (Fig. 7.41). In Snell's law, this corresponds to a negative index of refraction. A slab with $\epsilon = -1$, $\mu = -1$ in air acts as an unusual lens (Fig. 7.42).

Subjects closely related to Veselago's work had been in fact discussed in the literature well before his seminal publication – as early as in 1904. S.A. Tretyakov [Tre05], C.L. Holloway *et al.* [HKBJK03] and A. Moroz[39] provide the following references:

- A 1904 paper by H. Lamb[40] on waves in mechanical (rather than electromagnetic) systems.
- A. Schuster's monograph [Sch04], pp. 313–318; a 1905 paper by H.C. Pocklington[41].
- Negative refraction of electromagnetic waves was in fact considered by L.I. Mandelshtam more than two decades prior to Veselago's paper.[42] Man-

[38] Published in 1967 in Russian. In the English translation that appeared in 1968, the original Russian paper is mistakenly dated as 1964.

[39] http://www.wave-scattering.com/negative.html

[40] "On group-velocity," *Proc. London Math. Soc.* 1, pp. 473–479, 1904.

[41] H.C. Pocklington, Growth of a wave-group when the group velocity is negative, *Nature* 71, pp. 607–608, 1905.

[42] Leonid Isaakovich Mandelshtam (Mandelstam), 1879–1944, an outstanding Russian physicist. Studied at the University of Novorossiysk in Odessa and the University of Strasbourg, Germany. Together with G.S. Landsberg (1890–1957), observed Raman (in Russian – "combinatorial") scattering simultaneously or even before Raman did but published the discovery a little later than Raman. The 1930 Nobel Prize in physics went to Raman alone; for an account of these events, see I.L. Fabelinskii [Fab98], R. Singh & F. Riess [SR01] and E.L. Feinberg [Fei02].

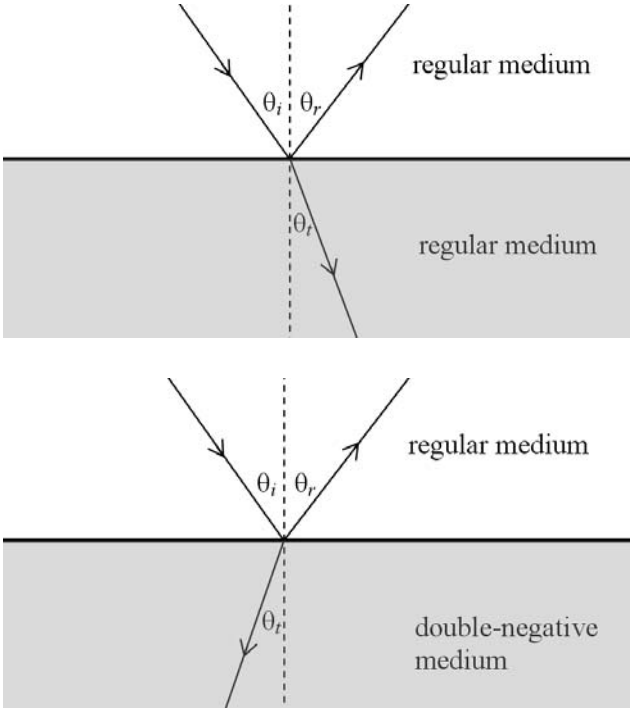


Fig. 7.41. At the interface between a regular medium and a double-negative medium light bends "the wrong way"; in Snell's law, this implies a negative index of refraction. Arrows indicate the direction of the Poynting vector that in the double-negative medium is opposite to the wave vector.

delshtam's short paper [Man45] and, even more importantly, his lecture notes [Man47, Man50] already described the most essential features of negative refraction. The 1945 paper, but not the lecture notes, is cited by Veselago.

- A number of papers on the subject appeared in Russian technical journals from the 1940s to the 1970s: by D.V. Sivukhin (1957) [Siv57], V.E. Pafomov (1959) [Paf59] and R.A. Silin (1959, 1978) [Sil59, Sil72].
- Silin's earlier review paper (1972) [Sil72], where he focuses on wave propagation in artificial periodic structures.

In one of his lectures cited above, Mandelshtam writes, in reference to a figure similar to Fig. 7.41 ([Man50], pp. 464–465):[43]

> "... at the interface boundary the tangential components of the fields ... must be continuous. It is easy to show that these conditions cannot

[43] My translation from the Russian. A similar quote is given by S.A. Tretyakov in [Tre05].

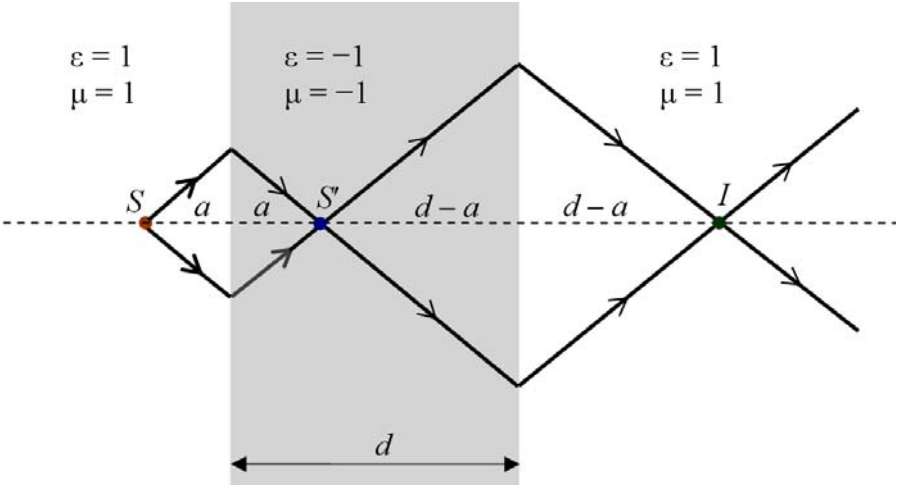


Fig. 7.42. The Veselago slab of a double-negative material acts as an unusual lens. Due to the negative refraction at both surfaces of the slab, a point source S located at a distance $a < d$ has a virtual image S' inside the slab and a real image I outside. The arrows indicate the direction of the Poynting vector, not the wave vector.

be satisfied with a reflected wave (or a refracted wave) alone. But with both waves present, the conditions can always be satisfied. From that, by the way, it does not at all follow that there must only be *three* waves and not more: the boundary conditions do allow one more wave, the fourth one, traveling at the angle $\pi - \phi_1$ in the second medium. Usually it is tacitly assumed that this fourth wave does not exist, i.e. it is postulated that only one wave propagates in the second medium. ... [the law of refraction] is satisfied at angle ϕ_1 as well as at $\pi - \phi_1$. The wave ... corresponding to ϕ_1 moves away from the interface boundary. ... The wave corresponding to $\pi - \phi_1$ moves toward the interface boundary. It is considered self-evident that the second wave cannot exist, as light impinges from the first medium onto the second one, and hence in the second medium *energy* must flow away from the interface boundary. But what does energy have to do with this? The direction of wave propagation is in fact determined by its *phase* velocity, whereas energy moves with *group* velocity. Here therefore there is a logical leap that remains unnoticed only because we are accustomed to the coinciding directions of propagation of energy and phase. If these directions do coincide, i.e. if group velocity is positive, then everything comes out correctly. If, however, we are dealing with the case of negative group velocity – quite a realistic case, as I already said, – then everything changes. Requiring as before that energy in the second medium flow *away* from the interface boundary, we arrive

at the conclusion that phase must run toward this boundary and, therefore, the direction of propagation of the refracted wave will be at the $\pi - \phi_1$ angle to the normal. However unusual this setup may be, there is, of course, nothing surprising about it, for phase velocity does not tell us anything about the direction of energy flow."

A quote from Silin's 1972 paper:

"Let a wave be incident from free space onto the dielectric. In principle one may construct two wave vectors β_2 and β_3 of the refracted wave ... Both vectors have the same projection onto the boundary of the dielectric and correspond to the same frequency. One of them is directed away from the interface, while the other is directed toward it. The waves corresponding to the vectors β_2 and β_3 are excited in media with positive and negative dispersion, respectively. In conventional dielectrics the dispersion is always positive, and a wave is excited that travels away from the interface. ...
The direction of the vector β_3 toward the interface in the medium with negative dispersion coincides with the direction of the phase velocity ... and is opposite to the group velocity $\mathbf{v}_{\mathrm{gr}}$. The velocity $\mathbf{v}_{\mathrm{gr}}$ is always directed away from the interfaces, so that the energy of the refracted wave always flows in the same direction as the energy of the incident wave."

Of the earlier contributions to the subject, a notable one was made by R. Zengerle in his PhD thesis on singly and doubly periodic waveguides in the late 1970s. His journal publication of 1987 [Zen87] contains, among other things, a subsection entitled "Simultaneous positive and negative ray refraction". Quote:

"Figure 10 shows refraction phenomena in a periodic waveguide whose effective index ... in the modulated region is ... higher than ... in the unmodulated region. The grating lines, however, are not normal to the boundaries. As a consequence of the boundary conditions, two Floquet-Bloch waves corresponding to the upper and lower branches of the dispersion contour ... are excited simultaneously ... resulting generally in two rays propagating in different directions. This ray refraction can be described by two effective ray indices: one for ordinary refraction ... and the other ... with a negative refraction angle ... "

The first publication on what today would be called a (quasi-)perfect cylindrical lens was a 1994 paper by N.A. Nicorovici *et al.* [NMM94] (now there are also more detailed follow-up papers by G.W. Milton *et al.* [MNMP05, MN06]).[44] These authors considered a coated dielectric cylinder, with the core of radius r_{core} and permittivity ϵ_{core}, the shell (coating) with the outer radius

[44] I am grateful to N.-A. Nicorovici for pointing these contributions out to me.

r_{shell} and permittivity ϵ_{shell}, embedded in a background medium with permittivity ϵ_{bg}. It turns out, first, that such a coated cylinder is completely transparent to the outside H-mode field (the H-field along the axis of the cylinder) under the quasistatic approximation if $\epsilon_{\text{core}} = \epsilon_{\text{bg}} = 1$, $\epsilon_{\text{shell}} \to -1$. (The limiting case $\epsilon_{\text{shell}} \to -1$ should be interpreted as the imaginary part of the permittivity tending to zero, while the real part is fixed at -1: $\epsilon_{\text{shell}} = -1 - \mathrm{i}\epsilon''_{\text{shell}}$, $\epsilon''_{\text{shell}} \to 0$.)[45] Second, under these conditions for the dielectric constants, many unusual imaging properties of coated cylinders are observed. For example, a line source placed outside the coated cylinder at a radius $r_{\text{src}} < r^3_{\text{shell}}/r^2_{\text{core}}$ would have an image *outside* the cylinder, at $r_{\text{image}} = r^4_{\text{shell}}/(r^2_{\text{core}} r_{\text{src}})$.

A turning point in the research on double-negative materials came in 1999–2000, when J.B. Pendry *et al.* [PHRS99] showed theoretically, and D.R. Smith *et al.* [SPV$^+$00] confirmed experimentally, negative refraction in an artificial material with split-ring resonators [SPV$^+$00]. A further breakthrough was Pendry's "perfect lens" paper in 2000 [Pen00]. It was known from Veselago's publications that a slab of negative index material could work as a lens focusing light from a point-like source on one side to a point on the other side.[46] Veselago's argument was based purely on geometric optics, however. Pendry's electromagnetic analysis showed, for the first time, that the evanescent part of light emitted by the source will be *amplified* by the slab, with the ultimate result of perfect transmission and focusing of both propagating and evanescent components of the wave.

The research field of negative refraction and superlensing has now become so vast that a more detailed review would be well beyond the scope of this book. Further reading may include J.B. Pendry & S.A. Ramakrishna [PR03], J.B. Pendry & D.R. Smith [PS04], S.A. Ramakrishna [Ram05], A.L. Pokrovsky & A.L. Efros [PE02, PE03], and references therein. Selected topics, however, will be examined in the remainder of this chapter.

7.13.2 Negative Permittivity and the "Perfect Lens" Problem

This section gives a numerical illustration of Pendry's "perfect lens" in the limiting case of a thin slab. If the thickness of the slab is much smaller than the wavelength, the problem becomes quasi-static and the electric and magnetic fields decouple. Analysis of the (decoupled) electric field brings us back from a brief overview of negative index materials to media with a negative real part of the dielectric permittivity. Rather than repeating J.B. Pendry's analytical calculation for a thin metal slab, let us, in the general spirit of this book, consider a numerical example illustrating the analytical result.

The problem, in the electrostatic limit, can be easily solved by Finite Element analysis. The geometric and physical setup is, for the sake of comparison,

[45] As a reminder, the $\exp(+\mathrm{i}\omega t)$ convention is used for complex phasors. See p. 352.

[46] V.G. Veselago remarks that this is not a lens "in the usual sense of the word" because it does not focus a parallel beam to a point.

chosen to be the same as in Pendry's paper [Pen00]. A FEMLAB^{TM} (Comsol Multiphysics^{TM}) mesh for 2D simulation is shown in Fig. 7.43. A metal slab of thickness 40 nm acts, under special conditions, as a lens. To demonstrate the lensing effect, two line charges (represented in the simulation by circles of 5 nm radius, not drawn exactly up to scale in the figure) are placed 20 nm above the surface of the slab, at points $(x, y) = (\pm 40, 40)$ nm. (The y axis is normal to the slab.) In the simulations reported below, the FE mesh has 30,217 nodes and 60,192 second-order triangular elements, with 120,625 degrees of freedom. Naturally, for the FE analysis the domain and the (theoretically infinite) slab had to be truncated sufficiently far away from the source charges.

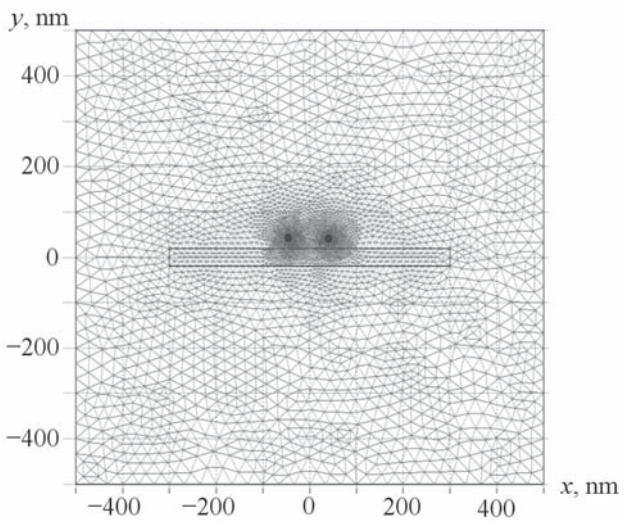


Fig. 7.43. A finite element mesh for Pendry's lens example with two line sources.

In Pendry's example ([Pen00], p. 3969), the relative permittivity of the slab is $\epsilon_{\text{slab}} \approx -0.98847 - 0.4\text{i}$,[47] which corresponds to silver at ~ 356 nm. The magnitude of the electric field in the source plane $y = 40$ nm is shown, as a function of x, in Fig. 7.44 and, as expected, exhibits two sharp peaks corresponding to the line sources.

The lensing effect of the slab is manifest in Fig. 7.45, where the field distributions with and without the slab are compared in the "image" plane ($y = -40$ nm).[48] Perfect lensing is a very subtle phenomenon and is extremely

[47] With the $\exp(+\text{i}\omega t)$ convention for phasors.

[48] A similar distribution of the electrostatic *potential* in the image plane has a flat maximum at $x = 0$ rather than two peaks. Note also that the maximum value theorem for the Laplace equation prohibits the potential from having a local

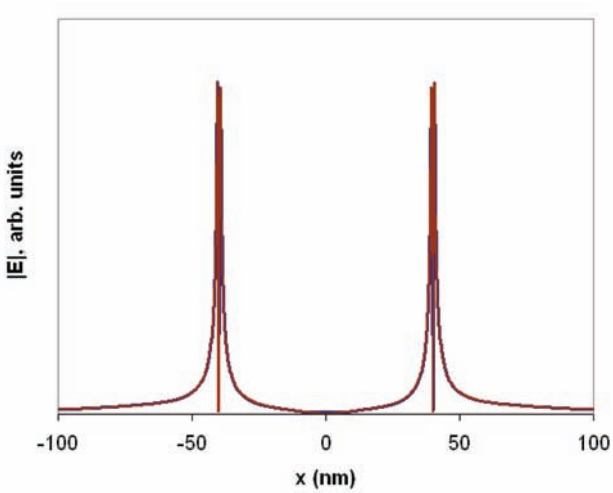


Fig. 7.44. The magnitude of the electric field in the source plane ($y = 40$ nm) as a function of x. The two line sources are manifest. (The field abruptly goes to zero at the very center of each cylindrical line of charge.)

sensitive to all physical and geometric parameters of the model. Ideally, the distance between the source and the surface of the slab has to be equal to half of the thickness of the slab; the relative permittivity has to be -1. In addition, if the thickness of the slab is not negligible relative to the wavelength, the permeability also has to be equal -1. R. Merlin [Mer04] (see also D.R. Smith *et al.* [SSR+03]) derived an analytical formula for the spatial resolution Δ of a slightly imperfect lens of thickness d and the refractive index $n = -(1-\delta)^{1/2}$, with δ small:

$$\Delta = \frac{2\pi d}{\left|\log \frac{\delta}{2}\right|} \tag{7.231}$$

According to this result, for a modest resolution Δ equal to the thickness of the slab, the deviation δ must not exceed ~ 0.0037. For $\Delta/d = 0.25$, δ must be on the order of 10^{-11}, i.e. the index of refraction must be almost perfectly equal to -1. This obviously imposes serious practical constraints on the design of the lens.

For a qualitative illustration of this sensitivity to parameters, let us turn to the electrostatic limit again and visualize how a slight variation of the numbers affects the potential distribution. In Figs. 7.46–7.48 the dielectric constant is purely real and takes on the values -0.9, -1, and -1.02; although these values are close, the results corresponding to them are completely different.

---

maximum (or minimum) strictly inside the domain with respect to all coordinates. Viewed as a function of *one* coordinate, with the other ones fixed, the potential can have a local maximum.

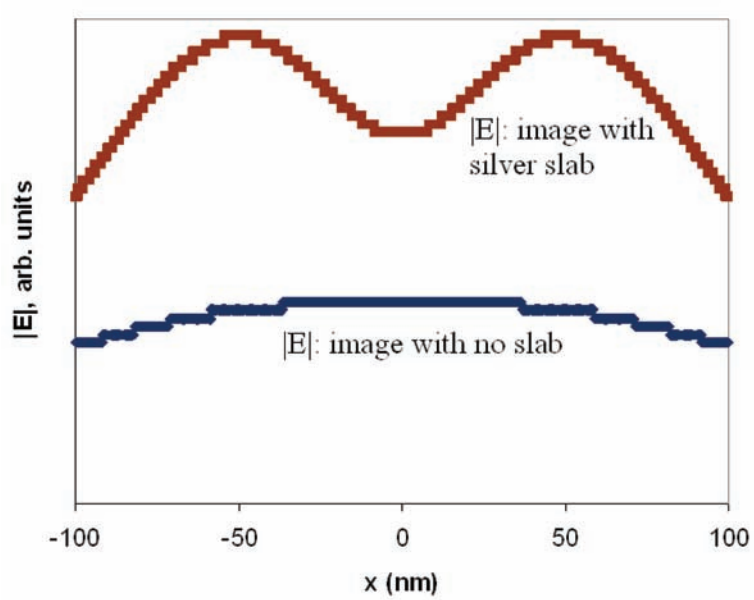


Fig. 7.45. The magnitude of the electric field in the image plane ($y = -40$ nm) as a function of x, with and without the silver slab. The lensing effect of the slab is evident. The staircase artifacts are caused by finite element discretization.

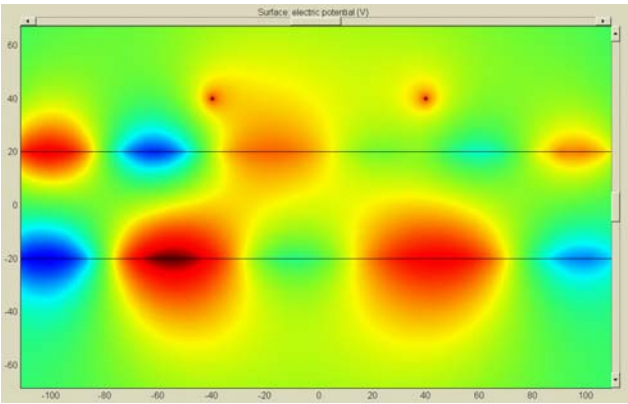


Fig. 7.46. The potential distribution for Pendry's lens example with two line sources; $\epsilon_{\text{slab}} = -0.9$.

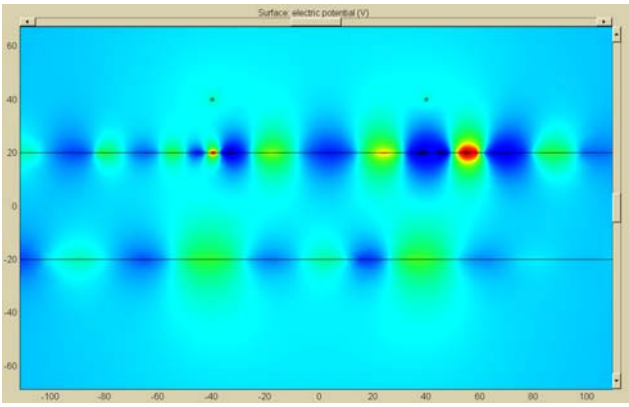


Fig. 7.47. The potential distribution for Pendry's lens example with two line sources; $\epsilon_{\text{slab}} = -1$.

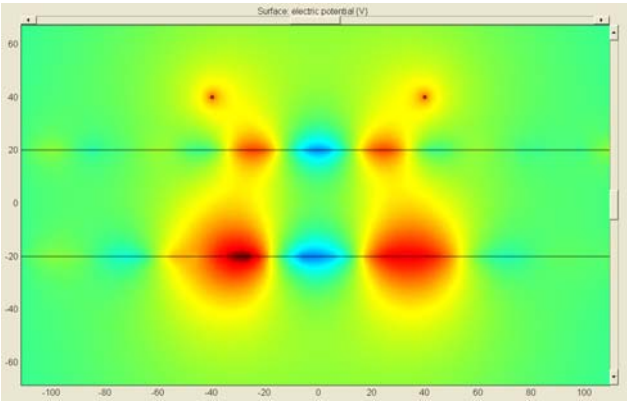


Fig. 7.48. The potential distribution for Pendry's lens example with two line sources; $\epsilon_{\text{slab}} = -1.02$.

Similarly, in Figs. 7.49–7.51 the imaginary part of the permittivity of the slab varies, with the real part fixed at -0.98847 as in Pendry's example. Again, the results are very different. As damping is increased, "multi-center" plasmon modes (no damping, Fig. 7.49) turn into two-center and then to one-center Teletubbies-like[49] distributions (Fig. 7.51).

[49] http://pbskids.org/teletubbies/parentsteachers/progmeet.html

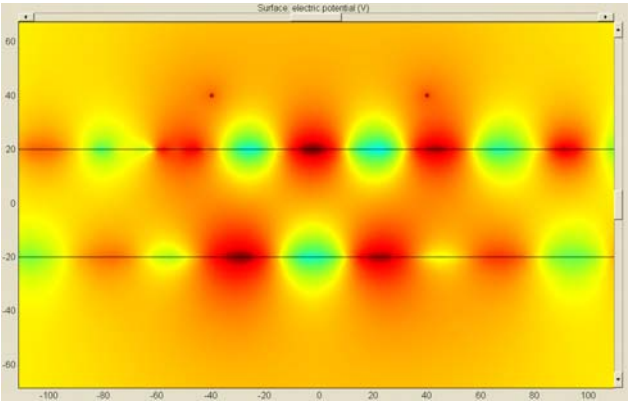


Fig. 7.49. The potential distribution for Pendry's lens example with two line sources; $\epsilon_{\text{slab}} = -0.98847$.

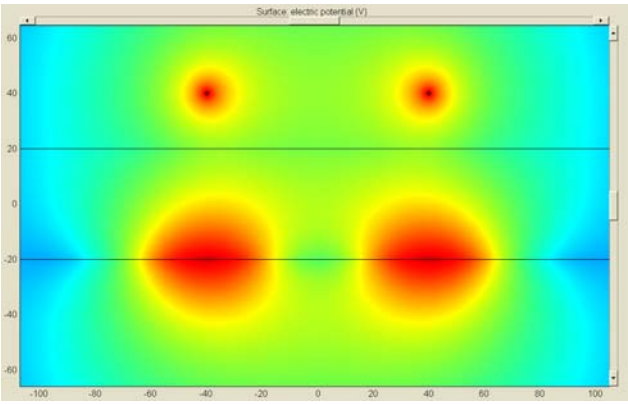


Fig. 7.50. The potential distribution for Pendry's lens example with two line sources; $\epsilon_{\text{slab}} = -0.98847 + 0.1\text{i}$.

7.13.3 Forward and Backward Plane Waves in a Homogeneous Isotropic Medium

In backward waves, energy and phase propagate in opposite directions (Section 7.13.1). We first examine this counterintuitive phenomenon in a hypothetical homogeneous isotropic medium with unusual material parameters (the "Veselago medium"). In subsequent sections, we turn to of forward and backward *Bloch* waves in periodic dielectric structures; plane-wave decomposition of Bloch waves will play a central role in that analysis.

Let us review the behavior of plane waves in a homogeneous isotropic medium with arbitrary constant complex parameters ϵ and μ at a given frequency. The only stipulation is that the medium be passive (no generation

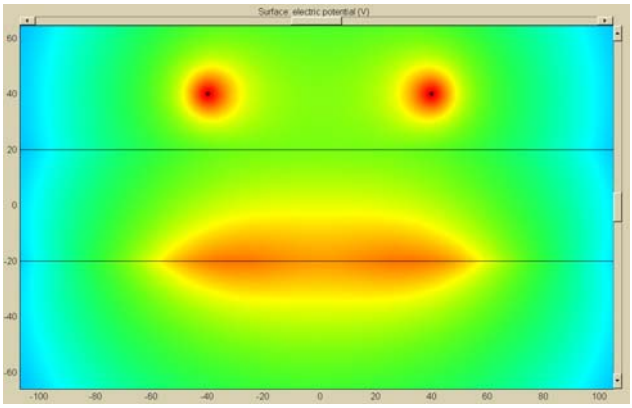


Fig. 7.51. The potential distribution for Pendry's lens example with two line sources; $\epsilon_{\text{slab}} = -0.98847 + 0.4\text{i}$.

of energy), which under the $\exp(+\text{i}\omega t)$ phasor convention implies negative imaginary parts of ϵ and μ. It will be helpful to assume that these imaginary parts are *strictly negative* and to view lossless materials as a limiting case of small losses: $\epsilon'' \to -0$, $\mu'' \to -0$. The goal is to establish conditions for the plane wave to be forward or backward. In the latter case, one has a "Veselago medium".

Let the plane wave propagate along the x axis, with $E = E_y$ and $H = H_z$. Then we have

$$E = E_y = E_0 \exp(-\text{i}kx) \tag{7.232}$$

$$H = H_z = H_0 \exp(-\text{i}kx) \tag{7.233}$$

where E_0, H_0 are some complex amplitudes. It immediately follows from Maxwell's equations that

$$H_0 = \frac{k}{\omega\mu} E_0 \tag{7.234}$$

$$k = \omega\sqrt{\mu\epsilon} \quad \text{(which branch of the square root?)} \tag{7.235}$$

Which branch of the square root "should" be implied in the formula for the wavenumber? In an unbounded medium, there is complete symmetry between the $+x$ and $-x$ directions, and waves corresponding to both branches of the root are equally valid. It is clear, however, that each of the waves is unbounded in one of the directions, which is not physical.

For a more physical picture, it is tacitly assumed that the unbounded growth is truncated: e.g. the medium and the wave occupy only half of the space, where the wave decays. With this in mind, let us focus on one of the two waves – say, the one with a negative imaginary part of k:

$$k'' < 0 \tag{7.236}$$

(The analysis for the other wave is completely analogous.) Splitting up the real and imaginary exponentials

$$\exp(-\mathrm{i}kx) \;=\; \exp(-\mathrm{i}(k' + \mathrm{i}k'')x) \;=\; \exp(k''x)\,\exp(-\mathrm{i}k'x)$$

we observe that this wave decays in the $+x$ direction. On physical grounds, one can argue that energy in this wave must flow in the $+x$ direction as well. This can be verified by computing the time-averaged Poynting vector

$$P \;=\; P_x \;=\; \frac{1}{2}\,\mathrm{Re}\,E_0 H_0^* \;=\; \frac{1}{2}\,\mathrm{Re}\,\frac{k}{\omega\mu}|E_0|^2 \tag{7.237}$$

To express P via material parameters, let

$$\epsilon \;=\; |\epsilon|\,\exp(-\mathrm{i}\phi_\epsilon); \qquad \mu \;=\; |\mu|\,\exp(-\mathrm{i}\phi_\mu); \qquad 0 < \phi_\epsilon, \phi_\mu < \pi$$

Then the square root with a negative imaginary part, consistent with the wave (7.236) under consideration, gives

$$k \;=\; \omega\sqrt{|\mu|\,|\epsilon|}\,\exp\left(-\mathrm{i}\,\frac{\phi_\epsilon + \phi_\mu}{2}\right) \tag{7.238}$$

Ignoring all positive real factors irrelevant to the sign of P in (7.237), we get

$$\mathrm{sign}\,P \;=\; \mathrm{sign}\,\mathrm{Re}\,\frac{k}{\mu} \;=\; \mathrm{sign}\,\cos\frac{\phi_\epsilon - \phi_\mu}{2}$$

The cosine, however, is always positive, as $0 < \phi_\epsilon, \phi_\mu < \pi$. Thus, as expected, P_x is positive, indicating that energy flows in the $+x$ direction indeed.

The type of the wave (forward vs. backward) therefore depends on the sign of phase velocity ω/k' – that is, on the sign of k'. As follows from (7.238),

$$\mathrm{sign}\,k' \;=\; \mathrm{sign}\,\cos\frac{\phi_\epsilon + \phi_\mu}{2}$$

and the wave is backward if and only if the cosine is negative, or

$$\phi_\epsilon + \phi_\mu \;>\; \pi \tag{7.239}$$

An algebraically equivalent criterion can be derived by noting that the cosine function is monotonically decreasing on $[0, \pi]$ and hence $\phi_\epsilon > \pi - \phi_\mu$ is equivalent to

$$\cos\phi_\epsilon \;<\; \cos(\pi - \phi_\mu)$$

or

$$\cos\phi_\epsilon \;+\; \cos\phi_\mu \;<\; 0$$

This coincides with the Depine–Lakhtakia condition [DL04] for backward waves:

$$\frac{\epsilon'}{|\epsilon|} + \frac{\mu'}{|\mu|} < 0 \tag{7.240}$$

This last expression is invariant with respect to complex conjugation and is therefore valid for both phasor conventions $\exp(\pm \mathrm{i}\omega t)$.

Note that the analysis above relies only on Maxwell's equations and the definitions of the Poynting vector and phase velocity. No considerations of causality, so common in the literature on negative refraction, were needed to establish the backward-wave conditions (7.239), (7.240).

7.13.4 Backward Waves in Mandelshtam's Chain of Oscillators

A classic case of backward waves in a chain of mechanical oscillators is due to L.I. Mandelshtam. His four-page paper [Man45][50] published by Mandelshtam's coworkers in 1945 after his death is very succinct, so a more detailed exposition below will hopefully prove useful. An electromagnetic analogy of this mechanical example (an optical grating) is the subject of the following section.

Consider an infinite 1D chain of masses, with the nearest neighbors separated by an equilibrium distance d and connected by springs with a spring constant f. Newton's equation of motion for the displacement ξ_n of the n-th mass m_n is

$$\ddot{\xi}(n) = \omega_n^2 \left[\xi(n-1) - 2\xi(n) + \xi(n+1)\right], \qquad \omega_n^2 = \frac{f}{m_n} \tag{7.241}$$

For brevity, dependence of ξ on time is not explicitly indicated. For waves at a given frequency ω, switching to complex phasors yields

$$\omega^2 \xi(n) + \omega_n^2 \left[\xi(n-1) - 2\xi(n) + \xi(n+1)\right] = 0 \tag{7.242}$$

Mandelshtam considers *periodic* chains of masses, focusing on the case with just two alternating masses, m_1 and m_2. The discrete analog of the Bloch wave then has the form

$$\xi(n) = \xi_{\mathrm{PER}}(n) \exp(-\mathrm{i}K_B n d) \tag{7.243}$$

where K_B is the Bloch wavenumber. ξ_{PER} is a periodic function of n with the period of two and can hence be represented by a Euclidean vector $\underline{\xi} \equiv (a, b) \in \mathbb{R}^2$, where a and b are the values of $\xi_{\mathrm{PER}}(n)$ for odd and even n, respectively.[51]

[50] The paper is also reprinted in Mandelshtam's lecture course [Man47].

[51] Alternatively and equally well, ξ_{PER} can be represented via its two-term Fourier sum, familiar from discrete-time signal analysis:

$$\xi_{\mathrm{PER}}(n) = \tilde{\xi}(0) + \tilde{\xi}(1)\exp(\mathrm{i}n\pi) = \tilde{\xi}(0) + (-1)^n \tilde{\xi}(1)$$

where

$$\tilde{\xi}(0) = \frac{1}{2}\left(\xi(0) + \xi(1)\right); \quad \tilde{\xi}(1) = \frac{1}{2}\left(\xi(0) - \xi(1)\right)$$

Substituting this discrete Bloch-type wave into the difference equation (7.242), we obtain

$$\begin{pmatrix} \omega_2^2(\lambda^2+1) & \lambda(\omega^2-2\omega_2^2) \\ \lambda(\omega^2-2\omega_1^2) & \omega_1^2(\lambda^2+1) \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = 0, \qquad \lambda \equiv \exp(-\mathrm{i}K_B d) \tag{7.244}$$

Hence (a, b) is the null vector of the 2×2 matrix in the left hand side of (7.244). Equating the determinant to zero yields two eigenfrequencies $\omega_{\mathrm{B1,B2}}$ of the Bloch wave

$$\omega_{\mathrm{B1,B2}} = \omega_1^2 + \omega_2^2 \pm \lambda^{-1}\sqrt{(\omega_1^2\,\lambda^2+\omega_2^2)\,(\omega_2^2\,\lambda^2+\omega_1^2)}$$

To analyze group velocity of Bloch waves, compute the Taylor expansion of these eigenfrequencies around $K_B = 0$ (keeping in mind that $\lambda = \exp(-\mathrm{i}K_B d)$):

$$\omega_{\mathrm{B1}} = 2\,\frac{d^2\omega_2^2\omega_1^2}{\omega_1^2+\omega_2^2}$$

$$\omega_{\mathrm{B2}} = 2(\omega_1^2+\omega_2^2) - 2\,\frac{d^2\omega_2^2\omega_1^2}{\omega_1^2+\omega_2^2}\,K_B^2$$

which coincides with Mandelshtam's formulas at the bottom of p. 476 of his paper. Group velocity $v_g = \partial\omega_{\mathrm{B}}/\partial K_B$ of long-wavelength Bloch waves is positive for the "acoustic" branch ω_{B1} but negative for the "optical" branch ω_{B2}.[52]

For $K_B = 0$ (i.e. $\lambda = 1$), simple algebra shows that the components of the second null vector $(a_{\mathrm{B2}}, b_{\mathrm{B2}})$ of (7.244) are proportional to the two particle masses:

$$\frac{a_{\mathrm{B2}}}{b_{\mathrm{B2}}} = -\frac{m_2}{m_1} \qquad (K_B = 0) \tag{7.245}$$

(The first null vector $a_{\mathrm{B1}} = b_{\mathrm{B1}}$ corresponding to the zero eigenfrequency for zero K_B represents just a translation of the chain as a whole and is uninteresting.)

Next, consider energy transfer along the chain. The force that mass $n-1$ exerts upon mass n is

$$F_{n-1,n} = [\xi(n-1) - \xi(n)]\,f$$

The mechanical "Poynting vector" is the power generated by this force:

$$P_{n-1,n}(t) = F_{n-1,n}(t)\,\dot{\xi}(n,t)$$

the time average of which, via complex phasors, is

$$\langle P_{n-1,n}\rangle = \frac{1}{2}\,\mathrm{Re}\,\{F_{n-1,n}\mathrm{i}\omega\xi(n)\}$$

[52] On the acoustic branch, by definition, $\omega \to 0$ as $K_B \to 0$; on optical branches, $\omega \nrightarrow 0$.

For the "optical" mode, i.e. the second eigenfrequency of oscillations, direct computation leads to Mandelshtam's expression

$$\langle P \rangle \;=\; \frac{1}{2}\, f\omega ab \sin(K_B d)$$

The subscripts for $\langle P \rangle$ have been dropped because the result is independent of n, as should be expected from physical considerations: no continuous energy accumulation occurs in any part of the chain.

We have now arrived at the principal point in this example. For small positive K_B ($K_B d \ll 1$), the Bloch wave has a long-wavelength component $\exp(-\mathrm{i}K_B nd)$. Phase velocity ω/K_B of the Bloch wave – in the sense discussed in more detail below – is positive. At the same time, the Poynting vector, and hence the group velocity, are negative because $a_{\mathrm{B}2}$ and $b_{\mathrm{B}2}$ have opposite signs in accordance with (7.245). Thus mechanical oscillations of the chain in this case propagate as a backward wave. An electromagnetic analogy of such a wave is mentioned very briefly in Mandelshtam's paper and is the subject of the following subsection.

Backward Waves in Mandelshtam's Grating

We now revisit Example 27 (p. 376) of a 1D volume grating, to examine the similarity with Mandelshtam's particle chain and the possible presence of backward waves. For definiteness, let us use the same numerical parameters as before and assume a periodic variation of the permittivity $\epsilon(x) = 2 + \cos 2\pi x$.

The Bloch–Floquet problem, in its algebraic eigenvalue form $\mathcal{K}^2 \underline{e} = \omega^2 \Xi \underline{e}$ (7.108), was already solved numerically in Example 27 (p. 375), and the band diagram was presented in Fig. 7.10.

We now discuss the splitting of the Poynting vector into the individual "Poynting components" $P_m = k_m |e_m|^2/(2\omega\mu)$ (7.118); this splitting has implications for the nature of the wave. The distribution of P_m for the first four Bloch modes in the grating is displayed in Fig. 7.52. The first mode shown in Fig. 7.52(a) is almost a pure plane wave ($P_{\pm 1}$ are on the order of 10^{-5}; $P_{\pm 2}$ are on the order of 10^{-13}, and so on) and does not exhibit any unusual behavior.

Let us therefore focus on mode #2 (upper right corner of the figure). There are four non-negligible harmonics altogether. The stems to the right of the origin ($K > 0$) correspond to plane wave components propagating to the right, i.e. in the $+x$-direction. Stems to the left of the origin correspond to plane waves propagating to the left, and hence their Poynting values are negative. It is obvious from the figure that the negative components dominate and as a result the total Poynting value for the Bloch wave is negative. The numerical values of the Poynting components and of the amplitudes of the plane wave harmonics are summarized in Table 7.3.

Now, the characterization of this wave as forward or backward hinges on the definition and sign of phase velocity. The smallest absolute value of the

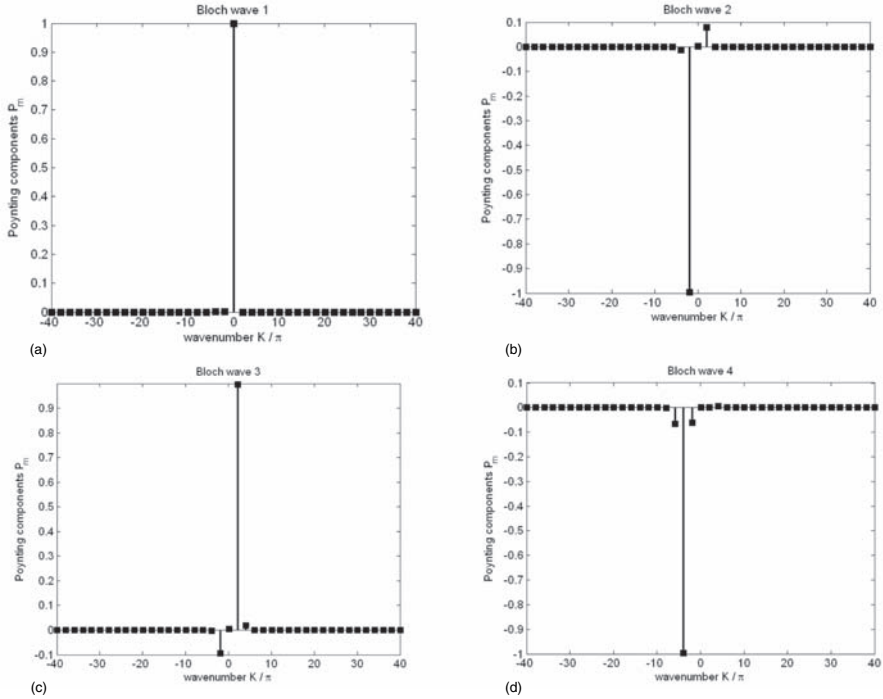


Fig. 7.52. The Poynting components P_m of the first four Bloch waves (a)–(d) for the volume grating with $\epsilon(x) = 2 + \cos 2\pi x$. Solution with 41 plane waves. $K_B x_0 = \pi/10$.

wavenumber in the Bloch "comb" $K_B = 0.1\pi$ determines the plane wave component with the longest wavelength (bold numbers in Table 7.3). If one defines phase velocity $v_{\text{ph}} = \omega/K_B$ based on $K_B = 0.1\pi$, then phase velocity is positive and, since the Poynting vector was found to be negative, one has a backward wave.

K/π	e_m	P_m
-5.9	-0.0023	-1.79×10^{-5}
-3.9	-0.0765	-0.013
−1.9	*−0.948*	*−0.997*
0.1	**0.174**	**0.00177**
2.1	0.253	0.0783
4.1	0.0179	0.000767
6.1	0.000495	8.73×10^{-7}

Table 7.3. The principal components of the second Bloch mode in the grating

However, the amplitude of the $K_B = 0.1\pi$ harmonic ($e_0 \approx 0.174$) is much smaller than that of the $K_B - \kappa_0 = -1.9\pi$ wave (italics in the Table). A common convention (P. Yeh [Yeh79], B. Lombardet *et al.* [LDFH05]) is to use this highest-amplitude component as a basis for defining phase velocity. If this convention is accepted in our present example, then phase velocity becomes negative and the wave is a forward one (since the Poynting vector is also negative).

One may then wonder what the value of phase velocity "really" is. This question is not a *mathematically* sound one, as one cannot truly argue about mathematical definitions. From the physical viewpoint, however, two aspects of the notion of phase velocity are worth considering.

First, boundary conditions at the interface between two *homogeneous* media are intimately connected with the values of phase velocities and indexes of refraction (defined for homogeneous materials in the usual unambiguous sense). Fundamentally, however, it is the wave vectors in both media that govern wave propagation, and it is the continuity of its tangential component that constrains the fields. Phase velocity plays a role only due to its direct connection with the wavenumber. For periodic structures, there is not one but a whole "comb" of wavenumbers that all need to be matched at the interface. We shall return to this subject in Section 7.13.5.

Second, in many practical cases phase velocity can be easily and clearly visualized. As an example, Fig. 7.53 shows two snapshots, at $t = 0$ and $t = 0.5$, of the second Bloch mode described above. For the visual clarity of this figure, low-pass filtering has been applied – without that filtering, the rightward motion of the wave is obvious in the animation but is difficult to present in static pictures. The Bloch wavenumber in the first Brillouin zone in this example is $K_B = 0.1\pi$ and the corresponding second eigenfrequency is $\omega \approx 4.276$. The phase velocity – if defined via the first Brillouin zone wavenumber – is $v_{\rm ph} = \omega/K_B \approx 4.276/0.1\pi \approx 13.61$. Over the time interval $t = 0.5$ between the snapshots, the displacement of the wave consistent with this phase velocity is $13.61 \times 0.5 \approx 6.8$. This corresponds quite accurately to the actual displacement in Fig. 7.53, proving that the first Brillouin zone wavenumber is indeed relevant to the perceived visual motion of the Bloch wave.

So, what is one to make of all this? The complete representation of a Bloch wave is given by a comb of wavenumbers $K_B - m\kappa_0$ and the respective amplitudes e_m of the Fourier harmonics. Naturally, one is inclined to distill this theoretically infinite set of data to just a few parameters that include the Poynting vector, phase and group velocities. While the Poynting vector and group velocity for the wave are rigorously and unambiguously defined, the same is *in general* not true for phase velocity.[53] However, there are practical

[53] As a mathematical trick, any finite (or even any countable) set of numbers can always be combined into a single one simply by intermixing the decimals: for example, $e = 2.71828\ldots$ and $\pi = 3.141592\ldots$ can be merged into $2.3711481258\ldots$. Of course this is not a serious proposition for us here.

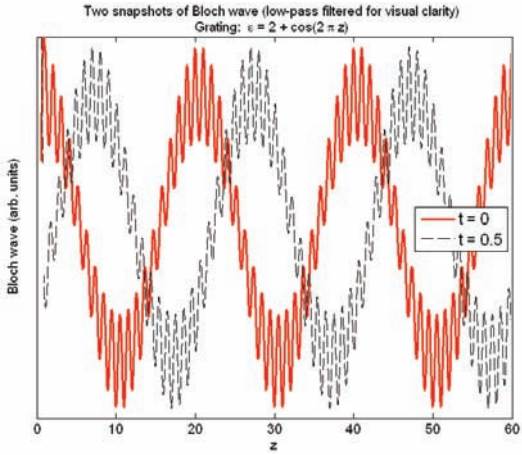


Fig. 7.53. Two snapshots, at $t = 0$ and $t = 0.5$, of the second Bloch mode. (Low-pass filtering applied for visual clarity.) The wave moves to the right with phase velocity corresponding to the smallest positive Bloch wavenumber $K_B = 0.1\pi$.

cases where phase velocity is meaningful. The situation is most clear-cut when the Bloch wave has a strongly dominant long-wavelength component. (This case will become important in Section 7.13.6.) Then the Bloch wave is, in a sense, close to a pure plane wave, but nontrivial effects may still arise. Even though the amplitudes of the individual higher-order harmonics may be small, it is possible for their *collective effect* to be significant. In particular, as the example in this section has shown, the higher harmonics taken together may carry more energy than the dominant component, and in the opposite direction. In this case one has a backward wave, where phase velocity is defined by the dominant long-wavelength harmonic, while the Poynting vector is due to a collective contribution of all harmonics.

An alternative generalization of phase velocity in 1D is the velocity v_{field} of points with a fixed magnitude of the E field. From the zero differential

$$dE \;=\; \frac{\partial E}{\partial x}\,dx \;+\; \frac{\partial E}{\partial t}\,dt \;=\; 0$$

one obtains

$$v_{\text{field}} = \frac{dx}{dt} \;=\; -\,\frac{\partial E}{\partial x}\Big/\frac{\partial E}{\partial t}$$

(see also equations (7.26), p. 354 and (7.37), p. 357). Unfortunately, this definition does not generalize easily to 2D and 3D, where an analogous velocity would be a tensor quantity (a separate velocity vector for each Cartesian component of the field).

7.13.5 Backward Waves and Negative Refraction in Photonic Crystals

Introduction

As already noted on p. 450, R. Zengerle in the late 1970s – early 1980s examined and observed negative refraction in singly and doubly periodic waveguides. In 2000, M. Notomi [Not00] noted similar effects in photonic crystals. For crystals with a sufficiently strong periodic modulation, there may exist a physically meaningful effective index of refraction within certain frequency ranges near the band edge. Under such conditions, anomalous refractive effects can arise at the surface of the crystal. Negative refraction is one of these possible effects. Another one is "open cavity" formation where light can run around closed paths in a structure with alternating positive-negative index of refraction (Fig. 7.54), even though there are no reflecting walls. Notomi's specific example involves TE modes in a 2D GaAs (index $n \approx 3.6$) hexagonal photonic crystal, with the diameter of the rods equal to 0.7 of the cell size.

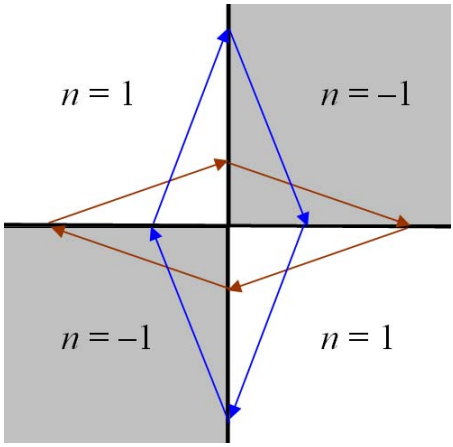


Fig. 7.54. [After M. Notomi [Not00].] "Open cavity" formation: light rays can form closed paths in a structure with alternating positive–negative index of refraction.

Since 2000, there have been a number of publications on negative refraction and the associated lensing effects in photonic crystals. To name just a few:

1. The photonic structure proposed by C. Luo *et al.* [LJJP02] is a bcc lattice of air cubes in a dielectric with the relative permittivity of $\epsilon = 18$. The dimension of the cubes is $0.75\,a$ and their sides are parallel to those of the lattice cell. The computation of the band diagram and equifrequency surfaces in the Bloch space, as well as FDTD simulations, demonstrate

"all-angle negative refraction" (AANR), i.e. negative refraction for all angles of the incident wave at the air–crystal interface. AANR occurs in the frequency range from $0.375(2\pi c/a)$ to $0.407(2\pi c/a)$ in the third band.

2. E. Cubukcu *et al.* [CAO+03] experimentally and theoretically demonstrate negative refraction and superlensing in a 2D photonic crystal in the microwave range. The crystal is a square array of dielectric rods in air, with the relative permittivity of $\epsilon = 9.61$, diameter 3.15 mm, and length 15 cm. The lattice constant is 4.79 mm. Negative refraction occurs in the frequency range from 13.10 to 15.44 GHz.
3. R. Moussa *et al.* [MFZ+05b] experimentally and theoretically studied negative refraction and superlensing in a triangular array of rectangular dielectric bars with $\epsilon = 9.61$. The dimensions of each bar are $0.40a \times 0.80a$, where the lattice constant $a = 1.5875$ cm. The length of each bar is 45.72 cm. At the operational frequency of 6.5 GHz, which corresponds to $\lambda_{\text{air}} \approx 4.62$ cm and $a/\lambda_{\text{air}} \approx 0.344$, the effective index is $n \approx -1$ with very low losses. Only TM modes are considered (the E field parallel to the rods.)
4. V. Yannopapas & A. Moroz [YM05] show that negative refraction can be achieved in a composite structure of polaritonic spheres occupying the lattice sites. A specific example involves $LiTaO_3$ spheres with the radius of 0.446 μm; the lattice constant is 1.264 μm, so that the fcc lattice is almost close-packed. Notably, the wavelength-to-lattice-size ratio is quite high, 14:1, but the relative permittivity of materials is also very high, on the order of 10^2.
5. M.S. Wheeler *et al.* [WAM06], independently of Yannopapas & Moroz, study a similar configuration. Wheeler *et al.* show that a collection of polaritonic spheres coated with a thin layer of Drude material can exhibit a negative index of refraction at infrared frequencies. The existence of negative effective magnetic permeability is due to the polaritonic material, while the Drude material is responsible for negative effective electric permittivity. The negative index region is centered at 3.61 THz, and the value of $n_{\text{eff}} = -1$, important for subwavelength focusing, is approached. The cores of the spheres are made of $LiTaO_3$ and their radius is 4 μm. The coatings have the outer radius of 4.7 μm, and their Drude parameters are $\omega_p/2\pi = 4.22$ THz, $\Gamma = \omega_p/100$. The filling fraction is 0.435.
6. S. Foteinopoulou & C.M. Soukoulis provide a general analysis of negative refraction at the air–crystal interfaces and, as a specific case, examine Notomi's example (a 2D hexagonal lattice of rods with permittivity 12.96 and the radius of 0.35 lattice size).
7. P.V. Parimi *et al.* [PLV+04] analyze and observe negative refraction and left-handed behavior of the waves in microwave crystals. The structure is a triangular lattice of cylindrical copper rods of height 1.26 cm and radius 0.63 cm. The ratio of the radius to lattice constant is 0.2. The TM-mode excitation is at frequencies up to 12 GHz. Negative refraction is observed, in particular, at 9.77 GHz.

For the analysis of anomalous wave propagation and refraction, it is important to distinguish *intrinsic* and *extrinsic* characteristics, as explained in the following subsection.

"Extrinsic" and "Intrinsic" Characteristics

This terminology, albeit not standard, reflects the nature of wave propagation and refraction in periodic structures such as photonic crystals and metamaterials. *Intrinsic* properties of the wave imply its characterization as either forward or backward; that is, whether the Poynting vector and phase velocity (if it can be properly defined) are in the same or opposite directions. (Or, more generally, at an acute or obtuse angle.)

Extrinsic properties refer to conditions at the interface of the periodic structure and air or another homogeneous medium. The key point is that refraction at the interface depends not only on the intrinsic characteristics of the wave in the bulk, but also on the way the Bloch wave is excited.

This can be illustrated as follows. Let the x axis run along the interface boundary between air and a material with x_0-periodic permittivity $\epsilon(x)$. For simplicity, we assume that ϵ does not vary along the normal coordinate y. Such a periodic medium can support Bloch E-modes of the form

$$E(\mathbf{r}) \;=\; \sum_{m=-\infty}^{\infty} e_m \exp(\mathrm{i}m\kappa_0 x)\, \exp(-\mathrm{i}K_{Bx}x)\, \exp(-\mathrm{i}K_y y)$$

Let the first-Brillouin-zone harmonic ($m = 0$) have an appreciable magnitude e_0, thereby defining phase velocity ω/K_{Bx} in the x-direction. For $K_{Bx} > 0$, this velocity is positive.

But any plane-wave component of the Bloch wave can serve as an "excitation channel"[54] for this wave, provided that it matches the x-component of the incident wave in the air:

$$K_{Bx} \;-\; \kappa_0 m \;=\; k_x^{\mathrm{air}}$$

First, suppose that the "main" channel ($m = 0$) is used, so that $K_{Bx} = k_x^{\mathrm{air}}$. If the Bloch wave in the material is a forward one, then the y-components of the Poynting vector P_y and the wave vector K_y are both directed *away* from the interface, and the usual positive refraction occurs. If, however, the wave is backward, then K_y is directed *toward* the surface (against the Poynting vector) and it can easily be seen that refraction is negative. This is completely consistent with Mandelshtam's explanation quoted on p. 448.

Exactly the opposite will occur if the Bloch wave is excited through an excitation channel where $K_{Bx} - \kappa_0 m$ is negative (say, for $m = 1$). The matching condition at the interface then implies that the x-component of the wave

[54] A lucid term due to B. Lombardet *et al.* [LDFH05].

vector in the air is negative in this case. Repeating the argument of the previous paragraph, one discovers that for a *forward* Bloch wave refraction is now *negative*, while for a backward wave it is positive.

In summary, refraction properties at the interface are a function of the intrinsic characteristics of the wave in the bulk *and* the excitation channel, with four substantially different combinations possible. This conclusion summarizes the results already available but dispersed in the literature [BST04, LDFH05, GMKH05].

Negative Refraction in Photonic Crystals: Case Study

To illustrate the concepts discussed in the sections above, let us consider, as one of the simplest cases, the structure proposed by R. Gajic, R. Meisels *et al.* [GMKH05, MGKH06]. Their photonic crystal is a 2D square lattice of alumina rods ($\epsilon_{\text{rod}} = 9.6$) in air. The radius of the rod is $r_{\text{rod}} = 0.61$ mm, the lattice constant $a = 1.86$ mm, so that $r_{\text{rod}}/a \approx 0.33$. The length of the rods is 50 mm. Gajic, Meisels *et al.* study various cases of wave propagation and refraction. In the context of this section, of most interest to us is negative refraction for small Bloch numbers in the second band of the H-mode.

The band diagram for the H-mode appears in Fig. 7.55. The diagram, computed using the plane wave method with 441 waves, is very close (as of course it should be) to the one provided by Gajic *et al.* Fig. 7.55 is plotted for the normalized frequency $\tilde{\omega} = \omega a/(2\pi c)$; in the Gajic paper, the diagram is for the absolute frequency $f = \omega/2\pi = \tilde{\omega}c/a$.

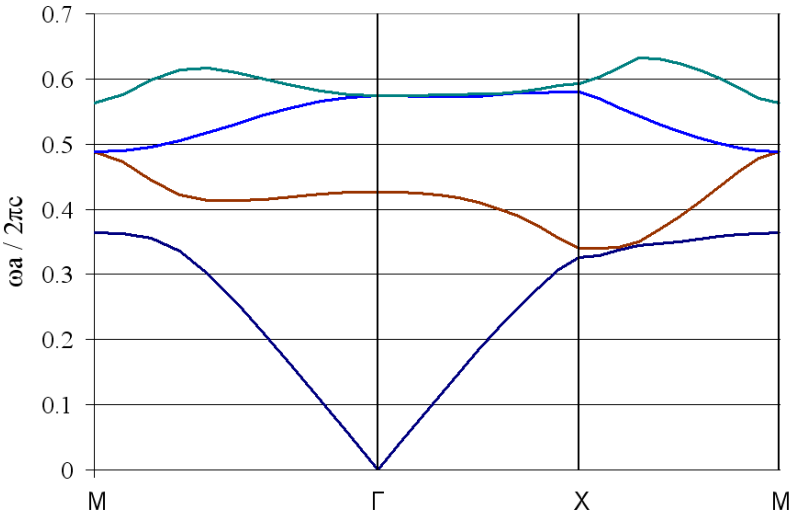


Fig. 7.55. The H-mode band diagram of the Gajic *et al.* crystal.

We observe that the TE2 dispersion curve is mildly convex around the Γ point ($K_B = 0$, $\tilde{\omega} \approx 0.427$), indicating a negative second derivative $\partial^2\omega/\partial K_B^2$ and hence a negative group velocity for small positive K_B and a possible backward wave. As we are now aware, an additional condition for a backward wave must also be satisfied: the plane-wave component corresponding to the small positive Bloch number must be appreciable (or better yet, dominant). Let us therefore consider the plane wave composition of the Bloch wave.

The amplitudes of the plane-wave harmonics for the Gajic *et al.* crystal are shown in Fig. 7.56. For $K_B = 0$ (i.e. at Γ) the spectrum is symmetric and characteristic of a standing wave. As K_B becomes positive and increases, the spectrum gets skewed, with the backward components ($K < 0$) increasing and the forward ones decreasing.

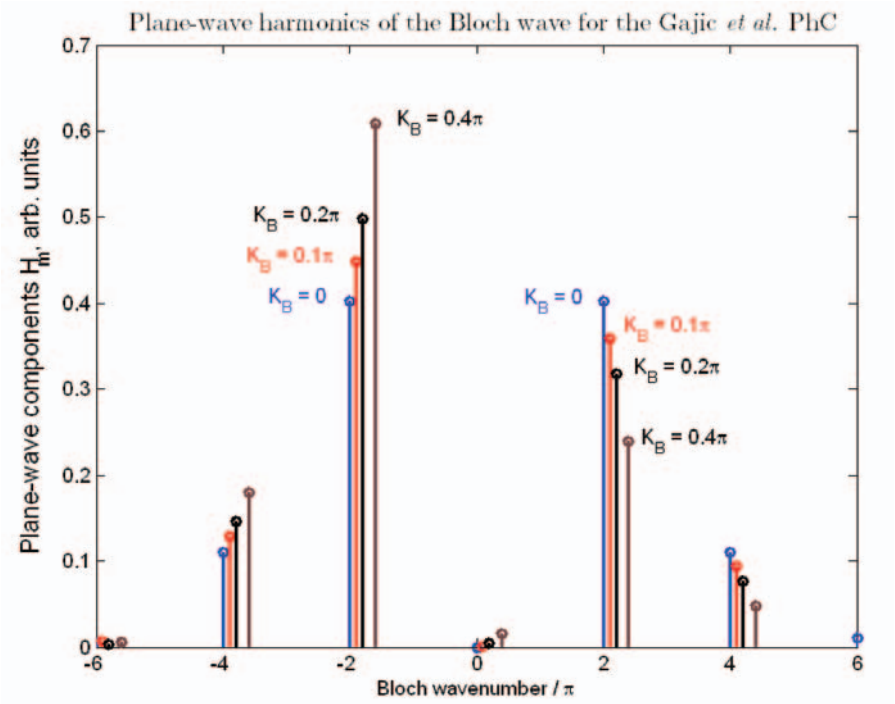


Fig. 7.56. Amplitudes h_m of the plane-wave harmonics for the Gajic *et al.* crystal (arb. units). Second H-mode (TE2) near the Γ point on the $\Gamma \to X$ line.

The numerical values of the amplitudes of a few spatial harmonics from Fig. 7.56 are also listed in Table 7.4 for reference. From the figure and table, it can be seen that the amplitudes of the spatial harmonics of this Bloch wave in the first Brillouin zone (the first four rows of numbers in the Table) are

quite small. It is therefore debatable whether a valid phase velocity can be attributed to this wave. The Bloch wave itself is pictured in Fig. 7.57 for illustration.

Normalized wavenumber $K_x a/\pi$	Amplitude h_m of the plane-wave harmonic
0	0
0.1	0.00124
0.2	0.00477
0.4	0.0159
2	0.4023
2.1	0.3589
2.2	0.3175

Table 7.4. Amplitudes of the spatial harmonics of the TE2 Bloch wave for the Gajic *et al.* photonic crystal.

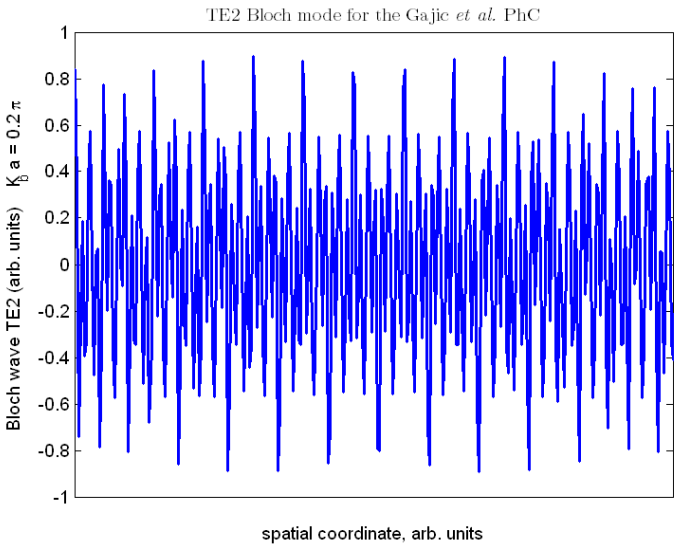


Fig. 7.57. The H field of the second H-mode (TE2, arb. units) for the Gajic *et al.* crystal. Point $K_B = 0.2\pi$ on the $\Gamma \to X$ line.

The distribution of Poynting components of the same wave and for the same set of values of the Bloch wavenumber is shown in Fig. 7.58. It is clear from the figure that the negative components outweigh the positive ones, so power flows in the negative direction.

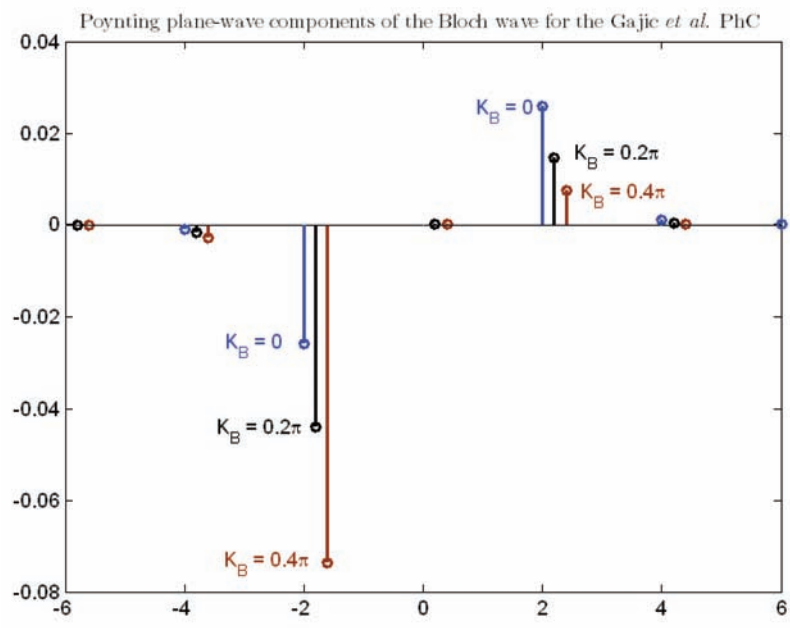


Fig. 7.58. The plane-wave Poynting components P_m for the Gajic *et al.* crystal (arb. units). Second H-mode (TE2) near the Γ point on the $\Gamma \to X$ line.

7.13.6 Are There Two Species of Negative Refraction?

Negative refraction is commonly classified as two species: first, homogeneous materials with double-negative effective material characteristics, as stipulated in Veselago's original paper [Ves68]; second, periodic dielectric structures (photonic crystals) capable of supporting modes with group and phase velocity at an obtuse angle to one another. The second category has been extensively studied theoretically, and negative refraction has been observed experimentally (see the list on p. 465 and Section 7.13.5).

Truly homogeneous materials, in the Veselago sense, are not currently known and could be found in the future only if some new molecular-scale magnetic phenomena are discovered. Consequently, much effort has been devoted to the development of artificial metamaterials capable of supporting backward waves and producing negative refraction. Selected developments of this kind are summarized in Table 7.5. (The numerical values in the Table are approximate.) The list is in no way exhaustive, and substantial further progress will almost certainly be made even before this book goes to press.

The right column of the table displays an important parameter: the ratio of the lattice cell size to the vacuum wavelength. One would hope that further improvements in nanofabrication and design could bring the cell size down to

Year	Publication	Design	f	λ	a	a/λ
2000	D.R. Smith *et al.* [SPV+00]	Copper SRR and wires	4.85 GHz	6.2 cm	8 mm	0.13
2001	R.A. Shelby *et al.* [SSS01]	Copper SRR and strips	10 GHz	3 cm	5 mm	0.17
2003	C.G. Parazzoli *et al.* [PGL+03]	A stack of SRRs with metal strips	12.6 GHz	2.38 cm	0.33 cm	0.14
2003	A.A. Houck *et al.* [HBC03]	Composite wire and SRR prisms	10 GHz	3 cm	0.6 cm	0.2
2004	D.R. Smith & D.C. Vier [SV04]	Copper SRR and strips	11 GHz	2.7 cm	3 mm	0.11
2005	V.M. Shalaev *et al.* [Sha06]	Pairs of nanorods	200 THz	1.5 m	$0.64 \times 1.8\ \mu$m	0.42–1.2
2005	S. Zhang *et al.* [ZFP+05]	Nano-fishnet (circular voids in metal)	150 THz	2 μm	0.838 μm	0.42
2005–2006	S. Zhang *et al.* [ZFM+05, ZFM+06]	Nano-fishnet with rectangular/ elliptical voids	215, 170 THz	1.4, 1.8 μm	0.801, 0.787 μm	0.57, 0.44
2006–2007	G. Dolling *et al.* [DEW+06, DWSL07]	Nano-fishnet with rectangular voids	210, 380 THz	1.45, 0.78 μm	0.6, 0.3 μm	0.41, 0.38

Table 7.5. Selected designs and parameters of negative-index metamaterials. The numerical values are approximate.

a small fraction of the wavelength, thereby approaching the Veselago case of a homogeneous material.

However, the main message of this section is that the cell size is constrained not only by the fabrication technologies. There are fundamental limitations on how small the lattice size can be for negative index materials. Homogeneous negative index materials may not in fact be realizable as a limiting case of spatially periodic dielectric structures with a small cell size.

The following analysis, available in a more detailed form in [Tsu07], shows that negative refraction disappears in the homogenization limit when the size of the lattice cells tends to zero, provided that other physical parameters, including frequency, are fixed. To streamline the mathematical development, let us focus on square Bravais lattice cells with size a in 2D and introduce dimensionless coordinates $\tilde{x} = x/a$, $\tilde{y} = y/a$, so that in these tilde-coordinates the 2D problem is set up in the unit square. (The 3D case is considered in [Tsu07].) The E-mode in the tilde-coordinates is described by the familiar 2D wave equation

$$\tilde{\nabla}^2 E \;+\; \tilde{\omega}^2 \epsilon_r E \;=\; 0, \tag{7.246}$$

where

$$\tilde{\omega} = \frac{\omega a}{c} \;=\; 2\pi \frac{a}{\lambda_0} \tag{7.247}$$

Here c and λ_0 are the speed of light and the wavelength in free space, respectively. The relative permittivity ϵ_r is a periodic function of coordinates over the lattice. The fundamental solutions of the field equation is a Bloch-Floquet wave; in the tilde-coordinates,

$$\mathbf{E}(\tilde{\mathbf{r}}) \;=\; \mathbf{E}_{\mathrm{PER}}(\tilde{\mathbf{r}}) \exp(-\mathrm{i}\tilde{\mathbf{K}}_B \cdot \tilde{\mathbf{r}}) \tag{7.248}$$

where $\tilde{\mathbf{r}}$ is the position vector. As in Section 7.6.2, it is convenient to view this Bloch wave as a suite of spatial Fourier harmonics (plane waves):

$$\mathbf{E}(\tilde{\mathbf{r}}) \;=\; \sum\nolimits_{\mathbf{n}} \mathbf{E}_{\mathbf{n}} \;\equiv\; \sum\nolimits_{\mathbf{n}} \tilde{\mathbf{e}}_{\mathbf{n}} \exp(\mathrm{i}2\pi \mathbf{n} \cdot \tilde{\mathbf{r}}) \exp(-\mathrm{i}\tilde{\mathbf{K}}_B \cdot \tilde{\mathbf{r}}) \tag{7.249}$$

(Summation in this and subsequent equations is over the integer lattice $\mathbb{Z}^2$.) As also noted in Section 7.6.2, the time- and cell-averaged Poynting vector $\langle \mathbf{P} \rangle \;=\; \frac{1}{2}\langle \mathrm{Re}\{\mathbf{E} \times \mathbf{H}^*\} \rangle$ can be represented as the sum of the Poynting vectors for the individual plane waves [LDFH05]:

$$\langle \mathbf{P} \rangle \;=\; \sum\nolimits_{\mathbf{n}} \mathbf{P}_{\mathbf{n}}; \qquad \mathbf{P}_{\mathbf{n}} \;=\; \frac{\pi n}{\tilde{\omega}\mu_0} |\tilde{\mathbf{e}}_{\mathbf{n}}|^2 \tag{7.250}$$

As we know, in Fourier space the scalar wave equation (7.246) becomes

$$\left|\tilde{\mathbf{K}}_B - 2\pi \mathbf{n}\right|^2 \tilde{e}_{\mathbf{n}} \;=\; \tilde{\omega}^2 \sum\nolimits_{\mathbf{m}} \tilde{\epsilon}_{\mathbf{n}-\mathbf{m}} \tilde{\mathbf{e}}_{\mathbf{m}}, \qquad \mathbf{n} \in \mathbb{Z}^2 \tag{7.251}$$

where $\tilde{\epsilon}_{\mathbf{n}}$ are the Fourier coefficients of the dielectric permittivity ϵ:

$$\epsilon = \sum\nolimits_{\mathbf{n}} \tilde{\epsilon}_{\mathbf{n}} \exp\left(\mathrm{i}2\pi\mathbf{n}\cdot\tilde{\mathbf{r}}\right) \tag{7.252}$$

The normalized band diagram, such as the one in Fig. 7.55, indicates that *negative refraction disappears in the homogenization limit* when the size of the lattice cells tends to zero, provided that other physical parameters, including frequency, are fixed. Indeed, the homogenization limit is obtained by considering the small cell size – long wavelength condition $a \to 0$, $\tilde{K} \to 0$ (see [SEK+05, Sj5] for additional mathematical details on Floquet-based homogenization theory for Maxwell's equations). As these limits are taken, the problem and the dispersion curves *in the normalized coordinates* remain unchanged, but the operating point $(\tilde{\omega}, \tilde{\mathbf{K}})$ approaches the origin along a fixed dispersion curve – the acoustic branch. In this case phase velocity in any given direction $\hat{l}$, $\omega/K_l = \tilde{\omega}/\tilde{K}_l$, is well defined and equal to group velocity $\partial\omega/\partial K_l$ simply by definition of the derivative. No backward waves can be supported in this regime.

This conclusion is not surprising from the physical perspective. As the size of the lattice cell diminishes, the operating frequency *increases*, so that it is not the absolute frequency ω but the normalized quantity $\tilde{\omega}$ that remains (approximately) constant. Indeed, a principal component of metamaterials with negative refraction is a resonating element [SPV+00, SV04, Ram05, Sha06] whose resonance frequency is approximately inverse proportional to size [LED+06].

It is pivotal here to make a distinction between *strongly* and *weakly* inhomogeneous cases of wave propagation. The latter is intended to resemble an ideal "Veselago medium," with the Bloch wave being as close as possible to a long-length plane wave. Toward this end, the following conditions characterizing the weakly inhomogeneous backward-wave regime are put forth:

- The first-Brillouin-zone component of the Bloch wave must be dominant; this component then defines the phase velocity of the Bloch wave.
- The other plane-wave components collectively produce energy flow at an obtuse angle to phase velocity.
- The lattice cell size a is small relative to the vacuum wavelength λ_0; $a/\lambda_0 \ll 1$.
- At the air-material interface, it is the long-wavelength, first-Brillouin-zone, plane wave component that serves as the excitation channel for the Bloch wave.

If any of the above conditions are violated, the regime will be characterized as *strongly inhomogeneous*: the EM wave can "see" the inhomogeneities of the material. By this definition, in the weakly inhomogeneous case the normalized Bloch wavenumber $\tilde{K}_B$ must be small, $\tilde{K}_B \equiv K_B a \ll \pi$. Larger values of K_B would indicate a strongly inhomogeneous (or, synonymously, "photonic crystal" or "grating") regime, where the lattice size is comparable with the Bloch wavelength. As we shall see, under reasonable physical assumptions, backward waves cannot be supported in the weakly inhomogeneous case; strong inhomogeneity is required.

As a preliminary step in the analysis, it is instructive to examine the direction of power flow for small $\tilde{K}_B$ in the lossless case (real ϵ). The average Poynting vector is, according to (7.250) and with a convenient normalization,

$$\tilde{P} \equiv 2\tilde{\omega}\mu_0\langle P\rangle = K_B \left|\tilde{e}_0(\tilde{K}_B)\right|^2 +$$

$$\sum\nolimits_{m=1}^{\infty}(\tilde{K}_B - 2\pi m)\left|\tilde{e}_m(\tilde{K}_B)\right|^2 + (\tilde{K}_B + 2\pi m)\left|\tilde{e}_{-m}(\tilde{K}_B)\right|^2 \tag{7.253}$$

The scalar form is used for notational convenience only; the vectorial case is quite similar. It is, however, essential to indicate explicitly that the Fourier amplitudes $\tilde{e}_m$ depend on the Bloch parameter $\tilde{K}_B$. Since the waves corresponding to $\pm\tilde{K}_B$ are complex conjugates of one another, we have $\tilde{e}_{-m}(\tilde{K}_B) = \tilde{e}_m^*(-\tilde{K}_B)$, and the expression for the Poynting vector becomes

$$\tilde{P} = \tilde{K}_B\left[\left|\tilde{e}_0(\tilde{K}_B)\right|^2 + \sum\nolimits_{m=1}^{\infty}\left(\left|\tilde{e}_m(\tilde{K}_B)\right|^2 + \left|\tilde{e}_m(-\tilde{K}_B)\right|^2\right)\right]$$

$$- 2\pi\sum\nolimits_{m=1}^{\infty} m\left(\left|\tilde{e}_m(\tilde{K}_B)\right|^2 - \left|\tilde{e}_m(-\tilde{K}_B)\right|^2\right) \tag{7.254}$$

The first line in (7.254) is directly proportional to $\tilde{K}_B$. To make this small parameter explicit in the second line as well, let us write

$$\tilde{P} = \tilde{K}_B\left[|\tilde{e}_0|^2 + \sum\nolimits_{m=1}^{\infty}\left(\left|\tilde{e}_m(\tilde{K}_B)\right|^2 + \left|\tilde{e}_m(-\tilde{K}_B)\right|^2\right)\right.$$

$$\left. - 2\pi\sum\nolimits_{m=1}^{\infty} m\,\frac{\partial\,|\tilde{e}_m|^2}{\partial\tilde{K}_B}\right] \tag{7.255}$$

For small $\tilde{\omega}$, the positive term $|\tilde{e}_0|^2$ in the square brackets tends to be dominant, making it difficult to produce a negative power flow and a backward wave. This is so because the magnitudes of all spatial harmonics *except for* $\tilde{e}_0$ are for small $\tilde{\omega}$ constrained by (7.251):

$$|\tilde{e}_n| \le \tilde{\omega}^2\left|\tilde{\mathbf{K}}_B - 2\pi\mathbf{n}\right|^{-2}\|\tilde{e}\|_{l_2}, \quad \|\tilde{e}\|_{l_2} = 1, \quad n \neq 0 \tag{7.256}$$

The arguments above suggest that there must be a lower bound for the relative cell size $a/\lambda_0 = \tilde{\omega}/2\pi$ when the medium could still support backward waves. Such a bound is derived below for *the weakly inhomogeneous backward-wave regime*, which implies that $\tilde{K}_B = K_B a \ll 1$. To simplify mathematical analysis, we focus on the limiting case $K_B = 0$, but the conclusions will apply, by physical continuity, to small $\tilde{K}_B$. We first turn to the E-mode governed by the 2D equation (7.246):

$$\epsilon E = -\eta\tilde{\nabla}^2 E, \quad \eta = \tilde{\omega}^{-2} \quad (\tilde{\omega} \neq 0) \tag{7.257}$$

Further analysis relies on the inversion of $\tilde{\nabla}^2$. To do this unambiguously, let us split E up into the zero-mean term $E_\perp$ and the remaining constant E_0: $E = E_0 + E_\perp$. Symbol '$\perp$' indicates orthogonality to the null space of the Laplacian (i.e. to constants). To eliminate the constant component E_0, we integrate (7.257) over the lattice cell. Integrating by parts and noting that the boundary term vanishes due to the periodic boundary conditions ($K_B = 0$), we get

$$E_0 = -\tilde{\epsilon}_0^{-1} \int_\Omega \epsilon E_\perp \, d\Omega, \qquad \tilde{\epsilon}_0 \neq 0$$

(The exceptional case $\tilde{\epsilon}_0 = 0$ is mathematically quite intricate and may constitute a special topic for future research.) With E_0 eliminated, the eigenvalue problem for $E_\perp$ becomes

$$\epsilon \left[E_\perp - \tilde{\epsilon}_0^{-1} \int_\Omega \epsilon \, E_\perp \, d\Omega \right] = -\eta \tilde{\nabla}^2 E_\perp$$

Since $E_\perp$ by definition is zero-mean,

$$\tilde{\nabla}_\perp^{-2} \left\{ \epsilon \left[E_\perp - \tilde{\epsilon}_0^{-1} (\epsilon, E_\perp) \right] \right\} = -\eta E_\perp \tag{7.258}$$

where $\tilde{\nabla}_\perp^{-2}$ is the zero-mean inverse of the Laplacian. Fourier analysis easily shows that this inverse is bounded (the Poincaré inequality): $\left\| \tilde{\nabla}_\perp^{-2} \right\| \leq (4\pi^2)^{-1}$. Then, taking the norm of both sides of (7.258), we get

$$|\eta| \leq (4\pi^2)^{-1} |\epsilon|_{\max} \left(1 + |\epsilon|_{\max} / |\tilde{\epsilon}_0|\right) \tag{7.259}$$

This result, that can be viewed as a generalization of the Poincaré inequality to cases with variable ϵ_r, leads to a simple lower bound for the lattice cell size, with the mean and maximum values of ϵ as parameters:

$$\left(\frac{a}{\lambda_0} \right)^2 = \frac{\tilde{\omega}^2}{4\pi^2} \geq \frac{1}{|\epsilon|_{\max} \left(1 + |\epsilon|_{\max} / |\tilde{\epsilon}_0|\right)} \tag{7.260}$$

This completes the theoretical derivation of lattice size bounds for the E-mode. Examples and analysis for the vector case can be found in [Tsu07]. The main conclusion is that for periodic structures capable of supporting backward waves and producing negative refraction, the lattice cell size, as a fraction of the vacuum wavelength and/or the Bloch wavelength, must be above certain thresholds. These thresholds depend on the maximum, minimum and mean values of the complex dielectric permittivity as key parameters. In the presence of good conductors (e.g. at microwave frequencies) such theoretical constraints are not very restrictive. However, at optical frequencies and/or for non-metallic structures the bounds on the cell size must be honored and may help to design metamaterials and photonic crystals with desired optical properties.

7.14 Appendix: The Bloch Transform

In Sections 7.4–7.10, we had an occasion to consider individual Bloch–Floquet waves in periodic structures. More generally, the electric or magnetic field can be represented via the *Bloch Transform* – a "continuous superposition" of Bloch waves. It can be viewed as a reformulation of the Fourier transform and is considered in this Appendix in 1D.

The general idea is rather simple: as we know, a Bloch wave in Fourier space is just a "comb" of plane waves at the spatial frequencies $K_B - m\kappa_0$ ($m = 0, \pm 1, \pm 2, \ldots$). A generic Fourier transform can be viewed as a continuous superposition of such combs for a varying K_B. The details are as follows.

Let a function $f(x)$ be expressed via its Fourier transform $F(k)$:

$$f(x) \;=\; \frac{1}{2\pi} \int_{-\infty}^{\infty} F(k) \exp(\mathrm{i}kx)\, dk \tag{7.261}$$

Let the k-axis be subdivided into intervals of equal length κ_0: $[m\kappa_0, (m+1)\kappa_0]$, $m = 0, \pm 1, \pm 2, \ldots$.

The Fourier transform $F(k)$ can then be viewed as a superposition of "Bloch combs"

$$\ldots,\; K_B - 2\kappa_0,\; K_B - \kappa_0,\; K_B,\; K_B + \kappa_0,\; K_B + 2\kappa_0,\; \ldots$$

with $K_B \in [0, \kappa_0]$; one such comb is shown in Fig. 7.59 for illustration. More

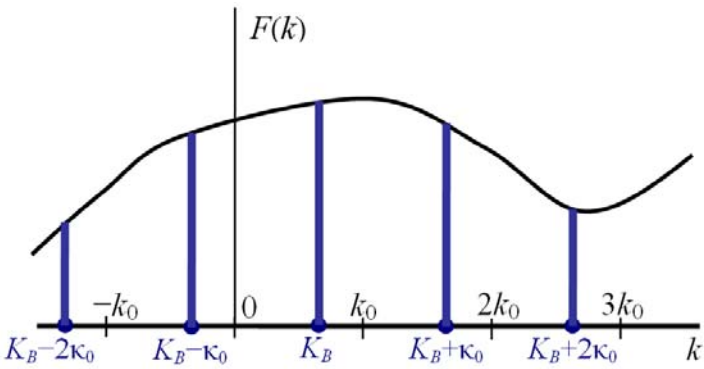


Fig. 7.59. From Fourier to Bloch: Fourier transform as a superposition of "Bloch combs".

precisely, Fourier integration can be broken up into "combs" by setting $k = K_B - m\kappa_0$, which yields

$$f(x) \;=\; \frac{1}{2\pi}\int_0^{\kappa_0}\left\{\sum_{m=-\infty}^{\infty} F(K_B - m\kappa_0)\exp(\mathrm{i}m\kappa_0 x)\right\}\exp(\mathrm{i}K_B x)\,dK_B \tag{7.262}$$

For a fixed K_B, the expression in the curly brackets is a Fourier series, with the coefficients $F_m = F(K_B - m\kappa_0)$. The sum of this series is a periodic function of x, with the period $x_0 = 2\pi/\kappa_0$. Let us denote this sum with $f_{\mathrm{PER}}(K_B, x)$, where "PER" implies periodicity with respect to x:

$$f_{\mathrm{PER}}(K_B, x) \;\equiv\; \sum_{m=-\infty}^{\infty} F(K_B - m\kappa_0)\exp(\mathrm{i}m\kappa_0 x) \tag{7.263}$$

With this in mind, the Bloch transform is, in essence, Fourier transform in terms of the Bloch variable K_B:

$$f(x) \;=\; \frac{1}{2\pi}\int_0^{\kappa_0} f_{\mathrm{PER}}(K_B, x)\exp(\mathrm{i}K_B x)\,dK_B \tag{7.264}$$

7.15 Appendix: Eigenvalue Solvers

The computational work in the bandgap structure calculation is dominated by the solution of the eigenvalue problem. This area of numerical analysis is remarkably rich in ideas and includes several classes of methods. An excellent compendium of computational strategies for solving eigenvalue problems is the *Templates* book [BDD$^+$00]. The following quote from this book identifies the key questions for choosing the best strategy:

1. Mathematical properties of the problem: Is it a Hermitian (real symmetric, self-adjoint) or a non-Hermitian eigenproblem? Is it a standard problem involving only one matrix or a generalized problem with two matrices?
2. Desired spectral information: Do we need just the smallest eigenvalue, a few eigenvalues at either end of the spectrum, a subset of eigenvalues "inside" the spectrum, or most eigenvalues? Do we want associated eigenvectors, invariant subspaces, or other quantities? How much accuracy is desired?
3. Available operations and their costs: Can we store the full matrix as an array and perform a similarity transformation on it? Can we solve a linear system with the matrix (or shifted matrix) by a direct factorization routine, or perhaps an iterative method? Or can we only multiply a vector by the matrix, and perhaps by its transpose? If several of these operations are possible, what are their relative costs?

The *Templates* book is perhaps the best starting point for studying eigenvalue algorithms. G.H. Golub & H.A. van der Vorst have written excellent reviews

of the subject [GvdV00, vdV04]. In addition to the classical monographs by J.H. Wilkinson [Wil65], G.H. Golub & C.F. van Loan [GL96] and B.N. Parlett [Par80], more recent books by Y. Saad [Saa92], J.W. Demmel [Dem97] and L.N. Trefethen & D. Bau [Tre97] are highly recommended.

This section summarizes the eigenvalue methods relevant to the PBG analysis. The answers to the *Templates* questions quoted above are different for Fourier-space and finite-element algorithms; we therefore deal with these two approaches separately.

Let us start with plane-wave expansion and assume that a regular (rather than generalized) eigenvalue problem is solved; this is the case when all coordinate-dependent quantities are incorporated into the differential operator on the left hand side of the eigenvalue equation and do not appear on the right. The system matrix is generally full because products in real space turn into convolutions in the Fourier space, with all spatial harmonics coupled. The matrix is Hermitian for the H-problem and non-Hermitian for the E-problem.

The standard solution procedure has two stages. At the first stage, the matrix is converted to a simpler form – usually *upper Hessenberg* form (upper triangular plus one lower subdiagonal) – by a sequence of orthogonal similarity transformations. The transformations are usually *Householder reflections* with respect to judiciously chosen hyperplanes. The purpose of conversion to upper Hessenberg form is to make QR iterations (see below) much more efficient.

Fig. 7.60 gives a geometric illustration of a Householder reflection. The picture is drawn just in two dimensions because the hypervolume of this book is limited. The hyperplane of reflection is orthogonal to a certain vector u that does not have to be a unit vector.

The projection of an arbitrary vector x onto u is easily calculated to be

$$x_u \;=\; u\,\frac{u^T x}{u^T u}$$

The "mirror image" of x with respect to the hyperplane is therefore

$$x' \;=\; x - 2x_u \;=\; x - 2u\,\frac{u^T x}{u^T u} \;=\; \left(I - 2\,\frac{uu^T}{u^T u}\right) x$$

The matrix in the parentheses formally defines the Householder reflection and is easily shown to be symmetric and orthogonal. It can also be demonstrated (see any of the monographs cited above) that a series of suitable Householder transforms will convert any matrix, column by column, to upper Hessenberg ("almost triangular") form. Since orthogonal similarity transforms preserve symmetry, for a symmetric (Hermitian in the complex case) matrix the upper Hessenberg form is also symmetric (Hermitian) and hence tridiagonal.

The second stage of the typical eigenvalue algorithm consists in applying QR-iterations to the upper Hessenberg (tridiagonal in the symmetric case)

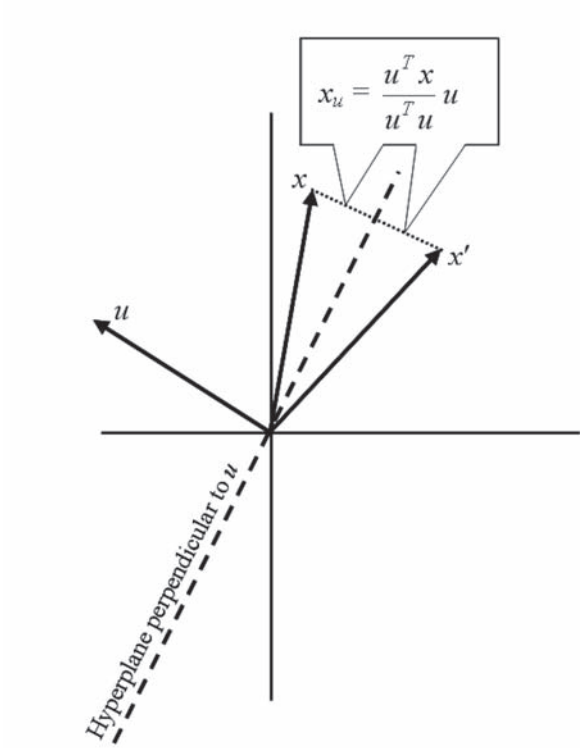


Fig. 7.60. A geometric illustration of Householder reflection. An arbitrary vector x is reflected with respect to a hyperplane orthogonal to some vector u.

matrix.[55] As known from linear algebra, any matrix A can be factored as

$$A = QR \tag{7.265}$$

where Q is an orthogonal matrix ($Q^T Q = I$) in the real case or a unitary matrix $Q^* Q = I$) in the complex case; R is an upper triangular matrix. This factorization is applicable not only to square matrices, but to rectangular $m \times n$ matrices with $m \geq n$ as well. However, we only need to consider square matrices here.

The QR decomposition is not a similarity transform and therefore is not by itself suitable for eigenvalue analysis. A similarity transform is obtained by multiplying Q and R in the opposite order:

$$A' = RQ = Q^* A Q \tag{7.266}$$

[55] QR *iterations* should not be confused with QR *decomposition*. The iterative process does have QR factorization as its central part but also involves RQ-multiplication and shifts, as summarized in the text below.

where the complex case is assumed for generality and R is expressed as Q^*A from (7.265). Remarkably, by repeating these two operations (QR-decomposition, then RQ-multiplication) one obtains a sequence of matrices A, A', A'', ... that usually converges very rapidly to a triangular matrix (diagonal in the symmetric/Hermitian case). [56] Since similarity transforms preserve the eigenvalues, the diagonal of this triangular matrix in fact contains the eigenvalues of the original matrix. The eigenvectors of the original matrix can be computed by "undoing" the orthogonal similarity transforms.

The first stage of the overall procedure – the transformation to upper Hessenberg (or tridiagonal, in the symmetric/Hermitian case) form – for a full $n \times n$ matrix requires $\mathcal{O}(n^3)$ operations. QR iterations, the second stage of the procedure, typically exhibit quadratic convergence at least; hence the number of iterations per eigenvalue is virtually independent of the size of the system, and the total operation count for the QR stage is $\mathcal{O}(n^2)$ if only the eigenvalues are sought. If the eigenvectors are also needed, the cost is $\mathcal{O}(n^3)$.

In summary, the total computational cost for the standard eigenvalue solver with Householder reflections and QR iterations is, as a rule, asymptotically proportional to the cube of the system size. In practical applications to photonic bandgap structure calculation, the computational cost limits the number of PWE terms, and hence the accuracy of the solution.

QR-based methods are usually viewed as direct solvers. An alternative is **iterative eigenvalue solvers**. Strictly speaking, *all* eigenvalue solvers for systems of dimension greater than four are, by necessity, iterative (as there is no general way of computing the roots of the corresponding characteristic polynomial in a finite number of operations). However, due to fast convergence of QR iterations, the direct part (Householder reflections) is dominant, and for practical purposes QR solvers are viewed as direct.

Iterative methods do not require explicit access to the matrix entries, as long as a matrix-vector multiplication procedure is available. The system matrix (or matrices, for the generalized eigenproblem) remains unchanged in the course of the iterations, and hence for sparse matrices no additional fill-in is created (i.e. zero entries remain zero). Moreover, in large-scale finite element simulations matrix-vector operations can be carried out on an element-by-element basis, without having to store the entries of the global matrix. However, additional memory is needed for an auxiliary set of orthogonal vectors, as described below.

The literature on iterative methods is vast and the algorithms are quite elaborate. Here I only highlight the main ideas, from the perspective of solving

[56] This algorithm, however, is not infallible. In practice, it is implemented with shifts: the QR factorization is applied not to the original matrix A itself but to $A - sI$, where the shift s may vary from iteration to iteration. Algorithms for choosing the shifts can be quite involved. A very interesting note on the Matlab implementation of the QR algorithm was written by Cleve Moler in 1995 and is posted on the Mathworks website: www.mathworks.com/company/newsletters/news_notes/pdf/sum95cleve.pdf

photonic bandgap problems; further details and references can be found in the monographs and review papers already cited on p. 478.

A key part of iterative algorithms is projection onto a suitably chosen subspace – in practice, with a small number of dimensions. Let us consider an eigenvalue problem $Ax = \lambda x$, where A is an $n \times n$ matrix and $x \in \mathbb{C}^n$, and assume that an orthonormal set of m vectors $q_k \in \mathbb{C}^n$, $k = 1, 2, \ldots, m$ is available. Construction of this set is the second key part of the procedure and a defining feature of particular classes of iterative methods.

To find an approximate solution of the eigenvalue problem within the subspace $\mathcal{Q}_m$ spanned by vectors q_k, a natural (but not the only) option is to apply the Galerkin method. The approximate solution x_m is a linear combination of m basis vectors q_k, and the same vectors are used to test the fidelity of this solution. In matrix form,

$$x_m = Q_m c_m, \qquad c_m \in \mathbb{C}^n$$

where the $n \times m$ matrix Q_m comprises the m column vectors q_k and c_m is a coefficient vector. The Galerkin equations are

$$(Ax_m,\ q_k) = \mu(x_m,\ q_k), \qquad k = 1, 2, \ldots, m$$

where μ is an approximate eigenvalue. Substituting $x_m = Q_m c_m$ and putting the system in matrix-vector form leads to

$$Q_m^* A Q_m c_m = \mu c_m \tag{7.267}$$

The right hand side got simplified due to the orthogonality of the columns of Q_m: $Q_m^* Q_m = I_m$. This eigenvalue problem reduced to subspace $\mathcal{Q}_m$ has m eigensolutions that are called the Ritz values and Ritz vectors of A with respect to $\mathcal{Q}_m$.

There are two main approaches for constructing the orthonormal sequence of vectors q_k. The first one involves the *Krylov spaces*. By definition, these spaces are spanned by the Krylov vectors y_1, $y_2 = Ay_1$, $y_3 = Ay_2 = A^2 y_1$, …, $y_m = A^{m-1} y_1$, where y_1 is some starting vector. The orthonormal basis q_k in the Krylov space is obtained by the modified Gram–Schmidt algorithm.[57]

Let us focus on the non-degenerate case where the n Krylov vectors, up to y_n, are linearly independent, and assemble these vectors as columns into an $n \times n$ matrix Y. Since the dimension of the whole space is n, vector $y_{n+1} = Ay_n$ must be a linear combination of the n Krylov vectors, with some coefficients $-s$. (The minus sign is included for compatibility with the conventional form of the "companion matrix" S below.) Then we have [Dem97]

$$AY = YS \tag{7.268}$$

[57] The modified version of the Gram–Schmidt algorithm is mathematically equivalent to the classical one but is more stable, due to the way the orthogonalization coefficients are calculated. See e.g. J.W. Demmel [Dem97], p. 107.

where

$$S = \begin{pmatrix} 0 & 0 & \ldots\ldots & 0 & -s_1 \\ 1 & 0 & \ldots\ldots & 0 & -s_2 \\ 0 & 1 & \ldots\ldots & 0 & -s_3 \\ \ldots & 0 & 1 \ \ldots & 0 & -s_4 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & 1 & -s_n \end{pmatrix} \tag{7.269}$$

This is just a matrix-form expression of two facts: (i) multiplying A with each column of matrix Y (i.e. with each Krylov vector) except for the very last one produces, by construction of the Krylov sequence, the next column; (ii) multiplying A by the last column of Y produces a vector that is a linear combination of all columns.

Since S is upper-Hessenberg, the matrix identity (7.268) can be interpreted as a transformation of the original matrix A to the upper Hessenberg form:

$$Y^{-1}AY = S \tag{7.270}$$

Although this transformation is non-orthogonal and computationally not robust, its conceptual importance is in establishing the connection between the Krylov vectors and upper Hessenberg matrices. An *orthogonal* transformation can be obtained by the QR-decomposition of Y, which is nothing other than the Gram–Schmidt procedure already mentioned:

$$Y = Q_Y R_Y$$

Upon substitution into (7.270), this yields

$$R_Y^{-1} Q_Y^* A Q_Y R_Y = S$$

or equivalently

$$Q_Y^* A Q_Y = H \equiv R_Y S R_Y^{-1} \tag{7.271}$$

where H is upper-Hessenberg because S is.

This conversion of the original matrix to upper Hessenberg form using the orthogonalized Krylov sequence is known as the *Arnoldi* algorithm. In the Hermitian case, this algorithm simplifies dramatically: the upper Hessenberg matrix H becomes tridiagonal and the orthogonal vector sequence can be generated by a simple three-term recurrence, as described in all texts on numerical linear algebra. This procedure for complex Hermitian or real symmetric matrices is known as the *Lanczos* method.

The Lanczos method is in exact arithmetic direct, in the sense that it orthogonally transforms the matrix to a tridiagonal one in a finite number of operations (after which the eigenproblem can be efficiently solved by QR iterations). However, in practice the Lanczos method and numerous other related Krylov-space algorithms are used almost exclusively as *iterative* solvers to compute the Ritz approximations (7.267) to the eigenpairs. The main reason

is that roundoff errors destroy orthogonality, so the procedure cannot be successfully completed in finite-precision arithmetic without additional measures such as reorthogonalization.

The highest and lowest eigenvalues in the Lanczos method converge faster as a function of m in (7.267) than the values in the interior of the spectrum. Applying the algorithm to a shifted and inverted matrix $(A - \sigma I)^{-1}$ instead of A will yield faster convergence for the eigenvalues closest to the shift σ. This, however, necessitates solving linear systems with matrix $(A - \sigma I)$, which may be prohibitively expensive for large matrices. Many preconditioning techniques, that can be viewed as approximate inverses in some sense, have been developed to overcome this difficulty. In addition to the review papers by H.A. van der Vorst [vdV04] and G.H. Golub & van der Vorst [GvdV00] already cited, see papers by A.V. Knyazev [Kny01], P. Arbenz [AHLT05] and references therein.

An interesting alternative to the Krylov-subspace solvers is the Jacobi–Davidson method. Van der Vorst [vdV04] notes the origin of this method in the 1846 paper by C.G.J. Jacobi [Jac46], the 1975 paper by E.R. Davidson [Dav75] and, finally, the 1996 paper by G.L.G. Sleijpen & H.A. van der Vorst [SvdV96].

An instructive way to view the Jacobi–Davidson method is as Newton–Raphson iterations [vdV04]. Let a unit vector y and a number θ (real for the Hermitian case) be an approximation to an eigenvector/eigenvalue pair of matrix A, so that

$$y^* A y \;=\; \theta \tag{7.272}$$

We are looking for suitable corrections Δy and $\Delta\theta$ that would improve this approximation. Since it does not make sense to update y along its own direction, Δy is sought to be orthogonal to y:

$$y^* \Delta y \;=\; 0 \tag{7.273}$$

The target condition for the corrections is

$$A(y + \Delta y) \;=\; (\theta + \Delta\theta)(y + \Delta y) \tag{7.274}$$

Ignoring the second order product $\Delta\theta\Delta y$ as in the Newton–Raphson linearization and moving the Δ-terms to the left hand side yields

$$(A - \theta I)\Delta y \;-\; \Delta\theta\, y \;=\; -\,(A - \theta I)y \tag{7.275}$$

Since the left hand side contains two unknown increments, a relationship between $\Delta\theta$ and Δy needs to be established to close the procedure. This can be done by pre-multiplying the equation for the corrections (7.273) with y^* [vdV04]:

$$y^* A(y + \Delta y) \;=\; y^*(\theta + \Delta\theta)(y + \Delta y)$$

After taking into account the orthogonality condition $y^*\Delta y = 0$, the eigenvalue condition $y^* A y = \theta$ and the unit length of y ($y^* y = 1$), one obtains

$$y^* A \, \Delta y \;=\; \Delta\theta \tag{7.276}$$

Substituting this $\Delta\theta$ into the Newton–Raphson formula (7.275) yields an equation for Δy alone:

$$(I - yy^*)(A - \theta I)\Delta y \;=\; -(A - \theta I)y$$

Since $y^*\Delta y = 0$, the last equation can be symmetrized:

$$(I - yy^*)(A - \theta I)(I - yy^*)\Delta y \;=\; -(A - \theta I)y \tag{7.277}$$

The entries of matrix yy^* are pairwise products of the components of y, and hence matrix $(I - yy^*)$ is in general full. However, it need not be computed explicitly; to carry out the iterative procedure, one only needs the product of $(I - yy^*)$ with an arbitrary vector z, which is easily calculated right-to-left: $(I - yy^*)z = z - y(y^*z)$.

The meaning of $(I - yy^*)$ as a projection operator is clear from Fig. 7.61. Indeed, y^*z (a scalar) is geometrically the length[58] of the projection of z onto y; $y(y^*z)$ is this projection itself; finally, $z - y(y^*z)$ is the projection of z onto the hyperspace orthogonal to y.

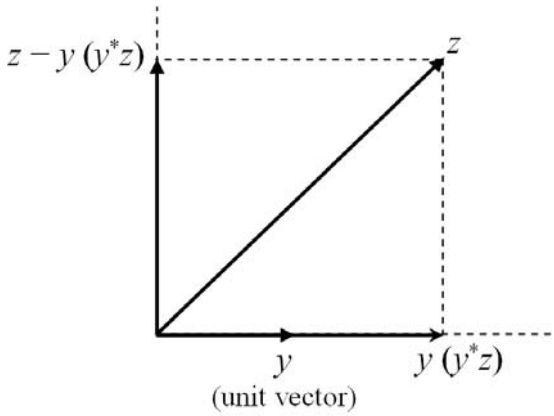


Fig. 7.61. A geometric illustration of the $(I - yy^*)$ projector.

In the Jacobi–Davidson method, the Newton–Raphson iteration outlined above defines a new direction Δy to be added to the Ritz approximation space. The complete Jacobi–Davidson algorithm is quite involved and includes, in addition to the computation of Δy: modified Gram–Schmidt orthogonalization; restarting procedures when the dimension of the Ritz subspace gets too

[58] For the geometric interpretation, vectors should be viewed as real.

large; solution of the eigenvalue problem in the subspace; deflation (i.e. projection onto a hyperspace orthogonal to the eigenvectors already found – a generalization of $I - yy^*$). Furthermore, preconditioning for eigensolvers in general, and within the Jacobi–Davidson algorithm in particular, is a more complicated matter than for linear system solvers. Parallel implementation of these algorithms is a very rich area of research as well. On all these subjects (except for parallel eigensolvers), the *Templates* book is again a comprehensive and condensed initial source of information; in addition, see papers by G.L.G. Sleijpen & F.W. Wubs [SW03], A.V. Knyazev [Kny98, Kny01], A. Basermann [Bas99, Bas00].

What is, then, the bottom line for the photonic bandgap computation as far as eigensolvers are concerned? Methods that are based on plane wave (or spherical/cylindrical wave) expansion lead to dense matrices. For problems of small or moderate size, direct methods are typically the best choice.

However, large-scale problems are common in computational practice: tens of thousands or more unknowns are often needed for adequate accuracy. At the same time, only a small number of low-frequency eigenmodes may be of interest. In such cases, iterative eigensolvers have an advantage. S.G. Johnson & J.D. Joannopoulos, in their highly cited paper [JJ01], apply preconditioned conjugate-gradient minimization of the block Rayleigh quotient or, alternatively, the Davidson method.

Finite element analysis leads to *generalized* eigenproblems, with the "stiffness" matrix on the left and the "mass" matrix on the right. Both matrices are sparse. For regular media with symmetric material tensors, the matrices are Hermitian; the stiffness matrix is nonnegative definite, and the mass matrix is positive definite. The use of direct solvers is justified only for small-size problems; the relevant direct solver for the generalized eigenproblem is the QZ method by C.B. Moler & G.W. Stewart [MS73]. A review of iterative methods for this type of problem is given in the *Templates* [BDD$^+$00]. W. Axmann & P. Kuchment [AK99] offer an interesting comparison of FEM with plane wave expansion; they use the "simultaneous coordinate overrelaxation method" as an eigenvalue solver for finite-element photonic (and acoustic) bandgap calculation in two dimensions.

8

Conclusion: "Plenty of Room at the Bottom" for Computational Methods

Google returns 57,000 references to R. Feynman's talk "There's Plenty of Room at the Bottom" [Fey59]. This talk, given on December 29, 1959 at the annual meeting of the American Physical Society, presaged the development of nanoscale science and technology.

As a 57,001st reference to Feynman, I argue in this book that there is "plenty of room at the bottom" for modeling and simulation. This point is illustrated with a number of examples: computation of long-range interactions between charged or polarized particles in free space and in heterogeneous media; plasmonic field enhancement by metal layers or cascades of particles; the bandgap structure and light waves in photonic crystals; scattering and enhancement of light in near-field optical microscopy; backward waves, negative refraction and perfect lensing.

The content of the book lies in the intersection of computational methods and applications, especially to problems on the nanoscale. Whenever possible, I have tried to give a common-sense explanation of mathematical, physical and computational ideas, hoping that sizeable portions of the text will be accessible and understandable to specialists in diverse areas of nanoscale science and technology: physicists, engineers, chemists, mathematicians, numerical analysts. Many sections in the book should be suitable for graduate students doing interdisciplinary research and to undergraduate students interested in simulation and in nanoscience.

Some of the material, however, is more advanced. For example, in addition to traditional techniques such as the Finite Element Method, the new finite-difference calculus of Flexible Local Approximation MEthods (FLAME) is presented and its applications in colloidal simulation and photonics are discussed. In many cases, the accuracy of FLAME is much higher than that of the standard finite-difference schemes and even of the finite element method. Another more advanced topic is finite element error estimates, with unconventional eigenvalue and singular value accuracy conditions presented in Section 3.14. Yet another example is the York–Yang "fast Fourier–Poisson"

method, including its versions *without* Fourier transforms; this method is known less well than its more conventional Ewald counterparts.

There is, unfortunately, less than "plenty of room" in this book, as nanofonts were not yet available at the time of its publication. Because of that, and due to the natural constraints of time and the limits of my own expertise, only a limited number of topics in nanoscale simulation could be included. For example, Boundary Integral methods, Fast Multipole algorithms and Finite Difference Time Domain schemes appear primarily in the review sections and as alternatives to other techniques that are discussed in much greater detail (Ewald summation, standard and generalized FD and FEM). On the application side, subjects that are not discussed in the book (nanotubes, nanodots, nanocomposites, and so on) would form a much longer list than the ones that *are* included (selected topics in molecular dynamics, colloidal systems and photonics). My goal will be achieved if some of this book's ideas in the intersection of numerical methods and nanoscale applications help to stimulate further analytical, computational and experimental work.

In closing, let me single out one particularly fascinating subject of current and future research: optical metamaterials. Such materials, with nanoscale feature sizes and very unusual optical properties, have intriguing applications from guiding light to nanofocusing to "cloaking". Progress in this field is impossible without support by physical experiments *and* simulation; especially important are analysis and control of effective electric and magnetic parameters at optical frequencies. Not only for technology, but also for theory and simulation there is indeed "plenty of room at the bottom".

References

[ABCM02] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini. Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Analysis*, 39(5):1749–1779, 2002.

[ABPS02] P. Alotto, A. Bertoni, I. Perugia, and D. Schötzau. Efficient use of the local discontinuous Galerkin method for meshes sliding on a circular boundary. *IEEE Trans. on Magn.*, 38:405–408, 2002.

[AD03] Mara G. Armentano and Ricardo G. Durán. Unified analysis of discontinuous Galerkin methods for elliptic problems. *Num. Meth. for Partial Diff. Equations*, 19(5):653–664, 2003.

[AF98] R. Albanese and R. Fresa. Upper and lower bounds for local electromagnetic quantities. *Int. J. for Numer. Meth. in Eng.*, 42(3):499–515, 1998.

[AF03] Robert A. Adams and John J.F. Fournier. *Sobolev Spaces.* Amsterdam; Boston : Academic Press, 2003.

[AFR00] R. Albanese, R. Fresa, and G. Rubinacci. Local error bounds for static and stationary fields. *IEEE Trans. Magn.*, 36(4):1615–1618, 2000.

[AG99] C. Ashcraft and R.G. Grimes. SPOOLES: an object-oriented sparse matrix library. *In Proc. 1999 SIAM Conf. Parallel Processing for Scientific Computing*, 1999.

[AHCN05] N. Anderson, A. Hartschuh, S. Cronin, and L. Novotny. Nanoscale vibrational analysis of single-walled carbon nanotubes. *J. Am. Chem. Soc.*, 127:2533–2537, 2005.

[AHLT05] Peter Arbenz, Ulrich L. Hetmaniuk, Richard B. Lehoucq, and Raymond S. Tuminaro. A comparison of eigensolvers for large-scale 3d modal analysis using amg-preconditioned iterative methods. *Int. J. for Numer. Meth. in Eng.*, 64:204–236, 2005.

[AK95] A. Ahagon and T. Kashimoto. 3-dimensional electromagnetic-wave analysis using high-order edge elements. *IEEE Trans. Magn.*, 31(3):1753–1756, 1995.

[AK99] W. Axmann and P. Kuchment. An efficient finite element method for computing spectra of photonic and acoustic band-gap materials - I. Scalar case. *J. of Comp. Phys.*, 150(2):468–481, 1999.

[AL98] C. Ashcraft and J. W.H. Liu. Robust ordering of sparse matrices using multisection. *SIAM J. on Matrix Analysis & Appl.*, 19(3):816–832, 1998.

[AM76] Neil W. Ashcroft and N. David Mermin. *Solid State Physics*. Fort Worth: Saunders College Publishing, 1976.

[And87] V.V. Andrievskii. On approximation of functions by harmonic polynomials. *Mathematics of the USSR-Izvestiya*, 51(1):1–13, 1987.

[AO00] Mark Ainsworth and J. Tinsley Oden. *A Posteriori Error Estimation in Finite Element Analysis*. John Wiley & Sons, 2000.

[AP98] U.M. Ascher and Linda Ruth Petzold. *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. Society for Industrial & Applied Mathematics, 1998.

[aPABS03] Jo ao Pedro A. Bastos and Nelson Sadowski. *Electromagnetic Modeling by Finite Element Methods*. New York : Marcel Dekker, 2003.

[App85] A. W. Appel. An efficient program for many-body simulation. *SIAM J. Sci. Stat. Comput.*, 6:85–103, 1985.

[Arf85] G. Arfken. *Mathematical Methods for Physicists*. Orlando, FL: Academic Press, 1985.

[Arn89] Vladimir Igorevich Arnol'd. *Mathematical Methods of Classical Mechanics*. New York : Springer-Verlag, 1989. 2nd ed.

[Arn04] Axel Arnold. *Computer simulations of charged systems in partially periodic systems*. PhD thesis, Johannes Gutenberg Universität, 2004.

[Art80] A.M. Arthurs. *Complementary Variational Principles*. Oxford : Clarendon Press ; New York : Oxford University Press, 1980.

[AS02] S.N. Atluri and S.P. Shen. The meshless local Petrov-Galerkin (MLPG) method: A simple & less-costly alternative to the finite element and boundary element methods. *CMES – Computer Modeling in Engineering & Sciences*, 3(1):11–51, 2002.

[Axe96] Owe Axelsson. *Iterative Solution Methods*. Cambridge University Press, 1996.

[AZ98] S.N. Atluri and T. Zhu. A new meshless local Petrov-Galerkin (MLPG) approach in computational mechanics. *Computational Mechanics*, 22:117–127, 1998.

[BA72] I. Babuška and A.K. Aziz. Survey lectures on the mathematical foundation of the finite element method. In *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*, pages 5–359. Academic Press, New York, 1972.

[BA76] I. Babuška and A.K. Aziz. On the angle condition in the finite element method. *SIAM J. Numer. Analysis*, 13(2):214–226, 1976.

[Bab58] Ivo Babuška. On the schwarz algorithm in the theory of differential equations of mathematical physics. *Tchecosl. Math. J.*, 8:328–342, 1958.

[Bab71] Ivo Babuška. Error bounds for the finite element method. *Numer. Math.*, 16:322–333, 1971.

[Bak66] Nikolai Sergeevitch Bakhvalov. On the convergence of a relaxation method under natural constraints on an elliptic operator. *Zhurnal vychislitel'noj matemaki i matematicheskoj fiziki*, 6:861–883, 1966.

[Bas99] Achim Basermann. Parallel preconditioned solvers for large sparse hermitian eigenvalue problems. In Jack Dongarra and Vicente Hernandez, editors, *Springer Series: Lecture Notes in Computer Science*, volume 1573, pages 72–85. Springer, 1999.

[Bas00] Achim Basermann. Parallel block ilut preconditioning for sparse eigenproblems and sparse linear systems. *Num. Linear Alg. with Appl.*, 7:635–648, 2000.

[BBO03] Ivo Babuška, Uday Banerjee, and John E. Osborn. Survey of meshless and generalized finite element methods: A unified approach. *Acta Numerica*, 12:1–125, 2003.

[BCO94] I. Babuška, G. Caloz, and J.E. Osborn. Special finite-element methods for a class of 2nd-order elliptic problems with rough coefficients. *SIAM J. Numer. Analysis*, 31(4):945–981, 1994.

[BDD+00] Zhaojun Bai, James Demmel, Jack Dongarra, Axel Ruhe, and Henk van der Vorst, editors. *Templates for the Solution of Algebraic Eigenvalue Problems : a Practical Guide*. Society for Industrial and Applied Mathematics: Philadelphia, PA, 2000.

[Bec00] Thomas L. Beck. Real-space mesh techniques in density-functional theory. *Rev. Mod. Phys.*, 72(4):1041–1080, Oct 2000.

[BEK93] Folkmar Bornemann, Bodo Erdmann, and Ralf Kornhuber. Adaptive multilevel methods in three space dimensions. *Int. J. for Numer. Meth. Eng.*, 36:3187–3203, 1993.

[Ber66] Stefan Bergman. Approximation of harmonic functions of three variables by harmonic polynomials. *Duke Math. J.*, 33(2):379–387, 1966.

[BFea99] D. Boffi, P. Fernandes, and L. Gastaldi et al. Computational models of electromagnetic resonators: Analysis of edge element approximation. *SIAM J. on Numer. Analysis*, 36(4):1264–1290, 1999.

[BFR98] F. Brezzi, L.P. Franca, and A. Russo. Further considerations on residual free bubbles for advective-diffusive equations. *Computer Meth. in Appl. Mech. & Eng.*, 166:25–33, 1998.

[BG] Rick Beatson and Leslie Greengard. A short course on fast multipole methods. http://www.math.nyu.edu/faculty/greengar/.

[BGTR04] M. Bathe, A. J. Grodzinsky, B. Tidor, and G. C. Rutledge. Optimal linearized poisson–boltzmann theory applied to the simulation of flexible polyelectrolytes in solution. *J. Chem. Phys.*, 121(16):7557–7561, 2004.

[BH70] J.H. Bramble and S.R. Hilbert. Estimation of linear functionals on sobolev spaces with applications to fourier transforms and spline interpolations. *SIAM J. Numer. Anal.*, 7:113–124, 1970.

[BH86] J. Barnes and P. Hut. A hierarchical $o(nlogn)$ force-calculation algorithm. *Nature*, 324:446–449, 1986.

[BHM00] William L. Briggs, Van Emden Henson, and Steve F. McCormick. *A Multigrid Tutorial*. Philadelphia, PA : Society for Industrial and Applied Mathematics, 2000.

[BKO+96] T. Belytschko, Y. Krongauz, D. Organ, M. Fleming, and P. Krysl. Meshless methods: an overview and recent developments. *Computer Meth. in Appl. Mech. & Eng.*, 139(1–4):3–47, 1996.

[BL58] G.M. Bell and S. Levine. Statistical thermodynamics of concentrated colloidal solutions ii. *Transactions of the Faraday Society*, 54:785–798, 1958.

[BLG94] T. Belytschko, Y. Y. Lu, and L. Gu. Element-free Galerkin methods. *Int. J. for Numer. Meth. Eng.*, 37:229–256, 1994.

[BM84a] I. Babuška and A.D. Miller. The post-processing approach in the finite element method, I: Calculations of displacements, stresses and other higher derivatives of the displacements. *Int. J. for Numer. Meth. Eng.*, 20:1085–1109, 1984.

[BM84b] I. Babuška and A.D. Miller. The post-processing approach in the finite element method, II: The calculation of stress intensity factors. *Int. J. Numer. Meth. Eng.*, 20:1111–1129, 1984.

[BM84c] I. Babuška and A.D. Miller. The post-processing approach in the finite element method, III: A posteriori error estimation and adaptive mesh selection. *Int. J. Numer. Meth. Eng.*, 20:2311–2324, 1984.

[BM89] A. Bossavit and I. Mayergoyz. Edge-elements for scattering problems. *IEEE Trans. Magn.*, 25(4):2816–2821, 1989.

[BM97] I. Babuška and J.M. Melenk. The partition of unity method. *Int. J. for Numer. Meth. in Eng.*, 40(4):727–758, 1997.

[BMM01] M. Bordag, U. Mohideen, and V.M. Mostepanenko. New developments in the casimir effect. *Physics Reports*, 353:1–205, 2001.

[BMS02] C.L. Bottasso, S. Micheletti, and R. Sacco. The discontinuous Petrov-Galerkin method for elliptic problems. *Computer Meth. in Appl. Mech. & Eng.*, 191(31):3391–3409, 2002.

[BN97] I. Babuška and R. Narasimhan. The babuška-brezzi condition and the patch test: an example. *Computer Meth. in Appl. Mech. & Eng.*, 140(1-2):183–199, 1997.

[Bof01] D. Boffi. A note on the de Rham complex and a discrete compactness property. *Appl. Math. Lett.*, 14(1):33–38, 2001.

[Boo01] Carl De Boor. *A Practical Guide to Splines.* New York: Springer, 2001.

[Bos88a] A. Bossavit. A rationale for "edge elements" in 3-d fields computations. *IEEE Trans. Magn.*, 24(1):74–79, 1988.

[Bos88b] A. Bossavit. Whitney forms: A class of finite elements for three-dimensional computations in electromagnetism. *IEE Proc. A*, 135:493–500, 1988.

[Bos90] A. Bossavit. Solving maxwell equations in a closed cavity, and the question of "spurious modes". *IEEE Trans. Magn.*, 26(2):702–705, 1990.

[Bos92] A. Bossavit. Edge-element computation of the force-field in deformable-bodies. *IEEE Trans. Magn.*, 28(2):1263–1266, 1992.

[Bos98] Alain Bossavit. *Computational Electromagnetism: Variational Formulations, Complementarity, Edge Elements.* San Diego: Academic Press, 1998.

[Boy01] John P. Boyd. *Chebyshev and Fourier Spectral Methods.* Publisher: Dover Publications, 2001.

[BP53] L. Brillouin and M. Parodi. *Wave Propagation in Periodic Structures.* Dover, New York, 1953.

[BPX90] J. H. Bramble, J. E. Pasciak, and J. Xu. Parallel multilevel preconditioners. *Math. Comp.*, 55:1–22, 1990.

[BR78a] I. Babuška and W.C. Rheinboldt. A-posteriori error estimates for the finite element method. *Int. J. for Numer. Meth. in Eng.*, 12(10):1597–1615, 1978.

[BR78b] I. Babuška and W.C. Rheinboldt. Error estimates for adaptive finite element computations. *SIAM J. on Numer. Analysis*, 15(4):736–754, Aug 1978.

[BR79] I. Babuška and W.C. Rheinboldt. On the reliability and optimality of the finite element method. *Computers & Structures*, 10:87–94, 1979.

[BR99] G. Binning and H. Rohrer. In touch with atoms. *Reviews of Modern Physics*, 71:S324–S330, 1999.

[BR01] Roland Becker and Rolf Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numerica*, 10:1–102, 2001.

[BR03] Wolfgang Bangerth and Rolf Rannacher. *Adaptive Finite Element Methods for Differential Equations*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 2003.

[Bra77] Achi Brandt. Multi-level adaptive solutions to boundary-value problems. *Mathematics of Computation*, 31(138):333–390, apr 1977.

[Bra93] James H. Bramble. *Multigrid Methods*. Harlow, Essex, England : Longman Scientific & Technical ; New York : Copubished in the U.S. with J. Wiley & Sons, 1993.

[Bre74] F. Brezzi. On the existence, uniqueness and approximation of saddle-point problems arising from lagrange multipliers. *R.A.I.R.O.*, 8:129–151, 1974.

[BRGW82] G. Binning, H. Rohrer, Ch. Gerber, and E. Weibel. Surface studies by scanning tunneling microscopy. *Phys. Rev. Lett.*, 49(1):57–61, Jul 1982.

[Bri60] L'eon Brillouin. *Wave Propagation and Group Velocity*. Academic Press, 1960.

[Bri92] J.L. Britton, editor. *Collected Works of A.M. Turing. Pure Mathematics. With a section on Turing's statistical work by I. J. Good*. Amsterdam, etc.: North-Holland, 1992.

[BS79] R.E. Bank and A.H. Sherman. The use of adaptive grid refinement for badly behaved elliptic partial differential equations. In R. Vichnevetsky and R. S. Stepleman, editors, *Advances in Computer Methods for Partial Differential Equations III*, pages 33–39. IMACS, New Brunswick, 1979.

[BS94] Ivo Babuška and Manil Suri. The p and h-p versions of the Finite Element Method, basic principles and properties. *SIAM Review*, 36:578–632, 1994.

[BS01] Ivo Babuška and Theofanis Strouboulis. *The Finite Element Method and Its Reliability*. Oxford, [England] : Clarendon Press ; New York : Oxford University Press, 2001.

[BS02] S.C. Brenner and L.R. Scott. *The Mathematical Theory of Finite Element Methods*. New York: Springer, 2002.

[BSB96] E.L. Briggs, D.J. Sullivan, and J. Bernholc. Real-space multigrid-based approach to large-scale electronic structure calculations. *Physical Review B*, 54(20):14362–14375, 1996.

[BSS97] Thomas C. Bishop, Robert D. Skeel, and Klaus Schulten. Difficulties with multiple time stepping and fast multipole algorithm in molecular dynamics. *J. Comp. Chem.*, 18:1785–1791, 1997.

[BSS+01] N.A. Baker, D. Sept, J. Simpson, M.J. Holst, and J.A. McCammon. Electrostatics of nanosystems: Application to microtubules and the ribosome. *PNAS*, 98(18):10037–10041, 2001.

[BST04] P.A. Belov, C.R. Simovski, and S.A. Tretyakov. Backward waves and negative refraction in photonic (electromagnetic) crystals. *J. of Communications Technology and Electronics*, 49(11):1199–1207, 2004.

[BSve] A.I. Bobenko and Yu.B. Suris. *Discrete Differential Geometry. Consistency as Integrability*. Preliminary version at arXiv: math.DG/0504358, 2007 (tentative).

[BSU+94] Ivo Babuška, T. Strouboulis, C. S. Upadhyay, S. K. Gangaraj, and K. Copps. Validation of a posteriori error estimators by numerical approach. *Int. J. for Numer. Methods in Eng.*, 37:1073–1123, 1994.

[BT05] Achim Basermann and Igor Tsukerman. Parallel generalized finite element method for magnetic multiparticle problems. In M. Daydé *et al.*, editor, *Springer Series: Lecture Notes in Computational Science and Engineering*, volume LNCS 3402, pages 325–339. Springer, 2005.

[But87] John C. Butcher. *The Numerical Analysis of Ordinary Differential Equations: Runge–Kutta and General Linear Methods.* John Wiley & Sons, 1987.

[But03] John C. Butcher. *Numerical Methods for Ordinary Differential Equations.* Hoboken, NJ: J. Wiley, 2003.

[BV82] A. Bossavit and J.C. Vérité. A mixed FEM-BIEM method to solve 3-D eddy current problem. *IEEE Trans. Magn.*, 18(2):431–435, 1982.

[BV83] A. Bossavit and J.C. Vérité. The Trifou code: solving the 3d eddy currents problems by using h as state variable. *IEEE Trans. Magn.*, 19:2465–2470, 1983.

[BY75] P. Barber and C. Yeh. Scattering of electromagnetic waves by arbitrarily shaped dielectric bodies. *Applied Optics*, 14(12):2864–2872, 1975.

[Byk72] V.P. Bykov. Spontaneous emission in a periodic structure. *Soviet physics, JETP (Journal of Experimental and Theoretical Physics)*, 35(2):269–273, 1972.

[Byk75] V.P. Bykov. Spontaneous emission from a medium with a band spectrum. *Sov. J. Quant. Electron.*, 4(7):861–871, 1975.

[Byk93] V.P. Bykov. *Radiation of Atoms in a Resonant Environment.* World Scientific, Singapore, 1993.

[CAO+03] E. Cubukcu, K. Aydin, E. Ozbay, S. Foteinopolou, and C. M. Soukoulis. Subwavelength resolution in a two-dimensional photonic-crystal-based superlens. *Phys. Rev. Lett.*, 91(20):207401, Nov 2003.

[CBPS00] P. Castillo, B.Cockburn, I. Perugia, and D. Schöotzau. An a priori error analysis of the local discontinuous Galerkin method for elliptic problems. *SIAM J. on Numer. Analysis*, 38(5):1676–1706, 2000.

[CCD+05] D.A. Case, T.E. Cheatham, T. Darden, H. Gohlke, R. Luo, K.M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R.J. Woods. The amber biomolecular simulation programs. *J. of Comput. Chem.*, 26(16):1668–1688, 2005.

[CCY+07] W. Cai, U.K. Chettiar, H.-K. Yuan, V.C. de Silva, A.V. Kildishev, V.P. Drachev, and V.M. Shalaev. Metamagnetics with rainbow colors. *Opt. Express*, 15:3333–3341, 2007.

[CD92] R.D. Coalson and A. Duncan. Systematic ionic screening theory of macroions. *J. of Chem. Phys.*, 97(8):5653–5661, 1992.

[CF97] C.M. Cortis and R.A. Friesner. Numerical solution of the poisson-boltzmann equation using tetrahedral finite-element meshes. *J. of Comput. Chem.*, 18(13):1591–1608, 1997.

[CFR00] S. Caorsi, P. Fernandes, and M. Raffetto. On the convergence of galerkin finite element approximations of electromagnetic eigenproblems. *SIAM J. on Numer. Analysis*, 38(2):580–607, 2000.

[CGR99] H. Cheng, L. Greengard, and V. Rokhlin. A fast adaptive multipole algorithm in three dimensions. *J. of Comp. Phys.*, 155(2):468–498, 1999.

[Cha13] D.L. Chapman. A contribution to the theory of electrocapillarity. *Philosophical Magazine*, 25(6):475–481, 1913.

[Chu02] T. J. Chung. *Computational Fluid Dynamics*. Cambridge University Press, 2002.

[Cia80] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems*. Amsterdam; New York: North-Holland Pub. Co., 1980.

[CJMS01] W.C. Chew, J.M. Jin, E. Michielssen, and J.M. Song, editors. *Fast and Efficient Algorithms in Computational Electromagnetics*. Artech House: Boston, MA, 2001.

[CK97] Q. Chen and A. Konrad. A review of finite element open boundary techniques for static and quasistatic electromagnetic field problems. *IEEE Trans. Magn.*, 33(1):663–676, 1997.

[CKS00] B. Cockburn, G.E. Karniadakis, and C.-W. Shu. The development of discontinuous Galerkin methods. In B. Cockburn, G.E.Karniadakis, and C.-W.Shu, editors, *Discontinuous Galerkin Methods. Theory, Computation and Applications*, volume 11 of *Lecture Notes in Comput. Sci. Engrg.*, pages 3–50. Springer-Verlag, New York, 2000.

[CKSU05] H. Cohn, R. Kleinberg, B. Szegedy, and C. Umans. Group-theoretic algorithms for matrix multiplication. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 379–388, 2005.

[CM96] L. Cordes and B. Moran. Treatment of material discontinuity in the element-free Galerkin method. *Comput. Meth. Appl. Mech. Engng.*, 139:75–89, 1996.

[CM98] Gary Cohen and Peter Monk. Gauss point mass lumping schemes for Maxwell's equations. *Num. Meth. for Partial Diff. Equations*, 14(1):63–88, 1998.

[Col66] Lothar Collatz. *The Numerical Treatment of Differential Equations*. New York: Springer, 1966.

[Col97] James B. Cole. High accuracy solution of Maxwell's equations using nonstandard finite differences. *Computers in Physics*, 11(3):287–292, 1997.

[Col04] James B. Cole. High-accuracy FDTD solution of the absorbing wave equation, and conducting Maxwell's equations based on a nonstandard finite-difference model. *IEEE Trans. on Antennas & Propagation*, 52(3):725–729, 2004.

[CR72] P. G. Ciarlet and P.-A. Raviart. General Lagrange and Hermite interpolation in $\mathbb{R}^n$ with applications to finite element methods. *Arch. Rational Mech. Anal.*, 46:177–199, 1972.

[CR73] M. Crouzeix and P.A. Raviart. Conforming and nonconforming finite element methods for solving the stationary Stokes equation. *RAIRO Anal. Numer.*, 7, R-3:33–76, 1973. MR 49:8401.

[CW71] P.G. Ciarlet and C. Wagschal. Multipoint taylor formulas and applications to the finite element method. *Numer. Math.*, 17:84–100, 1971.

[CW02] M. Clemens and T. Weiland. Magnetic field simulation using conformal FIT formulations. *IEEE Trans. Magn.*, 38(2):389–392, 2002.

[CXS93] Rongqing Chen, Zhizhan Xu, and Lan Sun. Finite-difference scheme to solve Schrödinger equations. *Phys. Rev. E*, 47(5):3799–3802, 1993.

[Dav75] E.R. Davidson. The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real symmetric matrices. *J. Comput. Phys.*, 17:87–94, 1975.

[Dav06] Timothy A. Davis. *Direct Methods for Sparse Linear Systems.* Philadelphia : SIAM, Society for Industrial and Applied Mathematics, 2006.

[DB01] Suvranua De and Klaus-Jrgena Bathe. Towards an efficient meshless computational technique: the method of finite spheres. *Engineering Computations*, 18(1–2):170–192, 2001.

[DBO00] C.A. Duarte, I. Babuška, and J.T. Oden. Generalized finite element methods for three-dimensional structural mechanics problems. *Computers & Structures*, 77(2):215–232, 2000.

[DD74] J. Douglas and T. Dupont. Galerkin approximations for the two point boundary problem using continuous, piecewise polynomial spaces. *Numer. Math.*, 22:99–109, 1974.

[DECB98] Costas D. Dimitropoulos, Brian J. Edwards, Kyung-Sun Chae, and Antony N. Beris. Efficient pseudospectral flow simulations in moderately complex geometries. *J. of Comp. Phys.*, 144:517–549, 1998.

[Dem97] James W. Demmel. *Applied Numerical Linear Algebra.* Society for Industrial & Applied Mathematics, 1997.

[Dem06] Leszek Demkowicz. *Computing with hp-Adaptive Finite Elements: Volume 1, One- and Two-Dimensional Elliptic and Maxwell Problems.* Chapman & Hall/Crc, 2006.

[DER89] I.S. Duff, A.M. Erisman, and J.K. Reid. *Direct Methods for Sparse Matrices.* Oxford University Press, 1989.

[DEW99] A. Doicu, Yu.A. Eremin, and T. Wriedt. Convergence of the t-matrix method for light scattering from a particle on or near a surface. *Optics Communications*, 159:266–277, 1999.

[DEW+06] Gunnar Dolling, Christian Enkrich, Martin Wegener, Costas M. Soukoulis, and Stefan Linden. Low-loss negative-index metamaterial at telecommunication wavelengths. *Optics Letters*, 31(12):1800–1802, 2006.

[DF] B. Draine and Piotr J. Flatau. User guide to the discrete dipole approximation code ddscat.6.0.

[DF94] B. Draine and P. Flatau. Discrete-dipole approximation for scattering calculations. *J. Opt. Soc. Am. A*, 11:1491–1499, 1994.

[DG05] M. Dorica and D.D. Giannacopoulos. Impact of mesh quality improvement systems on the accuracy of adaptive finite-element electromagnetics with tetrahedra. *IEEE Trans. Magn.*, 41(5):1692–1695, 2005.

[DH98a] Markus Deserno and Christian Holm. How to mesh up Ewald sums. I. a theoretical and numerical comparison of various particle mesh routines. *J. Chem. Phys.*, 109:7678–7693, 1998.

[DH98b] Markus Deserno and Christian Holm. How to mesh up Ewald sums. II. an accurate error estimate for the P3M algorithm. *J. Chem. Phys.*, 109:7694–7701, 1998.

[DH01] Markus Deserno and Christian Holm. Cell model and poisson-boltzmann theory: A brief introduction. In C. Holm, P. K'ekicheff, and R. Podgornik, editors, *NATO Science Series II – Mathematics, Physics and Chemistry*, volume 46. Kluwer, Dordrecht, 2001.

[DHM00] Markus Deserno, Christian Holm, and Sylvio May. Fraction of condensed counterions around a charged rod: comparison of Poisson-Boltzmann theory and computer simulations. *Macromolecules*, 33:199–205, 2000.

[DHM+04] J. Dobnikar, D. Haložan, M.Brumen, H.-H. von Grünberg, and R. Rzehak. Poisson–Boltzmann Brownian dynamics of charged colloids in suspension. *Computer Physics Communications*, 159:73–92, 2004.

[DK00] Z.H. Duan and R. Krasny. An Ewald summation based multipole method. *J. Chem. Phys.*, 113(9):3492–3495, 2000.

[DK01] Z.H. Duan and R. Krasny. An adaptive treecode for computing nonbonded potential energy in classical molecular systems. *J. Comp. Chem.*, 22(2):184–195, 2001.

[DL41] B.V. Derjaguin and L. Landau. Theory of the stability of strongly charged hydrophobic sols and of the adhesion of strongly charged particles in solutions of electrolytes. *Acta Physicochimica (USSR)*, 14:633–662, 1941.

[DL04] R.A. Depine and A. Lakhtakia. A new condition to identify isotropic dielectric-magnetic materials displaying negative phase velocity. *Microwave Opt. Technol. Lett.*, 41:315–316, 2004.

[dLPS80a] S. W. de Leeuw, J. W. Perram, and E. R. Smith. Simulation of electrostatic systems in periodic boundary conditions. I. lattice sums and dielectric constants. *Proc. Royal Soc. London A*, 373:27–56, 1980.

[dLPS80b] S. W. de Leeuw, J. W. Perram, and E. R. Smith. Simulation of electrostatic systems in periodic boundary conditions. II. Equivalence of boundary conditions. *Proc. Royal Soc. London A*, 373:57–66, 1980.

[dLPS86] S. W. de Leeuw, J. W. Perram, and E. R. Smith. Computer simulation of the static dielectric constant of systems with permanent electric dipoles. *Ann. Rev. Phys. Chem.*, 37:245–270, 1986.

[DM65] M. Danos and L.C. Maximon. Multipole matrix elements of the translation operator. *J. Math. Phys.*, 6:766–778, 1965.

[DM99] S. Dey and R. Mittra. A conformal finite-difference time-domain technique for modeling cylindrical dielectric resonators. *IEEE Trans. MTT*, 47(9):1737–1739, 1999.

[DO96] C.A. Duarte and J.T. Oden. h-p adaptive method using clouds. *Computer Meth. in Appl. Mech. & Eng.*, 139:237–262, 1996.

[Dod76] J. Dodziuk. Finite-difference approach to the Hodge theory of harmonic forms. *Amer. J. Math.*, 98(1):79–104, 1976.

[Dou96] Craig C. Douglas. Multigrid methods in science and engineering. *IEEE Comput. Sci. Eng.*, 3(4):55–68, 1996.

[DP01] D.C. Dobson and J.E. Pasciak. Analysis of an algorithm for computing electromagnetic Bloch modes using Nedelec spaces. *Comput. Meth. in Appl. Math.*, 1(2):138–153, 2001.

[DT06] Jianhua Dai and Igor Tsukerman. Flexible difference schemes with numerical bases for electrostatic particle interactions. In *Proceedings of Twelfth Biennial IEEE Conference on Electromagnetic Field Computation (CEFC 2006)*, 2006.

[DT07] Jianhua Dai and Igor Tsukerman. Flexible approximation schemes with adaptive grid refinement. In *Proceedings of Compumag2007, Aachen, Germany. Submitted to* IEEE Trans. Magn., 2007.

[DTP97] T. A. Darden, A. Toukmaji, and L. G. Pedersen. Long-range electrostatic effects in biomolecular simulations. *Journal de Chimie Physique et de Physico-Chimie Biologique*, 94:1346–1364, 1997.

[DTRS07] J. Dai, I. Tsukerman, A. Rubinstein, and S. Sherman. New computational models for electrostatics of macromolecules in solvents. *IEEE Trans. Magn.*, 43(4):1217–1220, 2007.

[Dud94] Donald G. Dudley. *Mathematical Foundations for Electromagnetic Theory.* Wiley-IEEE Press, 1994.

[DvG02] Markus Deserno and Hans-Hennig von Grünberg. Osmotic pressure of charged colloidal suspensions: A unified approach to linearized poisson-boltzmann theory. *Phys. Rev. E*, 66(1):011401, 2002.

[DvTS] Jianhua Dai, Frantisek Čajko, Igor Tsukerman, and Mark I. Stockman. Electrodynamic effects in plasmonic nanolenses. Submitted.

[DWSL07] G. Dolling, M. Wegener, C.M. Soukoulis, and S. Linden. Negative-index metamaterial at 780 nm wavelength. *Optics Letters*, 32(1):53–55, 2007.

[DYP93] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh Ewald: An N log(N) method for Ewald sums in large systems. *J. of Chem. Phys.*, 98(12):10089–10092, 1993.

[Eas73] M.S.P. Eastham. *The Spectral Theory of Periodic Differential Equations.* Scottish Academic Press: Edinburgh and London, 1973.

[EB05] G.V. Eleftheriades and K.G. Balmain. *Negative Refraction Metamaterials: Fundamental Principles and Applications.* Wiley-IEEE Press, 2005.

[EG84] Jr. E.C. Gartland. Computable pointwise error bounds and the Ritz method in one dimension. *SIAM Journal on Numerical Analysis*, 21(1):84–100, 1984.

[EJ97a] Alexandre Elmkies and Patrick Joly. Finite elements and mass lumping for Maxwell's equations: the 2D case. *Comptes Rendus de l'Academie des Sciences Series I Mathematics*, 324(11):1287–1293, 1997.

[EJ97b] Alexandre Elmkies and Patrick Joly. Finite elements and mass lumping for Maxwell's equations: the 3D case. *Comptes Rendus de l'Academie des Sciences Series I Mathematics*, 325(11):1217–1222, 1997.

[Ell93] R.S. Elliott. *An Introduction to Guided Waves and Microwave Circuits.* Prentice-Hall, 1993.

[EPB+95] U. Essmann, L. Perera, M.L. Berkowitz, T. Darden, H. Lee, and L.G. Pedersen. A smooth particle mesh Ewald method. *J. Chem. Phys.*, 103(19):8577–8593, 1995.

[ES83] C.R.I. Emson and J. Simkin. An optimal method for 3-d eddy currents. *IEEE Trans. Magn.*, 19(6):2450–2452, 1983.

[EVK06] R. Esteban, R. Vogelgesang, and K. Kern. Simulation of optical near and far fields of dielectric apertureless scanning probes. *Nanotechnology*, 17(2):475–482, 2006.

[Ewa21] P. Ewald. Die berechnung optischer und elektrostatischer gitterpotentiale. [evaluation of optical and electrostatic lattice potentials]. *Ann. Phys. Leipzig*, 64:253–287, 1921.

[Fab98] I.L. Fabelinskii. Seventy years of combination (Raman) scattering. *Physics Uspekhi*, 41(12):1229–1247, 1998.

[FAL99] Nordin Félidj, Jean Aubard, and Georges Lévi. Discrete dipole approximation for ultraviolet-visible extinction spectra simulation of silver and gold colloids. *J. of Chem. Phys.*, 111(3):1195–1208, 1999.

[FB97] F. Fogolari and J.M. Briggs. On the variational approach to Poisson-Boltzmann free energies. *Chemical Physics Letters*, 281:135–139, 1997.

[Fed61] Radii Petrovich Fedorenko. A relaxation method for solving elliptic difference equations. *Zhurnal Vychislitel'noj Matemaki i Matematicheskoj Fiziki*, 1:922–927, 1961. English translation: USSR Computational Math. and Math. Physics, vol. 1, 1962, pp. 1092–1096.

[Fed64] Radii Petrovich Fedorenko. The speed of convergence of one iteration process. *Zhurnal Vychislitel'noj Matemaki i Matematicheskoj Fiziki*, 4:559–563, 1964. English translation: USSR Computational Math. and Math. Physics, vol. 4, 1964, pp. 227–235.

[Fei02] Evgenii L. Feinberg. The forefather (about Leonid Isaakovich Mandelstam). *Physics-Uspekhi*, 45:81–100, 2002.

[FEVM01] F. Fogolari, G. Esposito, P. Viglino, and H. Molinari. Molecular mechanics and dynamics of biomolecules using a solvent continuum model. *J. of Comput. Chem.*, 22(15):1830–1842, 2001.

[Fey59] Richard Phillips Feynman. There's plenty of room at the bottom. In APS meeting*, December 29, 1959*, 1959.

[FF63] D.K. Faddeev and V.N. Faddeeva. *Computational Methods of Linear Algebra.* W. H. Freeman: San Francisco, 1963.

[FGW73] G.J. Fix, S. Gulati, and G.I. Wakoff. On the use of singular functions with finite elements approximations. *J. Comput. Phys.*, 13:209–228, 1973.

[FHF01] C. Farhat, I. Harari, and L.P. Franca. The discontinuous enrichment method. *Computer Meth. in Appl. Mech. & Eng.*, 190:6455–6479, 2001.

[FK04] Takeshi Fujisawa and Masanori Koshiba. Time-domain beam propagation method for nonlinear optical propagation analysis and its application to photonic crystal circuits. *J. of Lightwave Technology*, 22(2):684–691, 2004.

[Fla97] Piotr J. Flatau. Improvements in the discrete-dipole approximation method of computing scattering and absorption. *Optics Letters*, 22:1205–1207, 1997.

[FLS89] Richard Phillips Feynman, Robert B. Leighton, and Matthew L. Sands. *The Feynman Lectures on Physics.* Redwood City, Calif.: Addison-Wesley, 1989. vol. 1: Mechanics, radiation, and heat; vol. 2: Electromagnetism and matter; vol. 3: Quantum mechanics.

[FLSZ05] Nicholas Fang, Hyesog Lee, Cheng Sun, and Xiang Zhang. Sub-diffraction-limited optical imaging with a silver superlens. *Science*, 308(5721):534–537, 2005.

[FLZB97] F. Figueirido, R.M. Levy, R.H. Zhou, and B.J. Berne. Large scale simulation of macromolecules in solution: Combining the periodic fast multipole method with multiple time step integrators. *J. of Chem. Phys.*, 106(23):9835–9849, 1997.

[FM67] George E. Forsythe and Cleve B. Moler. *Computer Solution of Linear Algebraic Systems.* Englewood Cliffs, N.J. : Prentice-Hall, 1967.

[FM03] D.R. Fredkin and I.D. Mayergoyz. Resonant behavior of dielectric objects (electrostatic resonances). *Phys. Rev. Lett.*, 91(25):253902, 2003.

[Fra69] U. Kreibig & C. Von Fragstein. Limitation of electron mean free path in small silver particles. *Zeitschrift fuer Physik*, 224(4):307–23, 1969.

[Fri05] Gary Friedman. Private communication, 2002–2005.

[Fus92] M. Fushiki. Molecular dynamics simulations for charged colloidal dispersions. *J. Chem. Phys.*, 97(2):6700–6713, 1992.

[Gan59] Felix R. Gantmakher. *The Theory of Matrices.* New York, Chelsea Pub. Co., 1959.

[Gan88] F.R. Gantmakher. *Teoriia Matrits.* Nauka, 1988.

[Gbu03] G. Gbur. Nonradiating sources and other "invisible" objects. *Progress in Optics*, 45:273–315, 2003.

[GD03] Nail A. Gumerov and Ramani Duraiswami. Recursions for the computation of multipole translation and rotation coefficients for the 3-D Helmholtz equation. *SIAM J. Sci. Comput.*, 25(4):1344–1381, 2003.

[GDLM93] M.K. Gilson, M.E. Davis, B.A. Luty, and J.A. McCammon. Computation of electrostatic forces on solvated molecules using the Poisson-Boltzmann equation. *J. Phys. Chem.*, 97:3591–3600, 1993.

[Gea67] C. William Gear. The numerical integration of ordinary differential equations. *Math. Comp.*, 21(98):146–156, 1967.

[Gea71] C. William Gear. *Numerical Initial Value Problems in Ordinary Differential Equations.* Englewood Cliffs, N.J., Prentice-Hall, 1971.

[GH02] Leslie F. Greengard and Jingfang Huang. A new version of the Fast Multipole Method for screened Coulomb interactions in three dimensions. *J. Comp. Phys.*, 180:642–658, 2002.

[GL81] Alan George and Joseph W-H Liu. *Computer Solution of Large Sparse Positive Definite Systems.* Englewood Cliffs, N.J., Prentice-Hall, 1981.

[GL89] Alan George and Joseph W.H. Liu. The evolution of the minimum degree ordering algorithm. *SIAM Review*, 31(1):1–19, 1989.

[GL96] Gene H. Golub and Charles F. Van Loan. *Matrix Computations.* The Johns Hopkins University Press: Baltimore, MD, 1996.

[GLe] Cleve Ashcraftand Roger Grimes, Joseph Liu, and Jim Patterson *et al.* Spooles 2.2 : SParse object oriented linear equations solver.

[GLOP96] M. Gyimesi, D. Lavers, D. Ostergaard, and T. Pawlak. Hybrid finite element – Trefftz method for open boundary anslysis. *IEEE Trans. Magn.*, 32(3):671–674, 1996.

[GMKH05] R. Gajic, R. Meisels, F. Kuchar, and K. Hingerl. Refraction and rightness in photonic crystals. *Opt. Express*, 13:8596–8605, 2005.

[GNS02] A. Yu. Grosberg, T. T. Nguyen, and B. I. Shklovskii. Colloquium: The physics of charge inversion in chemical and biological systems. *Reviews of Modern Physics*, 74(2):329–345, 2002.

[GNV02] N. Garcia and M. Nieto-Vesperinas. Left-handed materials do not make a perfect lens. *Phys. Rev. Lett.*, 88(20):207403, May 2002.

[Gou10] G. Gouy. Sur la constitution de la charge électrique à la surface d'un électrolyte. *Journal de physique théorique et appliqué*, 9:457–468, 1910.

[GPN01] J. A. Grant, B. T. Pickup, and A. Nicholls. A smooth permittivity function for Poisson–Boltzmann solvation methods. *J. Comp. Chem.*, 22(6):608–640, 2001. and references therein.

[GR87a] S.K. Godunov and V.S. Ryabenkii. *Difference Schemes: an Introduction to the Underlying Theory.* Amsterdam; New York: Elsevier Science Pub. Co., 1987. ISBN 0444702334.

[GR87b] L. Greengard and V. Rokhlin. A fast algorithm for particle simulations. *J. Comput. Phys.*, 73:325–348, 1987.

[GR97] L. Greengard and V. Rokhlin. A new version of the fast multipole method for the laplace equation in three dimensions. *Acta Numerica*, 6:229–269, 1997.

[Gre48] H.J. Greenberg. The determination of upper and lower bounds for the solution of the Dirichlet problem. *J. Math. Phys.*, 27:161–182, 1948.

[Gre87] Leslie Greengard. *The rapid evaluation of potential fields in particle systems.* PhD thesis, The Massachusetts Institute of Technology, 1987. Association for Computing Machinery distingished dissertations.

[GRSP95] Griffin K. Gothard, Sadasiva M. Rao, Tapan K. Sarkar, and Magdalena Salazar Palma. Finite element solution of open region electrostatic problems incorporating the measured equation of invariance. *IEEE Microwave and Guided Wave Letters*, 5(8):252–254, 1995.

[GS00] M. Griebel and M. A. Schweitzer. A particle-partition of unity method for the solution of elliptic, parabolic and hyperbolic PDE. *SIAM J. Sci. Comp.*, 22(3):853–890, 2000.

[GS02a] M. Griebel and M. A. Schweitzer. A particle-partition of unity method-part II: efficient cover construction and reliable integration. *SIAM J. Sci. Comp.*, 23(5):1655–1682, 2002.

[GS02b] M. Griebel and M. A. Schweitzer. A particle-partition of unity method-part III: a multilevel solver. *SIAM J. Sci. Comp.*, 24(2):377–409, 2002.

[GvdV00] Gene H. Golub and Henk A. van der Vorst. Eigenvalue computation in the 20th century. *Journal of Computational and Applied Mathematics*, 123:35–65, 2000.

[GWO01] M. Gyimesi, Jian-She Wang, and D. Ostergaard. Hybrid p-element and Trefftz method for capacitance computation. *IEEE Trans. Magn.*, 37(5):3680–3683, 2001.

[GWP97] R. Graglia, D.R. Wilton, and A.F. Peterson. Higher order interpolatory vector bases for computational electromagnetics. *IEEE Trans. Antennas and Prop.*, 45:329–342, 1997.

[Hac85] Wolfgang Hackbusch. *Multi-grid Methods and Applications.* Berlin ; New York : Springer-Verlag, 1985.

[Had02a] G. Ronald Hadley. High-accuracy finite-difference equations for dielectric waveguide analysis I: uniform regions and dielectric interfaces. *Journal of Lightwave Technology*, 20(7):1210–1218, 2002.

[Had02b] G. Ronald Hadley. High-accuracy finite-difference equations for dielectric waveguide analysis II: dielectric corners. *Journal of Lightwave Technology*, 20(7):1219–1231, 2002.

[Haf99a] Christian Hafner. *MaX-1: A Visual Electromagnetics Platform for PCs.* Wiley, 1999.

[Haf99b] Christian Hafner. *Post-modern Electromagnetics: Using Intelligent MaXwell Solvers.* Wiley, 1999.

[Har01] Roger F. Harrington. *Time-Harmonic Electromagnetic Fields.* Wiley-IEEE Press, 2001.

[HBC03] Andrew A. Houck, Jeffrey B. Brock, and Isaac L. Chuang. Experimental observations of a left-handed material that obeys snell's law. *Phys. Rev. Lett.*, 90(13):137401, Apr 2003.

[HCS90] K.M. Ho, C.T. Chan, and C.M. Soukoulis. Existence of a photonic gap in periodic dielectric structures. *Phys. Rev. Lett.*, 65(25):3152–3155, Dec 1990.

[HE88] R. W. Hockney and J. W. Eastwood. *Computer Simulation Using Particles.* Taylor & Francis, Inc.: Bristol, PA, USA, 1988.

[Her00] Ismael Herrera. Trefftz method: A general theory. *Numer. Methods Partial Differential Eq.*, 16:561–580, 2000.

[HFT04] Derek Halverson, Gary Friedman, and Igor Tsukerman. Local approximation matching for open boundary problems. *IEEE Trans. Magn.*, 40(4):2152–2154, 2004.

[HGR96] L. Tobiska H.-G. Roos, M. Stynes. *Numerical Methods for Singularly Perturbed Differential Equations: Convection-Diffusion and Flow Problems (Springer Series in Computational Mathematics.* Springer, 1996.

[Hig02] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms.* Society for Industrial & Applied Mathematics, 2002.

[Hip01] R. Hiptmair. Discrete Hodge operators. *Numer. Math.*, 90:265–289, 2001.

[HJ90] Roger A. Horn and Charles R. Johnson. *Matrix Analysis.* Cambridge [England]; New York: Cambridge University Press, 1990.

[HJ94] Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis.* Cambridge University Press, 1994.

[HKBJK03] Christopher L. Holloway, Edward F. Kuester, James Baker-Jarvis, and Pavel Kabos. A double negative (dng) composite medium composed of magnetodielectric spherical particles embedded in a matrix. *IEEE Transactions on Antennas and Propagation*, 51(10):2596–2603, 2003.

[HN95] B. Honig and A. Nicholls. Classical electrostatics in biology and chemistry. *Science*, 268(5214):1144–1149, 1995.

[HR98a] John H. Henderson and Sadasiva M. Rao. Electrostatic solution for three-dimensional arbitrarily shaped conducting bodies using finite element and measured equation of invariance. *IEEE Transactions on Antennas and Propagation*, 46(11):1660–1664, 1998.

[HR98b] Tomasz Hrycak and Vladimir Rokhlin. An improved fast multipole algorithm for potential fields. *SIAM Journal on Scientific Computing*, 19(6):1804–1826, 1998.

[HrW93] E. Hairer, S.P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations – Stiff and Differential-Algebraic Problems.* Berlin; New York: Springer-Verlag, 1993.

[HT95] I. Harari and E. Turkel. Fourth order accurate finite difference methods for time-harmonic wave propagation. *J. Comp. Phys.*, 119:252–270, 1995.

[HTK02] R. Hillenbrand, T. Taubner, and F. Keilmann. Phonon-enhanced light-matter interaction at the nanometre scale. *Nature*, 418(6894):159–162, 2002.

[Hug95] T.J.R. Hughes. Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid-scale models, bubbles and the origins of stabilized methods. *Computer Meth. in Appl. Mech. & Eng.*, 127:387–401, 1995.

[HY04] Louis A. Hageman and David M. Young. *Applied Iterative Methods.* Dover Publications, 2004.

[Iag62] A.M. Iaglom. *An Introduction to the Theory of Stationary Random Functions.* New York, Dover Publications, 1973, c1962.

[IB] P. Ingelstrom and A. Bondeson. Goal-oriented error estimation and h-adaptivity for Maxwell's equations. *Computer Meth. in Appl. Mech. & Eng.*, 192(22).

[IHH+04] Taro Ichimura, Norihiko Hayazawa, Mamoru Hashimoto, Yasushi Inouye, and Satoshi Kawata. Tip-enhanced coherent anti-stokes raman scattering for vibrational nanoimaging. *Phys. Rev. Lett.*, 92(22):220801, 2004.

[Ise96] Arieh Iserles. *A First Course in the Numerical Analysis of Differential Equations.* Cambridge [England]; New York: Cambridge University Press, 1996.

[Isr92] Jacob Israelachvili. *Intermolecular and Surface Forces with Applications to Colloidal and Biological Systems.* London; San Diego, CA : Academic Press, 1992.

[ITJ05] Akira Ishimaru, John Rhodes Thomas, and Sermsak Jaruwatanadilok. Electromagnetic waves over half-space metamaterials of arbitrary permittivity and permeability. *IEEE Transactions on Antennas and Propagation*, 53(3):915–921, 2005.

[Jac46] C.G.J. Jacobi. Über ein leichtes verfahren, die in der theorie der sä"cularstörungen vorkommenden gleichungen numerisch aufzulösen. *J. reine angew. Math*, 30:51–94, 1846.

[Jac99] John David Jackson. *Classical Electrodynamics.* New York: Wiley, 1999.

[Jam70] Pierre Jamet. On the convergence of finite-difference approximations to one-dimensional singular boundary-value problems. *Numer. Math.*, 14:355–378, 1970.

[Jam76] Pierre Jamet. Estimations de l'erreur pour des éléments finis droits preque dégénérés. *Rairo Anal. Numer.*, 10:43–60, 1976.

[JC72] P.B. Johnson and R.W. Christy. Optical constants of the noble metals. *Phys. Rev. B*, 6(12–15):4370–4379, 1972.

[Jin02] Jianming Jin. *The Finite Element Method in Electromagnetics.* Wiley – IEEE Press, 2002.

[Jir78] J. Jirousek. Basis for development of large finite elements locally satisfying all field equations. *Comp. Meth. Appl. Mech. Eng.*, 14:65–92, 1978.

[JJ01] Steven G. Johnson and J.D. Joannopoulos. Block-iterative frequency-domain methods for maxwell's equations in a planewave basis. *Opt. Express*, 8(3):173–190, 2001.

[JL77] J. Jirousek and N. Leon. A powerful finite element for plate bending. *Comp. Meth. Appl. Mech. Eng.*, 12:77–96, 1977.

[Joh87] Sajeev John. Strong localization of photons in certain disordered dielectric superlattices. *Phys. Rev. Lett.*, 58(23):2486–2489, Jun 1987.

[JZ97] J. Jirousek and A.P. Zielinski. Survey of Trefftz-type element formulations. *Computers & Structures*, 63:225–242, 1997.

[Kam99] A. Kameari. Symmetric second order edge elements for triangles and tetrahedra. *IEEE Trans. Magn.*, 35(3):1394–1397, 1999.

[Kap04] Igor Kaporin. The aggregation and cancellation techniques as a practical tool for faster matrix multiplication. *Theoretical Computer Science*, 315(2-3):469–510, 2004.

[KB96] L.R. Petzold K.E. Brenan, S.L. Campbell. *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations.* Philadelphia: Society for Industrial and Applied Mathematics, 1996.

[KB98] Y. Krongauz and T. Belytschko. EFG approximation with discontinuous derivatives. *Int. J. Numer. Meth. Engng.*, 41:1215–1233, 1998.

[KBA01] George Karniadakis, Ali Beskok, and Narayan Aluru. *Micro Flows.* Springer, 2001.

[KCZS03] K. Lance Kelly, E. Coronado, L.L. Zhao, and G.C. Schatz. The optical properties of metal nanoparticles: The influence of size, shape, and dielectric environment. *Journal of Physical Chemistry B*, 107(3):668–677, 2003.

[KH04] F. Keilmann and R. Hillenbrand. Near-field microscopy by elastic light scattering from a tip. *Philosophical Transactions of the Royal Society of London Series A – Mathematical, Physical and Engineering Sciences*, 362(1817):787–805, 2004.

[KM01] Jörg P. Kottmann and Olivier J. F. Martin. Retardation-induced plasmon resonances in coupled nanoparticles. *Optics Letters*, 26:1096–1098, 2001.

[Kny98] A.V. Knyazev. Preconditioned eigensolvers — an oxymoron? *Electronic Transactions on Numerical Analysis*, 7:104–123, 1998.

[Kny01] A.V. Knyazev. Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method. *SIAM Journal on Scientific Computing*, 23(2):517–541, 2001.

[Kom75] J. Komorowski. On finite-dimensional approximations of the exterior differential, codifferential and laplacian on a riemannian manifold. *Bull. Acad. Polonaise Sc. (Math., Astr., Phys.)*, 23(9):999–1005, 1975.

[Kor94] J. Korringa. Early history of multiple scattering theory for ordered systems. *Phys. Rep.*, 238(6):341–360, 1994.

[Kot85] Peter Robert Kotiuga. *Hodge Decompositions and Computational Electromagnetics.* PhD thesis, McGill University, Montreal, Canada, 1985.

[KR54] W. Kohn and N. Rostoker. Solution of the schrödinger equation in periodic lattices with an application to metallic lithium. *Phys. Rev.*, 94(5):1111–1120, Jun 1954.

[KTH00] M. Koshiba, Y. Tsuji, and M. Hikari. Time-domain beam propagation method and its application to photonic crystal circuits. *IEEE J. of Lightwave Technology*, 18(1):102–110, 2000.

[Kum03] Manoj Kumar. A new finite difference method for a class of singular two-point boundary value problems. *Applied Mathematics and Computation*, 143(2-3):551–557, 2003.

[Kř92] M. Křižek. On the maximum angle condition for linear tetrahedral elements. *SIAM J. Numer. Analysis*, 29(2):513–520, 1992.

[KV95] Uwe Kreibig and Michael Vollmer. *Optical Properties of Metal Clusters.* Springer, 1995.

[KWC80] M. Kerker, D.S. Wang, and H. Chew. Surface enhanced raman-scattering (sers) by molecules adsorbed at spherical particles. *Applied Optics*, 19:3373–3388, 1980.

[Lad69] O.A. Ladyzhenskaya. *The Mathematical Theory of Viscous Incompressible Flows.* Gordon and Breach, London, 1969.

[Lam91] John Denholm Lambert. *Numerical methods for Ordinary Differential Systems: the initial value problem.* Chichester; New York : Wiley, 1991.

[Lam99] S.K. Lamoreaux. Resource letter cf-1: Casimir force. *American Journal of Physics*, 67(10):850–861, 1999.

[LDFH05] B. Lombardet, L. A. Dunbar, R. Ferrini, and R. Houdré. Fourier analysis of Bloch wave propagation in photonic crystals. *J. Opt. Soc. Am. B*, 22:1179–1190, 2005.

[LED+06] Stefan Linden, Christian Enkrich, Gunnar Dolling, Matthias W. Klein, Jiangfeng Zhou, Thomas Koschny, Costas M. Soukoulis, Sven Burger, Frank Schmidt, and Martin Wegener. Photonic metamaterials: Magnetism at optical frequencies. *IEEE J. of Selected Topics in Quantum Electronics*, 12(6):1097–1105, 2006.

[Lek96] J. Lekner. Optical properties of isotropic chiral media. *Pure and Applied Optics: Journal of the European Optical Society, Part A*, 5(4):417–443, 1996.

[Leo06] Ulf Leonhardt. Optical conformal mapping. *Science*, 312(5781):1777–1780, 2006.

[LeV96] Randall J. LeVeque. *Numerical Methods for Conservation Laws.* Birkhauser, 1996.

[LeV02a] Randall J. LeVeque. *Finite Volume Methods for Hyperbolic Problems.* Cambridge [England]; New York: Cambridge University Press, 2002.

[Lev02b] Yan Levin. Electrostatic correlations: from plasma to biology. *Reports on Progress in Physics*, 65(11):1577–1632, 2002.

[Lie93a] A. Liebsch. Surface-plasmon dispersion and size dependence of mie resonance: Silver versus simple metals. *Phys. Rev. B*, 48(15):11317–11328, Oct 1993.

[Lie93b] A. Liebsch. Surface plasmon dispersion of ag. *Phys. Rev. Lett.*, 71(1):145–148, Jul 1993.

[Lif56] E.M. Lifshitz. The theory of molecular attractive forces between solids. *Sov. Phys. JETP*, 2:73–83, 1956.

[Liu02] G. R. Liu. *Mesh Free Methods: Moving Beyond the Finite Element Method.* CRC Press, 2002. See Chapter 7 for Meshless Local Petrov-Galerkin method.

[LJJP02] Chiyan Luo, Steven G. Johnson, J. D. Joannopoulos, and J. B. Pendry. All-angle negative refraction without negative effective index. *Phys. Rev. B*, 65(20):201104, May 2002.

[LJZ95] W. Liu, S. Jun, and Y. Zhang. Reproducing kernel particle methods. *Int. J. Numer. Meth. Fluids*, 20:1081–1106, 1995.

[LL84] L.D. Landau and E.M. Lifshitz. *Electrodynamics of Continuous Media.* Oxford; New York: Pergamon, 1984.

[LLN03] Larry A. Lambe, Richard Luczak, and John W. Nehrbass. A new finite difference method for the Helmholtz equation using symbolic computation. *International Journal of Computational Engineering Science*, 4(1):121–144, 2003.

[LM03] A. Lakhtakia and G. Mulholland. On two numerical techniques for light scattering by dielectric agglomerated structures. *J. of Research of the Nat. Inst. of Standards and Tech.*, 98(6):699–716, 2003.

[LR67] M.L. Levin and S.M. Rytov. *Teoriia Ravnovesnykh Teplovykh Fluktuatsii v Elektrodinamike. (Theory of steady-state thermal fluctuations in electrodynamics.).* Moskva, Nauka, 1967.

[LR80] R.E. Lynch and J.R. Rice. A high-order difference method for differential equations. *Math. Comp.*, 34:333–372, 1980.

[LSB03] Kuiru Li, Mark I. Stockman, and David J. Bergman. Self-similar chain of metal nanospheres as an efficient nanolens. *Phys. Rev. Lett.*, 91(22):227402, 2003.

[LSB06] Kuiru Li, Mark I. Stockman, and David J. Bergman. Li, Stockman, and Bergman reply. *Phys. Rev. Lett.*, 97(7):079702, 2006.

[LSC91] J.F. Lee, D.K. Sun, and Z.J. Cendes. Tangential vector finite-elements for electromagnetic-field computation. *IEEE Trans. Magn.*, 27(5):4032–4035, 1991.

[LVH04] Domenico Lahaye, Stefan Vandewalle, and Kay Hameyer. An algebraic multilevel preconditioner for field-circuit coupled problems. *J. Comput. Appl. Math.*, 168(1-2):267–275, 2004.

[LW95] P.T.S. Liu and J.P. Webb. Analysis of 3d microwave cavities using hierarchal vector finite elements. *IEE Proceedings - Microwaves, Antennas and Propagation*, 142(5):373–378, 1995.

[lX95] Y. l. Xu. Electromagnetic scattering by an aggregate of spheres. *Appl. Opt.*, 34:4573–4588, 1995.

[LYX06] Zhipeng Li, Zhilin Yang, and Hongxing Xu. Comment on "Self-similar chain of metal nanospheres as an efficient nanolens". *Phys. Rev. Lett.*, 97(7):079701, 2006.

[LZCS03] Jensen Li, Lei Zhou, C. T. Chan, and P. Sheng. Photonic band gap from a stack of positive and negative index materials. *Phys. Rev. Lett.*, 90(8):083901, 2003.

[M̈78] W. Müller. Analytic torsion and R-torsion of Riemannian manifolds. *Advances in Mathematics*, 28:233–305, 1978.

[MA05] Stefan A. Maier and Harry A. Atwater. Plasmonics: Localization and guiding of electromagnetic energy in metal/dielectric structures. *Journal of Applied Physics*, 98:011101, 2005.

[Mac91] Daniel W. Mackowski. Analysis of radiative scattering for multiple sphere configurations. *Proc. Royal Soc. London A*, 433:599–614, 1991.

[Mai07] Stefan A. Maier. *Plasmonics: Fundamentals and Applications*. Springer, 2007.

[Mal99] Stéphane Mallat. *A Wavelet Tour of Signal Processing (Wavelet Analysis & Its Applications)*. Academic Press, 1999. See p.28 for the Poisson summation formula.

[Man45] L.I. Mandelshtam. Group velocity in crystalline arrays. *Zh. Eksp. Teor. Fiz.*, 15:475–478, 1945.

[Man47] L.I. Mandelshtam. *Polnoe Sobranie Trudov, v. 2*. Akademiia Nauk SSSR, 1947.

[Man50] L.I. Mandelshtam. *Polnoe Sobranie Trudov, v. 5*. Akademiia Nauk SSSR, 1950.

[Map50] C.B. Maple. The Dirichlet problem: bound at a point for the solution and its derivatives. *Quart. Appl. Math.*, 8:213–228, 1950.

[Mat97] C. Mattiussi. An analysis of finite volume, finite element, and finite difference methods using some concepts from algebraic topology. *J. of Comp. Phys.*, 133(2):289–309, 1997.

[May03] I.D. Mayergoyz. *Mathematical Models of Hysteresis and Their Applications*. Amsterdam; Boston: Elsevier Academic Press, 2003.

[Maz05] Martial Mazars. Lekner summations and ewald summations for quasi-two-dimensional systems. *Molecular Physics*, 103(9):1241–1260, 2005.

[MB96] J.M. Melenk and I. Babuška. The partition of unity finite element method: Basic theory and applications. *Comput. Methods Appl. Mech. Engrg.*, 139:289–314, 1996.

[MB05] D.O.S. Melville and R.J. Blaikie. Super-resolution imaging through a planar silver layer. *Optics Express*, 13(6):2127–2134, 2005.

[McC89] Stephen Fahrney McCormick. *Multilevel Adaptive Methods for Partial Differential Equations.* Philadelphia, PA : Society for Industrial and Applied Mathematics, 1989.

[MD01] P. Monk and L. Demkowicz. Discrete compactness and the approximation of Maxwell's equations in $\mathbb{R}^3$. *Math. of Comp.*, 70(234):507–523, 2001.

[MDH+99] S. Moskow, V. Druskin, T. Habashy, P. Lee, and S. Davydycheva. A finite difference scheme for elliptic equations with rough coefficients using a Cartesian grid nonconforming to interfaces. *SIAM J. Numer. Analysis*, 36(2):442–464, 1999.

[MEHV02] Esteban Moreno, Daniel Erni, Christian Hafner, and Rüdiger Vahldieck. Multiple multipole method with automatic multipole setting applied to the simulation of surface plasmons in metallic nanostructures. *Opt. Soc. Am. A*, 19(1):101–111, 2002.

[Mel99] J.M. Melenk. Operator adapted spectral element methods I: harmonic and generalized harmonic polynomials. *Numer. Math.*, 84:35–69, 1999.

[Mer04] R. Merlin. Analytical solution of the almost-perfect-lens problem. *Applied Physics Letters*, 84(8):1290–1292, 2004.

[Mes02] René Messina. Image charges in spherical geometry: Application to colloidal systems. *The J. of Chem. Phys.*, 117(24):11062–11074, 2002.

[Meu07] Gérard Meunier, editor. *The Finite Element Method for Electromagnetic Modeling.* ISTE Publishing Company, 2007.

[MFZ05a] Isaak D. Mayergoyz, Donald R. Fredkin, and Zhenyu Zhang. Electrostatic (plasmon) resonances in nanoparticles. *Physical Review B*, 72(15):155412, 2005.

[MFZ+05b] R. Moussa, S. Foteinopoulou, Lei Zhang, G. Tuttle, K. Guven, E. Ozbay, and C. M. Soukoulis. Negative refraction and superlens behavior in a two-dimensional photonic crystal. *Physical Review B (Condensed Matter and Materials Physics)*, 71(8):085106, 2005.

[MGKH06] R. Meisels, R. Gajic, F. Kuchar, and K. Hingerl. Negative refraction and flat-lens focusing in a 2d square-lattice photonic crystal at microwave and millimeter wave frequencies. *Opt. Express*, 14:6766–6777, 2006.

[MGS01] J.M. Melenk, K. Gerdes, and C. Schwab. Fully discrete hp-finite elements: fast quadrature. *Comput. Methods Appl. Mech. Engrg.*, 190:4339–4364, 2001.

[MHW01] Yves C. Martin, Hendrik F. Hamann, and H. Kumar Wickramasinghe. Strength of the electric field in apertureless near-field optical microscopy. *J. of Appl. Phys.*, 89(10):5774–5778, 2001.

[Mic94] Ronald E. Mickens. *Nonstandard Finite Difference Models of Differential Equations.* Singapore; River Edge, N.J.: World Scientific, 1994.

[Mic00] Ronald E. Mickens, editor. *Applications of Nonstandard Finite Difference Schemes.* Singapore; River Edge, N.J.: World Scientific, 2000.

[Mik64] S.G. Mikhlin. *Variational Methods in Mathematical Physics.* Oxford, New York, Pergamon Press, 1964.

[Mik65] S.G. Mikhlin. *The Problem of the Minimum of a Quadratic Functional.* San Francisco, Holden-Day, 1965.

[Mil70] W.E. Milne. *Numerical Solution of Differential Equations.* New York, Dover Publications, 1970.

[Mil94] P.W. Milonni. *The Quantum Vacuum : An Introduction to Quantum Electrodynamics.* Boston : Academic Press, 1994.

[Mil04] P.W. Milonni. *Fast Light, Slow Light and Left-Handed Light.* Taylor & Francis, 2004.

[Min03] J.R. Minkel. Left-handed materials debate heats up. *Phys. Rev. Focus*, 9:755–760, 2003.

[Mit89] William F. Mitchell. A comparison of adaptive refinement techniques for elliptic problems. *ACM Trans. Math. Softw.*, 15(4):326–347, 1989.

[Mit92] William F. Mitchell. Optimal multilevel iterative methods for adaptive grids. *SIAM J. on Sci & Stat. Computing*, 13(1):146–167, 1992.

[MKSD01] Michelle Duval Malinsky, K. Lance Kelly, George C. Schatz, and Richard P. Van Duyne. Nanosphere lithography: effect of substrate on the localized surface plasmon resonance spectrum of silver nanoparticles. *J. Phys. Chem. B*, 105:2343–2350, 2001.

[ML78] C. Moler and C. Van Loan. Nineteen dubious ways to compute the exponential of a matrix. *SIAM Review*, 20:801–836, 1978.

[ML03] C. Moler and C. Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45(1):3–49, 2003.

[MLH+05] D. Mehtani, N. Lee, R.D. Hartschuh, A. Kisliuk, M.D. Foster, A. P. Sokolov, and J.F. Maguire. Nano-raman spectroscopy with side-illumination optics. *Journal of Raman Spectroscopy*, 36(11):1068–1075, 2005.

[MLH+06] D. Mehtani, N. Lee, R.D. Hartschuh, A. Kisliuk, M.D. Foster, A.P. Sokolov, F. Čajko, and I. Tsukerman. Optical properties and enhancement factors of the tips for apertureless near-field optics. *Journal of Optics A: Pure and Applied Optics*, 8:S183–S190, 2006.

[MM96] D.W. Mackowski and M.I. Mishchenko. Calculation of the t matrix and the scattering matrix for ensembles of spheres. *J. Optic. Soc. Amer. A*, 13:2266–2277, 1996.

[MN76] J. Mahanty and B.W. Ninham. *Dispersion Forces.* London ; New York : Academic Press, 1976.

[MN06] Graeme W. Milton and Nicolae-Alexandru P. Nicorovici. On the cloaking effects associated with anomalous localized resonance. *Proc. R. Soc. Lond. A*, 2006.

[MNB05] R.C. McPhedran, N.A. Nicorovici, and L.C. Botten. Neumann series and lattice sums. *J. of Math. Phys.*, 46(8):083509, 2005.

[MNMP05] Graeme W. Milton, Nicolae-Alexandru P. Nicorovici, Ross C. McPhedran, and Viktor A. Podolskiy. A proof of superlensing in the quasistatic regime, and limitations of superlenses in this regime due to anomalous localized resonance. *Proc. R. Soc. Lond. A*, 461(2064):3999–4034, Dec 2005.

[MNS02] Pedro Morin, Ricardo H. Nochetto, and Kunibert G. Siebert. Convergence of adaptive finite element methods. *SIAM Rev.*, 44(4):631–658, 2002.

[Mon03] Peter Monk. *Finite Element Methods for Maxwell's Equations.* Oxford: Clarendon Press, 2003, 2003.

[Mor02] Alexander Moroz. Metallo-dielectric diamond and zinc-blende photonic crystals. *Phys. Rev. B*, 66(11):115109, Sep 2002.

[MPC+94] K. K. Mei, R. Pous, Z. Chen, Y. W. Liu, and M. D. Prouty. Measured equation of invariance: A new concept in field computation. *IEEE Trans. Antennas Propagat.*, 42:320–327, 1994.

[MRB+93] R.D. Meade, A.M. Rappe, K.D. Brommer, J.D. Joannopoulos, and O.L. Alerhand. Accurate theoretical analysis of photonic band-gap materials. *Phys. Rev. B*, 48(11):8434–8437, Sep 1993.

[MRB+97] R.D. Meade, A.M. Rappe, K.D. Brommer, J.D. Joannopoulos, and O.L. Alerhand. Erratum: Accurate theoretical analysis of photonic band-gap materials [phys. rev. b 48, 8434 (1993)]. *Phys. Rev. B*, 55(23):15942, Jun 1997.

[MS73] C.B. Moler and G.W. Stewart. An algorithm for generalized matrix eigenvalue problems. *SIAM J. Numer. Analysis*, 10(2):241–256, 1973.

[MT98] M.I. Mishchenko and L.D. Travis. Capabilities and limitations of a current fortran implementation of the t-matrix method for randomly oriented, rotationally symmetric scatterers. *J. Quant. Spectrosc. Radiat. Transfer*, 60:309–324, 1998.

[MTL02] M.I. Mishchenko, L.D. Travis, and A.A. Lacis. *Scattering, Absorption, and Emission of Light by Small Particles.* Cambridge University Press, 2002.

[MTL06] M.I. Mishchenko, L.D. Travis, and A.A. Lacis. *Multiple Scattering of Light by Particles: Radiative Transfer and Coherent Backscattering.* Cambridge University Press, 2006.

[MTM96] M.I. Mishchenko, L.D. Travis, and D.W. Mackowski. T-matrix computations of light scattering by nonspherical particles: A review. *J. Quant. Spectrosc. Radiat. Transfer*, 55:535–575, 1996.

[Mun00] E.H. Mund. A short survey on preconditioning techniques in spectral calculations. *Applied Numer. Math.*, 33:61–70, 2000.

[MVB+04] M.I. Mishchenko, G. Videen, V.A. Babenko, N.G. Khlebtsov, and T. Wriedt. T-matrix theory of electromagnetic scattering by particles and its applications: A comprehensive reference database. *J. Quant. Spectrosc. Radiat. Transfer*, 88:357–406, 2004.

[MW79] Wilhelm Magnus and Stanley Winkler. *Hill's Equation.* New York: Dover Publications, 1979. See p.28 for the Poisson summation formula.

[MZ95] S.A. Meguid and Z.H. Zhu. A novel finite element for treating inhomogeneous solids. *Int. J. for Numer. Meth. Eng.*, 38:1579–1592, 1995.

[Né80] Jean-Claude Nédélec. Mixed finite elements in $\mathbb{R}^3$. *Numer. Math.*, 35:315–341, 1980.

[Né86] Jean-Claude Nédélec. A new family of mixed finite elements in $\mathbb{R}^3$. *Numer. Math.*, 50:57–81, 1986.

[Neh96] John W. Nehrbass. *Advances in finite difference methods for electromagnetic modeling.* PhD thesis, Ohio State University, 1996.

[NGG+04] C.L. Nehl, N.K. Grady, G.P. Goodrich, F. Tam, N.J. Halas, and J.H. Hafner. Scattering spectra of single gold nanoshells. *Nano Letters*, 4(12):2355–2359, 2004.

[NGS00] T.T. Nguyen, A. Yu. Grosberg, and B. I. Shklovskii. Macroions in salty water with multivalent ions: giant inversion of charge. *Phys. Rev. Lett.*, 85:1568–1571, 2000.

[NMM94] N.A. Nicorovici, R.C. McPhedran, and G.W. Milton. Optical and dielectric properties of partially resonant composites. *Phys. Rev. B*, 49(12):8479–8482, Mar 1994.

[NO99] R.R. Netz and H. Orland. Field theory for charged fluids and colloids. *Europhysics Letters*, 45(6):726–732, 1999.

[NO00] R.R. Netz and H. Orland. Beyond poisson-boltzmann: Fluctuation effects and correlation functions. *The European Phys. J. E*, 1:203–214, 2000.

[Not00] M. Notomi. Theory of light propagation in strongly modulated photonic crystals: Refractionlike behavior in the vicinity of the photonic band gap. *Phys. Rev. B*, 62(16):10696–10705, Oct 2000.

[NSR04] C.C. Neacsu, G.A. Steudle, and M.B. Raschke. Plasmonic light scattering from nanoscopic metal tips. *Appl. Phys. B*, 80:295–300, 2004.

[NVG03] M. Nieto-Vesperinas and N. Garcia. Nieto-vesperinas and garcia reply:. *Phys. Rev. Lett.*, 91(9):099702, 2003.

[OBB98] J.T. Oden, I. Babuška, and C.E. Baumann. A discontinuous hp finite element method for diffusion problems. *J. of Comp. Phys.*, 146:491–519, 1998.

[Ohs94a] Hiroyuki Ohshima. Electrostatic interaction between a hard-sphere with constant surface-charge density and a soft-sphere – polarization effect of a hard-sphere. *J. Colloidal & Interface Sci.*, 168(1):255–265, 1994.

[Ohs94b] Hiroyuki Ohshima. Electrostatic interaction between two dissimilar spheres: an explicit analytic expression. *J. of Colloid & Interface Sci*, 162(2):487–495, 1994.

[Ohs95] Hiroyuki Ohshima. Electrostatic interaction between two dissimilar spheres with constant surface charge density. *J. of Colloid & Interface Sci*, 170(2):432–439, 1995.

[OLB+83] M.A. Ordal, L.L. Long, R.J. Bell, S.E. Bell, R.R. Bell, R.W. Alexander, and C.A. Ward. Optical properties of the metals Al, Co, Cu, Au, Fe, Pb, Ni, Pd, Pt, Ag, Ti and W in the infrared and far-infrared. *Appl. Opt.*, 22:1099–1119, 1983.

[OP01] J. Tinsley Oden and S. Prudhomme. Goal-oriented error estimation and adaptivity for the finite element method. *Computers & Math. with Appl.*, 41:735–756, 2001.

[Ors80] S.A. Orszag. Spectral methods for problems in complex geometries. *J. Comp. Phys.*, 37(1):70–92, 1980.

[OS97] Walter Oevel and Mark Sofroniou. Symplectic runge–kutta-schemes ii: Classification of symmetric methods, 1997.

[Ott96] Hans Christian Ottinger. *Stochastic Processes in Polymeric Fluids: Tools and Examples for Developing Simulation Algorithms.* Springer, 1996.

[Paf59] V.E. Pafomov. K voprosu o perehodnom izluchenii i izluchenii Vavilova-Cherenkova. *Zh. Eksp. Teor. Fiz.*, 36(6):1853–1858, 1959.

[Pan84] Victor Pan. How can we speed up matrix multiplication? *SIAM Rev.*, 26(3):393–415, 1984.

[Par80] B.N. Parlett. *The Symmetric Eigenvalue Problem.* Prentice-Hall, Englewood Cliffs, N.J., 1980.

[Par06] Vozken Adrian Parsegian. *Van der Waals Forces : A Handbook for Biologists, Chemists, Engineers, and Physicists.* New York : Cambridge University Press, 2006.

[Pat80] S.V. Patankar. *Numerical Heat Transfer and Fluid Flow.* John Benjamins Publishing Co., 1980.

[PB02] John A. Pelesko and David H. Bernstein. *Modeling MEMS and NEMS.* CRC Press, 2002.

[PDL84] D.W. Pohl, W. Denk, and M. Lanz. Optical stethoscopy: Image recording with resolution lambda/20. *Applied Physics Letters*, 44(7):651–653, 1984.

[PE02] A.L. Pokrovsky and A.L. Efros. Sign of refractive index and group velocity in lefthanded media. *Solid State Communications*, 124:283–287, 2002.

[PE03] A.L. Pokrovsky and A.L. Efros. Diffraction theory and focusing of light by a slab of left-handed material. *Physica B: Condensed Matter*, 338:333–337, 2003.

[Pel96] J. Peltoniemi. Electromagnetic scattering by irregular grains usingvariational volume integral equation method. *J. Quant. Spectrosc. Radiat. Transfer*, 55(5):637–647, 1996.

[Pen00] J.B. Pendry. Negative refraction makes a perfect lens. *Phys. Rev. Lett.*, 85(18):3966–3969, Oct 2000.

[Pen01] John Pendry. Pendry replies:. *Phys. Rev. Lett.*, 87(24):249704, Nov 2001.

[Pet95] Henrik G. Petersen. Accuracy and efficiency of the particle mesh ewald method. *The J. of Chem. Phys.*, 103(9):3668–3679, 1995.

[PGL+03] C. G. Parazzoli, R. B. Greegor, K. Li, B. E. C. Koltenbah, and M. Tanielian. Experimental verification and simulation of negative index of refraction using snell's law. *Phys. Rev. Lett.*, 90(10):107401, Mar 2003.

[PHRS99] J.B. Pendry, A.J. Holden, D.J. Robbins, and W.J. Stewart. Magnetism from conductors and enhanced nonlinear phenomena. *IEEE Trans. on Microwave Theory & Tech.*, 47(11):2075–2084, Nov 1999.

[Pis84] Sergio Pissanetzky. *Sparse Matrix Technology.* London : Academic Press, 1984.

[PLV+04] P. V. Parimi, W. T. Lu, P. Vodo, J. Sokoloff, J. S. Derov, and S. Sridhar. Negative refraction and left-handed electromagnetism in microwave photonic crystals. *Phys. Rev. Lett.*, 92(12):127401, 2004.

[PM96] M.A. Paesler and P.J. Moyer. *NearField Optics: Theory, Instrumentation and Applications.* New York: John Wiley & Sons, Inc., 1996.

[PO02] S. Prudhomme and J. Tinsley Oden. Computable error estimators and adaptive techniques for fluid flow problems. In T. Barth and H. Deconinck, editors, *Error Estimation and Adaptive Discretization Methods in Computational Fluid Dynamics*, Lecture Notes in Computational Science and Engineering, vol. 25, pages 207–268. Springer-Verlag, Heidelberg, 2002.

[PP62] Wolfgang K.H. Panofsky and Melba Phillips. *Classical Electricity and Magnetism.* Reading, Mass., Addison-Wesley Pub. Co., 1962.

[PPL+98] B. Palpant, B. Prével, J. Lermé, E. Cottancin, M. Pellarin, M. Treilleux, A. Perez, J.L. Vialle, and M. Broyer. Optical properties of gold clusters in the size range 2˘4 nm. *Phys. Rev. B*, 57(3):1963–1970, Jan 1998.

[PR03] J.B. Pendry and S.A. Ramakrishna. Focussing light using negative refraction. *J. Phys.: Condens. Matter*, 15:6345–6364, 2003.

[PR04] Richard Pasquetti and Francesca Rapetti. Spectral element methods on triangles and quadrilaterals: comparisons and applications. *J. Comp. Phys.*, 198:349–362, 2004.

[Pra03] Paras N. Prasad. *Introduction to Biophotonics.* Wiley-Interscience, 2003.

[Pra04] Paras N. Prasad. *Nanophotonics.* Wiley-Interscience, 2004.

[PRM98] Andrew F. Peterson, Scott L. Ray, and Raj Mittra. *Computational Methods for Electromagnetics.* Oxford University Press, 1998.

[PS04] J.B. Pendry and D.R. Smith. Reversing light with negative refraction. *Phys. Today*, 57:37–43, 2004.

[PSS06] J.B. Pendry, D. Schurig, and D.R. Smith. Controlling Electromagnetic Fields. *Science*, 312(5781):1780–1782, 2006.

[PT02] L. Proekt and I. Tsukerman. Method of overlapping patches for electromagnetic computation. *IEEE Trans. Magn.*, 38(2):741–744, 2002.

[PTFY03] A. Plaks, I. Tsukerman, G. Friedman, and B. Yellen. Generalized Finite Element Method for magnetized nanoparticles. *IEEE Trans. Magn.*, 39(3):1436–1439, 2003.

[PTPT00] A. Plaks, I. Tsukerman, S. Painchaud, and L. Tabarovsky. Multigrid methods for open boundary problems in geophysics. *IEEE Trans. Magn.*, 36(4):633–638, 2000.

[PWT07] H. Pinheiro, J.P. Webb, and I. Tsukerman. Flexible local approximation models for wave scattering in photonic crystal devices. *IEEE Trans. Magn.*, 43(4):1321–1324, 2007.

[Qui96] Michael Quinten. Optical constants of gold and silver clusters in the spectral range between 1.5 ev and 4.5 ev. *Zeitschrift für Physik B Condensed Matter*, 101(2):211–217, 1996.

[Qui99] Michael Quinten. Optical effects associated with aggregates of clusters. *J. of Cluster Science*, 10(2):319–358, 1999.

[QV99] Alfio Quarteroni and Alberto Valli. *Domain Decomposition Methods for Partial Differential Equations.* Oxford; New York: Clarendon Press, 1999.

[RÖ3] Ulrich Rüde. Fully adaptive multigrid methods. *SIAM J. Numer. Anal.*, 30(1):230–248, 1993.

[RAH01] W. Rocchia, E. Alexov, and B. Honig. Extending the applicability of the nonlinear Poisson–Boltzmann equation: Multiple dielectric constants and multivalent ions. *J. Phys. Chem. B*, 105(28):6507–6514, 2001.

[Rak72] Yu.V. Rakitskii. A methodology for a systematic time step increase in the numerical integration of ordinary differential equations. *Doklady Akademii Nauk SSSR (Mathematics). Proceedings of the Academy of Sciences of the USSR. Comptes rendus de l'Académie des sciences de l'URSS*, 207(4):793–795, 1972.

[Ram05] S. Anantha Ramakrishna. Physics of negative refractive index materials. *Rep. Prog. Phys.*, 68:449–521, 2005.

[RDS97] Tamar Schlick Robert D. Skeel, Guihua Zhang. A family of symplectic integrators: stability, accuracy, and molecular dynamics applications. *SIAM J. Sci. Comput.*, 18:203–222, 1997.

[Rek80] Karel Rektorys. *Variational Methods in Mathematics, Science, and Engineering.* Dordrecht ; Boston : D. Reidel, 1980.

[RI00] Z. Ren and N. Ida. Solving 3d eddy current problems using second order nodal and edge elements. *IEEE Trans. Magn.*, 36(4):746–750, 2000.

[Ric03] David Richards. Near-field microscopy: throwing light on the nanoworld. *Phil. Trans. R. Soc. Lond. A*, 361(1813):2843–2857, 2003.

[RKG88] M.O. Robbins, K. Kremer, and G. Grest. Phase diagram and dynamics of Yukawa systems. *J. Chem. Phys.*, 88:3286–3312, 1988.

[Rob05] Sara Robinson. Toward an optimal algorithm for matrix multiplication. *SIAM News*, 38(9), November 2005.

[Rod83] D. Rodger. Finite-element method for calculating power frequency 3-dimensional electromagnetic-field distributions. *IEE Proceedings-A: Science, Measurement and Technology*, 130(5):233–238, 1983.

[RR90] Eric S. Reiner and Clayton J. Radke. Variational approach to the electrostatic free energy in charged colloidal suspensions: general theory for open systems. *J. Chem. Soc., Faraday Trans.*, 86:3901–3912, 1990.

[RS76] G.W. Reddien and L.L. Schumaker. On a collocation method for singular two-point boundary value problems. *Numerische Mathematik*, 25:427–432, 1976.

[RS04] A. Rubinstein and S. Sherman. Influence of the solvent structure on the electrostatic interactions in proteins. *Biophys J.*, 87:1544–1557, 2004.

[RSY+85] Yu.V. Rakitskii, E.D. Shchukin, V.S. Yushchenko, I.A. Tsukerman, Yu.B. Suris, and A.I. Slutsker. Mechanism of the formation of energy fluctuation and a method for studying it. *Doklady. Physical chemistry: Proceedings of the Academy of Sciences of the USSR. Comptes rendus de l'Académie des sciences de l'URSS*, 285(4):1204–1207, 1985.

[RT74] G.D. Raithby and K.E. Torrance. Upstream weighted differencing schemes and their application to elliptic problems involving fluid flow. *J. of Computers & Fluids*, 2:191–206, 1974.

[RT75] Donald J. Rose and R. Endre Tarjan. Algorithmic aspects of vertex elimination. In *STOC '75: Proceedings of seventh annual ACM symposium on Theory of computing*, pages 245–254, New York, NY, USA, 1975. ACM Press.

[RT77] P.-A. Raviart and J.M. Thomas. A mixed finite element method for 2nd order elliptic problems. *Lecture Notes in Mathematics, Springer, Berlin*, 606:292–315, 1977.

[RUC79] IU.V. Rakitskii, S.M. Ustinov, and I.G. Chernorutskii. *Chislennye Metody Resheniia Zhestkikh Sistem [in Russian]. (Numerical Methods for Stiff Systems).* Moskva: Nauka, 1979.

[Rud76] Walter Rudin. *Principles of Mathematical Analysis.* McGraw-Hill, 1976.

[Ryt53] S.M. Rytov. *Teoriia Elektricheskikh Fluktuatsii i Teplovogo Izlucheniia (Theory of Electric Fluctuations and Thermal Radiation).* Moskva, Izd-vo Akademii nauk SSSR, 1953.

[Saa92] Yousef Saad. *Numerical Methods for Large Eigenvalue Problems.* Manchester University Press, Manchester, UK, 1992.

[Saa03] Yousef Saad. *Iterative Methods for Sparse Linear Systems.* Philadelphia: Society for Industrial and Applied Mathematics, 2003.

[Sad97] John Elie Sader. Accurate analytic formulae for the far field effective potential and surface charge density of a uniformly charged sphere. *J. of Colloid & Interface Sci*, 188:508–510, 1997. Article No. CS974776.

[Sak05] Kazuaki Sakoda. *Optical Properties of Photonic Crystals.* Berlin; New York: Springer, 2005.

[Sam01] A.A. Samarskii. *The Theory of Difference Schemes.* New York: M. Decker, 2001.

[SB91] Barna Szabó and Ivo Babuška. *Finite Element Analysis.* New York : Wiley, 1991.

[SBC00] T. Strouboulis, I. Babuška, and K.L. Copps. The design and analysis of the Generalized Finite Element Method. *Computer Meth. in Appl. Mech. & Eng.*, 181(1–3):43–69, 2000.

[SC99] Sheppard Salon and M. V.K. Chari. *Numerical Methods in Electromagnetism.* Academic Press, 1999.

[SC00] Gwenaël Salin and Jean-Michel Caillol. Ewald sums for Yukawa potentials. *J. of Chem. Phys.*, 113(23):10459–10463, 2000.

[Sch04] A. Schuster. *Introduction to the Theory of Optics.* Edward Arnold, London, 1904.

[Sch66] Laurent Schwartz. *Mathematics for the Physical Sciences.* Addison-Wesley Pub. Co., 1966.

[Sch73] I.J. Schoenberg. *Cardinal Spline Interpolation (CBMS-NSF Regional Conference Series in Applied Mathematics No. 12).* SIAM, 1973.

[Sch97] Lawrence S. Schulman. *Time's Arrows and Quantum Measurement.* Cambridge University Press, 1997.

[Sch02] Tamar Schlick. *Molecular Modeling and Simulation.* Springer, 2002.

[SD99] Celeste Sagui and Thomas A. Darden. Molecular dynamics simulations of biomolecules: long-range electrostatic effects. *Annu. Rev. Biophys. Biomol. Struct.*, 28:155–179, 1999.

[SD01] Celeste Sagui and Thomas A. Darden. Multigrid methods for classical molecular dynamics simulations of biomolecules. *J. Chem. Phys.*, 114(15):6578–6591, 2001.

[SEK+05] Daniel Sjöberg, Christian Engstrom, Gerhard Kristensson, David J. N. Wall, and Niklas Wellander. A floquet–bloch decomposition of maxwell's equations applied to homogenization. *Multiscale Modeling & Simulation*, 4(1):149–171, 2005.

[Sel84] Siegfried Selberherr. *Analysis and Simulation of Semiconductor Devices.* Springer-Verlag, 1984.

[SF73] G. Strang and G.J. Fix. *An Analysis of the Finite Element Method.* Prentice-Hall, Englewood Cliffs, 1973.

[SF90] P.P. Silvester and R.L. Ferrari. *Finite Elements for Electrical Engineers.* Cambridge ; New York : Cambridge University Press, 2nd ed., 1990.

[SH90] Kim A. Sharp and Barry Honig. Calculating total electrostatic energies with the nonlinear poisson-boltzmann equation. *J. Phys. Chem.*, 94:7684–7692, 1990.

[Sha06] Vladimir M. Shalaev. Optical negative-index metamaterials. *Nature Photonics*, 1:41–48, 2006.

[She94] N. Al Shenk. Uniform error estimates for certain narrow lagrange finite elements. *Math. Comput.*, 63(207):105–119, 1994.

[Sil59] R.A. Silin. Waveguiding properties of two-dimensional periodical slow-wave systems. *Voprosy Radioelektroniki, Elektronika*, 4:11–33, 1959.

[Sil72] R.A. Silin. Optical properties of artificial dielectrics. *Izvestia VUZov Radiofizika*, 15:809–820, 1972.

[Sim03] Thomas Simonson. Electrostatics and dynamics of proteins. *Reports on Progress in Physics*, 66(5):737–787, 2003. and references therein.

[Siv57] D.V. Sivukhin. The energy of electromagnetic waves in dispersive media. *Optika i Spektroskopia*, 3:308–312, 1957.

[Sj5] Daniel Sjöberg. Homogenization of dispersive material parameters for maxwell's equations using a singular value decomposition. *Multiscale Modeling & Simulation*, 4(3):760–789, 2005.

[SJS06] Patrick Stoller, Volker Jacobsen, and Vahid Sandoghdar. Measurement of the complex dielectric constant of a single gold nanoparticle. *Optics Letters*, 31(16):2474–2476, 2006.

[SKLL93] Z.S. Sacks, D.M. Kingsland, R. Lee, and J.-F. Lee. A perfectly matched anisotropic absorber for use as an absorbing boundary condition. *IEEE Trans. on Antennas and Propag.*, 43(12):1460–1463, 1993.

[SL86] K. Stüben and J. Linden. Multigrid methods: An overview with emphasis on grid generation processes. In J. Häuser, editor, *Proc. First Internat. Conference on Numerical Grid Generations in Computational Fluid Dynamics*, Swansea, 1986. Pinerige Press.

[SL91] K. E. Schmidt and M. A. Lee. Implementing the fast multipole method in three dimensions. *J. Stat. Phys.*, 63(5/6):1223–1235, 1991.

[SL00] A.K. Soh and Z.F. Long. Development of two-dimensional elements with a central circular hole. *Comput. Methods Appl. Mech. Engrg.*, 188:431–440, 2000.

[Smi81] E. R. Smith. Electrostatic energy in ionic crystals. *Proc. Roy. Soc. London A*, 375:475–505, 1981.

[SMJ+06] D. Schurig, J.J. Mock, B. J. Justice, S. A. Cummer, J. B. Pendry, A. F. Starr, and D. R. Smith. Metamaterial Electromagnetic Cloak at Microwave Frequencies. *Science*, 314(5801):977–980, 2006.

[Smy89] William B. Smythe. *Static and Dynamic Electricity.* John Benjamins Publishing Co, 1989.

[SP96] J.S. Savage and A.F. Peterson. Higher-order vector finite elements for tetrahedral cells. *IEEE Trans. on Microwave Theory Tech.*, 44(6):874–879, 1996.

[Spa72] D.B. Spalding. A novel finite-difference formulation for differential expressions involving both first and second derivatives. *Int. J. for Numer. Meth. Eng.*, 4:551–559, 1972.

[SPV+00] D.R. Smith, Willie J. Padilla, D.C. Vier, S.C. Nemat-Nasser, and S. Schultz. Composite medium with simultaneously negative permeability and permittivity. *Phys. Rev. Lett.*, 84(18):4184–4187, May 2000.

[SR01] Rajinder Singh and Falk Riess. The 1930 Nobel Prize for physics: a close decision? *Notes and Records of the Royal Society of London*, 55:267–283, 2001.

[SS97] Kazuaki Sakoda and Hitomi Shiroma. Numerical method for localized defect modes in photonic lattices. *Phys. Rev. B*, 56(8–15):4830–4835, 1997.

[SSC94] J.M. Sanz-Serna and M.P. Calvo. *Numerical Hamiltonian Problems.* Chapman and Hall, Londonr, 1994.

[SSK05] Ilya V. Shadrivov, Andrey A. Sukhorukov, and Yuri S. Kivshar. Complete band gaps in one-dimensional left-handed periodic structures. *Phys. Rev. Lett.*, 95(19):193903, 2005.

[SSR+03] David R. Smith, David Schurig, Marshall Rosenbluth, Sheldon Schultz, S. Anantha Ramakrishna, and John B. Pendry. Limitations on subdiffraction imaging with a negative refractive index slab. *Applied Physics Letters*, 82(10):1506–1508, 2003.

[SSS01] R.A. Shelby, D.R. Smith, and S. Schultz. Experimental verification of a negative index of refraction. *Science*, 292(5514):77–79, 2001.

[ST98] I. Singer and E. Turkel. High order finite difference methods for the Helmholtz equation. *Computer Meth. in Appl. Mech. & Eng.*, 163:343–358, 1998.

[ST06] Lucía B Scaffardi and Jorge O Tocho. Size dependence of refractive index of gold nanoparticles. *Nanotechnology*, 17(5):1309–1315, 2006.

[Str41] J.A. Stratton. *Electromagnetic Theory.* McGraw-Hill: New York, 1941.

[Str69] Volker Strassen. Gaussian elimination is not optimal. *Numerische Mathematik*, 13:354–356, 1969.

[Str72] G. Strang. Variational crimes in the finite element method. In A.K.Aziz, editor, *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*, pages 689–710. Academic Press, New York, 1972.

[Str04] J.C. Strikwerda. *Finite Difference Schemes and Partial Differential Equations.* Society for Industrial and Applied Mathematics, 2004.

[Stü83] K. Stüben. Algebraic multigrid (AMG): experiences and comparisons. *Appl. Math. Comput.*, 13:419–452, 1983.

[Stü00] K. Stüben. An introduction to algebraic multigrid. In U. Trottenberg, C. W. Oosterlee, and A. Schüller, editors, *Multigrid*, pages 413–532. Academic Press, London, 2000. Appendix A.

[SU04] Gennady Shvets and Yaroslav A. Urzhumov. Engineering the electromagnetic properties of periodic nanostructures using electrostatic resonances. *Phys. Rev. Lett.*, 93(24):243902, 2004.

[Sub90] Yu. N. Subbotin. Dependence of estimates of a multidimensional piecewise polynomial approximation on the geometric characteristics of the triangulation. *Proceedings of the Steklov Institute of Mathematics*, 189(4):135–159, 1990.

[Sur87] Yu.B. Suris. On some properties of methods for numerical integration of systems $\ddot{x} = f(x)$. *USSR J. Comput. Math. and Math. Phys.*, 27:149–156, 1987.

[Sur90] Yu.B. Suris. Hamiltonian methods of Runge–Kutta type and their variational interpretation [in Russian]. *Math. Modeling*, 2:78–87, 1990.

[Sur96] Yu.B. Suris. Partitioned Runge–Kutta methods as phase-volume preserving integrators. *Phys. Lett. A*, 220:63–69, 1996.

[SV04] D. R. Smith and D. C. Vier. Design of metamaterials with negative refractive index. In M. Razeghi and G. J. Brown, editors, *Quantum Sensing and Nanophotonic Devices. Edited by Razeghi, Manijeh; Brown, Gail J. Proceedings of the SPIE, Volume 5359, pp. 52-63 (2004).*, pages 52–63, 2004.

[SvdV96] G.L.G. Sleijpen and H.A. van der Vorst. A Jacobi-Davidson iteration method for linear eigenvalue problems. *SIAM J. Matrix Anal. Appl.*, 17:401–425, 1996.

[SvdV00] Yousef Saad and Henk A. van der Vorst. Iterative solution of linear systems in the 20th century. *J. Comput. Appl. Math.*, 123(1-2):1–33, 2000.

[SW03] G.L.G. Sleijpen and F.W. Wubs. Exploiting multilevel preconditioning techniques in eigenvalue computations. *SIAM J. Sci. Comput.*, 25:1249–1272, 2003.

[SW04] R. Schuhmann and T. Weiland. Recent advances in finite integration technique for high frequency applications. In S.H.M.J. Houben W.H.A. Schilders, Jan W. ter Maten, editor, *Springer Series: Mathematics in Industry*, volume 4, pages 46–57, 2004.

[SWZ+03] K.-H. Su, Q.-H. Wei, X. Zhang, J.J. Mock, D.R. Smith, and S. Schultz. Interparticle coupling effects on plasmon resonances of nanogold particles. *Nanoletters*, 3(8):1087–1090, 2003.

[Syn28] E.H. Synge. A suggested method for extending microscopic resolution into the ultramicroscopic region. *Philosophical Magazine*, 6:356–362, 1928.

[Syn32] E.H. Synge. An application of piezoelectricity to microscopy. *Philosophical Magazine*, 13:297–300, 1932.

[Syn57] J.L. Synge. *The Hypercircle in Mathematical Physics; a Method for the Approximate Solution of Boundary Value Problems.* Cambridge: University Press, 1957.

[sZZ83] Ole Østerby & Zahari Zlatev. *Direct Methods for Sparse Matrices.* Berlin ; New York : Springer-Verlag, 1983.

[TB96] Abdulnour Y. Toukmaji and John Board. Ewald summation techniques in perspective: a survey. *Computer Physics Communications*, 95:73–92, 1996.

[TH05] Allen Taflove and Susan C. Hagness. *Computational Electrodynamics: The Finite-Difference Time-Domain Method.* Artech House Publishers, 2005.

[Tho01] Vidar Thomée. From finite differences to finite elements: A short history of numerical analysis of partial differential equations. *J. of Comp. and Applied Math.*, 128:1–54, 2001.

[Ton02] E. Tonti. Finite formulation of electromagnetic field. *IEEE Trans. Magn.*, 38(2):333–336, 2002.

[Tox94] Søren Toxvaerd. Hamiltonians for discrete dynamics. *Phys. Rev. E*, 50(3):2271–2274, Sep 1994.

[TP98] Igor Tsukerman and Alexander Plaks. Comparison of accuracy criteria for approximation of conservative fields on tetrahedra. *IEEE Trans. Magn.*, 34(5):3252–3255, 1998.

[TP99a] Igor Tsukerman and Alexander Plaks. Hierarchical basis multilevel preconditioners for 3d magnetostatic problems. *IEEE Trans. Magn.*, 35(3):1143–1146, 1999.

[TP99b] Igor Tsukerman and Alexander Plaks. Refinement strategies and approximation errors for tetrahedral elements. *IEEE Trans. Magn.*, 35(3):1342–1345, 1999.

[TPB98] Igor Tsukerman, Alexander Plaks, and H. Neal Bertram. Multigrid methods for computation of magnetostatic fields in magnetic recording problems. *J. of Applied Phys.*, 83(11):6344–6346, 1998.

[Tre85] Lloyd N. Trefethen. Three mysteries of gaussian elimination. *ACM SIGNUM Newsletter*, 20(4):2–5, 1985.

[Tre97] Lloyd Nicholas Trefethen. *Numerical Linear Algebra.* Philadelphia : Society for Industrial and Applied Mathematics, 1997.

[Tre05] S.A. Tretyakov. Research on negative refraction and backward-wave media: A historical perspective. In *Proceedings of the EPFL Latsis Symposium, Lausanne*, 2005.

[Tsu90] I.A. Tsukerman. Error estimation for finite-element solutions of the eddy currents problem. *COMPEL*, 9(2):83–98, 1990.

[Tsu94] Igor Tsukerman. Application of multilevel preconditioners to finite element magnetostatic problems. *IEEE Trans. Magn.*, 30(5):3562–3565, 1994.

[Tsu95] Igor Tsukerman. Accurate computation of ripple solutions on moving finite element meshes. *IEEE Trans. Magn.*, 31(3):1472–1475, 1995.

[Tsu98a] Igor Tsukerman. Approximation of conservative fields and the element edge shape matrix. *IEEE Trans. Magn.*, 34(5):3248–3251, 1998.

[Tsu98b] Igor Tsukerman. A general accuracy criterion for finite element approximation. *IEEE Trans. Magn.*, 34(5):2425–2428, 1998.

[Tsu98c] Igor Tsukerman. How flat are flat elements? *The International Compumag Society Newsletter*, 5(1):7–12, March 1998.

[Tsu99] Igor Tsukerman. Finite element matrices and interpolation errors. *unpublished*, March 1999.

[Tsu03] Igor Tsukerman. Symbolic algebra as a tool for understanding edge elements. *IEEE Trans. Magn.*, 39(3):1111–1114, 2003.

[Tsu04a] Igor Tsukerman. Efficient computation of long-range electromagnetic interactions without Fourier Transforms. *IEEE Trans. Magn.*, 40(4):2158–2160, 2004.

[Tsu04b] Igor Tsukerman. Flexible local approximation method for electro- and magnetostatics. *IEEE Trans. Magn.*, 40(2):941–944, 2004.

[Tsu04c] Igor Tsukerman. Toward Generalized Finite Element Difference Methods for electro- and magnetostatics. In S.H.M.J. Houben W.H.A. Schilders, Jan W. ter Maten, editor, *Springer Series: Mathematics in Industry*, volume 4, pages 58–77, 2004.

[Tsu05a] I. Tsukerman. Electromagnetic applications of a new finite-difference calculus. *IEEE Trans. Magn.*, 41(7):2206–2225, 2005.

[Tsu05b] I. Tsukerman. A new FD calculus: simple grids for complex problems. *The International Compumag Society Newsletter*, 12(2):3–17, 2005.

[Tsu06] I. Tsukerman. A class of difference schemes with flexible local approximation. *J. Comput. Phys.*, 211(2):659–699, 2006.

[Tsu07] Igor Tsukerman. Negative refraction requires strong inhomogeneity. *arXiv:0710.0011. Submitted for publication.*, September 2007.

[Tur48] A.M. Turing. Rounding-off errors in matrix processes. *Q J Mechanics Appl Math*, 1(1):287–308, 1948.

[Tv07] I. Tsukerman and F. Čajko. Photonic band structure computation using FLAME. In *Proceedings of Compumag'2007, Aachen, Germany. Submitted to* IEEE Trans. Magn., June 2007.

[TW67] W.F. Tinney and J.W. Walker. Direct solutions of sparse network equations by optimally ordered triangular factorization. *Proc. IEEE*, 55:1801–1809, 1967.

[TW05] Andrea Toselli and Olof Widlund. *Domain Decomposition Methods – Algorithms and Theory.* Springer Series in Computational Mathematics , Vol. 34. Springer, 2005.

[Twe52] Victor Twersky. Multiple scattering of radiation by an arbitrary configuration of parallel cylinders. *J. Acoust. Soc. Am.*, 24:42–46, 1952.

[Var00] Richard S. Varga. *Matrix Iterative Analysis.* Berlin; New York: Springer, 2000.

[VCK98] John L. Volakis, Arindam Chatterjee, and Leo C. Kempel. *Finite Element Method for Electromagnetics: Antennas, Microwave Circuits, and Scattering Applications.* Wiley – IEEE Press, 1998.

[VD99] L. Vardapetyan and L. Demkowicz. *hp*-adaptive finite elements in electromagnetics. *Comput. Methods Appl. Mech. Engrg.*, 169:331–344, 1999.

[vDR89] L.L. van Dommelen and E.A. Rundensteiner. Fast, adaptive summation of point forces in the two-dimensional poisson equation. *J. Comput. Phys.*, 83:126–147, 1989.

[vdV03a] Henk A. van der Vorst. *Iterative Krylov Methods for Large Linear Systems.* Cambridge, UK; New York: Cambridge University Press, 2003.

[vdV03b] Henk A. van der Vorst. *Iterative Krylov Methods for Large Linear Systems.* Cambridge University Press, 2003.

[vdV04] Henk A. van der Vorst. Modern methods for the iterative computation of eigenpairs of matrices of high dimension. *Z. Angew. Math. Mech.*, 84(7):444–451, 2004.

[Ver96] Rüdiger Verfürth. *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques.* Wiley-Teubner, Stuttgart, 1996.

[Ves68] V.G. Veselago. Electrodynamics of substances with simultaneously negative values of ϵ and μ. *Sov Phys Uspekhi*, 10(4):509–514, 1968.

[VK84] V.V. Voevodin and Iu.A. Kuznetsov. *Matritsy i Vychisleniia.* Moskva: Nauka, 1984. [in Russian].

[Vla84] V.S. Vladimirov. *Equations of Mathematical Physics.* Moscow: Mir, 1984. translated from the Russian by Eugene Yankovsky.

[VO48] E. J. W. Verwey and J. Th. G. Overbeek. *Theory of the Stability of Lyophobic Colloids.* Elsevier, Amsterdam, 1948.

[vT07] F. Čajko and I. Tsukerman. Flexible Approximation schemes for wave refraction in negative index materials. In *Proceedings of Compumag'2007, Aachen, Germany. Submitted to* IEEE Trans. Magn., June 2007.

[VWV02] P.M. Valanju, R.M. Walser, and A.P. Valanju. Wave refraction in negative-index media: Always positive and very inhomogeneous. *Phys. Rev. Lett.*, 88(18):187401, Apr 2002.

[VWV03] P.M. Valanju, R.M. Walser, and A.P. Valanju. Valanju, walser, and valanju reply:. *Phys. Rev. Lett.*, 90(2):029704, 2003.

[WAHFS01] Jr. William A. Harris, Jay P. Fillmore, and Donald R. Smith. Matrix exponentials. another approach. *SIAM Review*, 43(4):694–706, 2001.

[Wal98a] S. Waldron. The error in linear interpolation at the vertices of a simplex. *SIAM J. Numer. Analysis*, 35(3):1191–1200, 1998.

[Wal98b] Wolfgang Walter. *Ordinary Differential Equations.* New York: Springer, 1998.

[WAM06] Mark S. Wheeler, J. Stewart Aitchison, and Mohammad Mojahedi. Coated nonmagnetic spheres with a negative index of refraction at infrared frequencies. *Physical Review B*, 73(4):045105, 2006.

[Was53] K. Washizu. Bounds for solutions of boundary value problems in elasticity. *J. Math. Phys.*, 32:117–128, 1953.

[Was03] Takumi Washio. private communication, 2002-2003.

[WB00] A. Wiegmann and K.P. Bube. The explicit-jump immersed interface method: Finite difference methods for PDEs with piecewise smooth solutions. *SIAM J. Numer. Analysis*, 37(3):827–862, 2000.

[WD00] T. Wriedt and A. Doicu. T-matrix method for light scattering from a particle on or near an infinite surface. In F. Moreno and F. González, editors, *Springer Lecture Notes in Physics*, volume 534, pages 113–132. Springer-Verlag, 2000.

[Web93] J.P. Webb. Edge elements and what they can do for you. *IEEE Trans. Magn.*, 29(2):1460–1465, 1993.

[Web99] J.P. Webb. Hierarchal vector basis functions of arbitrary order for triangular and tetrahedral finite elements. *IEEE Trans. on Antennas & Propagation*, 47(8):1244–1253, 1999.

[Web02] J.P. Webb. P-adaptive methods for electromagnetic wave problems using hierarchal tetrahedral edge elements. *Electromagnetics*, 22(5):443–451, 2002.

[Web05] J.P. Webb. Using adjoint solutions to estimate errors in global quantities. *IEEE Trans. Magn.*, 41(5):1728–1731, 2005.

[Web07] Jon Webb. Private communication, 2004–2007.

[Wei74] Robert Weinstock. *Calculus of Variations: with applications to physics and engineeri.* New York, Dover Publications, 1974.

[Wes85] J.E. Wessel. Surface-enhanced optical microscopy. *J. Opt. Soc. Am. B*, 2:1538–1541, 1985.

[Wes91] Pieter Wesseling. *An Introduction to Multigrid Methods.* Chichester [England] ; New York : J. Wiley, 1991.

[WF93] J.P. Webb and B. Forghani. Hierarchical scalar and vector tetrahedra. *IEEE Trans. Magn.*, 29(2):1495–1498, 1993.

[WH01] Zuowei Wang and Christian Holm. Estimate of the cutoff errors in the ewald summation for dipolar systems. *J. of Chem. Phys.*, 115(14):6351–6359, 2001.

[Whi57] H. Whitney. *Geometric Integration Theory.* Princeton, NJ: Princeton Univ. Press, 1957.

[Wil61] James Hardy Wilkinson. Error analysis of direct methods of matrix inversion. *J. of the ACM*, 8(3):281–330, 1961.

[Wil65] James Hardy Wilkinson. *The Algebraic Eigenvalue Problem.* Oxford ; New York : Oxford University Press, 1988 (c1965).

[Wil94] James Hardy Wilkinson. *Rounding Errors in Algebraic Processes.* Dover Publications (Reprint edition), 1994.

[Wil01] John Michael Williams. Some problems with negative refraction. *Phys. Rev. Lett.*, 87(24):249703, Nov 2001.

[WKLL97] Jo-Yu Wu, D.M. Kingsland, Jin-Fa Lee, and R. Lee. A comparison of anisotropic pml to berenger's pml and its application to the finite-element method for em scattering. *IEEE Trans. on Antennas & Propagation*, 45(1):40–50, 1997.

[WOL75] J.H. Weaver, C.G. Olson, and D.W. Lynch. Optical properties of crystalline tungsten. *Phys. Rev. B*, 12:1293–97, 1975.

[WPL02] Z.J. Wang, A.J. Przekwas, and Yen Liu. A fv-td electromagnetic solver using adaptive cartesian grids. *Computer Physics Communications*, 148:17–29, 2002.

[Wri99] Thomas Wriedt, editor. *Generalized Multipole Techniques for Electromagnetic and Light Scattering.* Amsterdam ; London : Elsevier, 1999.

[Xu89] J. Xu. *Theory of Multilevel Methods.* PhD thesis, Cornell University, 1989.

[Xu92] Jinchao Xu. Iterative methods by space decomposition and subspace correction. *SIAM Review*, 34:581–613, 1992.

[Xu97] Jinchao Xu. An introduction to multigrid convergence theory. In Tony F. Chan Raymond H. Chan and Gene H. Golub, editors, *Iterative Methods in Scientific Computing*, pages 169–242, Singapore ; New York, 1997. Springer.

[XZ03] Jinchao Xu and Ludmil Zikatanov. Some observations on babuška and brezzi theories. *Numerische Mathematik*, 94(1):195–202, March 2003.

[Yab87] Eli Yablonovitch. Inhibited spontaneous emission in solid-state physics and electronics. *Phys. Rev. Lett.*, 58(20):2059–2062, May 1987.

[Yas06] Kiyotoshi Yasumoto, editor. *Electromagnetic theory and applications for photonic crystals.* Boca Raton, FL : CRC/Taylor & Francis, 2006.

[YC97] K. S. Yee and J.S. Chen. The finite-difference time-domain (FDTD) and the finite-volume time-domain (FVTD) methods in solving maxwell's equations. *IEEE Trans. Antennas Prop.*, 45(3):354–363, 1997.

[Yee66] K. S. Yee. Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media. *IEEE Trans. Antennas & Prop.*, AP-14(3):302–307, 1966.

[Yeh79] Pochi Yeh. Electromagnetic propagation in birefringent layered media. *J. Opt. Soc. Am.*, 69(5):742–756, 1979.

[Yeh05] Pochi Yeh. *Optical Waves in Layered Media.* Hoboken, N.J.: John Wiley, 2005.

[YF04] B. Yellen and G. Friedman. Programmable assembly of heterogeneous colloidal particle arrays. *Adv. Mater.*, 16(2):111–115, 2004.

[YFB04] B.B. Yellen, G. Friedman, and K.A. Barbee. Programmable self-aligning ferrofluid masks for lithographic applications. *IEEE Trans Magn.*, 40(4):2994–2996, 2004.

[YG89] E. Yablonovitch and T.J. Gmitter. Photonic band structure: The face-centered-cubic case. *Phys. Rev. Lett.*, 63(18):1950–1953, Oct 1989.

[YM01] W. Yu and R. Mittra. A conformal finite difference time domain technique for modeling curved dielectric surfaces. *IEEE Microwave Wireless Comp. Lett.*, 11:25–27, 2001.

[YM05] Vassilios Yannopapas and Alexander Moroz. Negative refractive index metamaterials from inherently non-magnetic materials for deep infrared to terahertz frequency ranges. *J. of Physics: Condensed Matter*, 17(25):3717–3734, 2005.

[You03] David M. Young, editor. *Iterative Solution of Large Linear Systems.* Dover Publications, 2003.

[Yse86] H. Yserentant. On the multilevel splitting of finite-element spaces. *Numerische Mathematik*, 49(4):379–412, 1986.

[YT97] T.V. Yioultsis and T.D. Tsiboukis. Development and implementation of second and third order vector finite elements in various 3-d electromagnetic field problems. *IEEE Trans Magn.*, 33(2):1812–1815, 1997.

[YWK+02] S. Yamada, Y. Watanabe, Y. Katayama, X. Y. Yan, and J. B. Cole. Simulation of light propagation in two-dimensional photonic crystals with a point defect by a high-accuracy finite-difference time-domain method. *J. of Applied Phys.*, 92(3):1181–1184, 2002.

[YY94] Darrin M. York and Weitao Yang. The fast Fourier Poisson method for calculating Ewald sums. *J. Chem. Phys.*, 101:3298–3300, 1994.

[ZC00] J.S. Zhao and W.C. Chew. Integral equation solution of Maxwell's equations from zero frequency to microwave frequencies. *IEEE Trans. Antennas & Prop.*, 48(10):1635–1645, 2000.

[Zen87] R. Zengerle. Light propagation in singly and doubly periodic planar waveguides. *J. Mod. Optics*, 34:1589–1617, 1987.

[ZFM+05] Shuang Zhang, Wenjun Fan, K.J. Malloy, S.R. Brueck, N.C. Panoiu, and R.M. Osgood. Near-infrared double negative metamaterials. *Optics Express*, 13(13):4922–4930, 2005.

[ZFM+06] Shuang Zhang, Wenjun Fan, Kevin J. Malloy, Steven R.J. Brueck, Nicolae C. Panoiu, and Richard M. Osgood. Demonstration of metal-dielectric negative-index metamaterials with improved performance at optical frequencies. *J. Opt. Soc. Am. B*, 23(3):434–438, 2006.

[ZFP+05] Shuang Zhang, Wenjun Fan, N. C. Panoiu, K. J. Malloy, R. M. Osgood, and S. R. J. Brueck. Experimental demonstration of near-infrared negative-index metamaterials. *Phys. Rev. Lett.*, 95(13):137404, 2005.

[Zha95] S.Y. Zhang. Successive subdivisions of tetrahedra and multigrid methods on tetrahedral meshes. *Houston J. of Math.*, 21(3):541–556, 1995.

[Zl8] M. Zlámal. On the finite element method. *Numer. Math.*, 12:394–409, 1968.

[ZMC+97] F.-X. Zgainski, Y. Maréchal, J.-L. Coulomb, M.G. Vanti, and A. Raizer. An a priori indicator of finite element quality based on the condition number of the stiffness matrix. *IEEE Trans. Magn.*, 33(2):1748–1751, 1997.

[ZSW03] I.A. Zagorodnov, R. Schuhmann, and T. Weiland. A uniformly stable conformal FDTD-method in Cartesian grids. *Int. J. Numer. Model.*, 16:127–141, 2003.

[ZT05] O.C. Zienkiewicz and R.L. Taylor. *The Finite Element Method for Solid And Structural Mechanics.* Oxford ; Burlington, Mass. : Elsevier Butterworth-Heinemann, 2005.

[ZTZ05] O.C. Zienkiewicz, R.L. Taylor, and J.Z. Zhu. *The Finite Element Method : Its Basis and Fundamentals.* Oxford ; Boston : Elsevier Butterworth-Heinemann, 2005.

[ZZ87] O.C. Zienkiewicz and J.Z. Zhu. A simple error estimator and adaptive procedure for practical engineerng analysis. *Int. J. for Numer. Meth. Eng.*, 24(2):337–357, 1987.

[ZZ92a] O.C. Zienkiewicz and J.Z. Zhu. The superconvergent patch recovery and a-posteriori error-estimates. 1. the recovery technique. *Int. J. for Numer. Meth. Eng.*, 33(7):1331–1364, 1992.

[ZZ92b] O.C. Zienkiewicz and J.Z. Zhu. The superconvergent patch recovery and a-posteriori error-estimates. 2. error estimates and adaptivity. *Int. J. for Numer. Meth. Eng.*, 33(7):1365–1382, 1992.

Index

A

A posteriori error estimates, 151
Adams methods, 21, 24, 30
Adaptive mesh refinement, 100
Adaptive refinement, 147, 148, 300, 302, 310
Algebraic multigrid (AMG) schemes, 158
Apertureless SNOM, 438, 441
Aperture-limited SNOM, 437, 438
Approximation
  accuracy, 158–160
  analytical, 325
  and condition number, 179, 180
  finite element, 127, 128
  flexibility vs. conformity, 193–195
  local, 89–91
  Taylor, 42, 162–164, 166, 170, 192, 193
Approximation accuracy, and element shape, 158–160
Atomic Force Microscopy (AFM), 436

B

Backward Differentiation Formulae (BDF), 29, 30
Backward waves, 10, 389
  and group velocity, 358, 378
  historical notes, 446–450
  homogeneous isotropic medium, 456, 458, 459
  in Mandelshtam's chain of oscillators, 459–464
  in photonic crystals, 465–470
  and volume grating, 376
  weakly inhomogeneous regime, 474–476
Bessel functions, spherical, 195, 285, 315, 322, 323
Bloch Transform, 477, 478
Bloch-Floquet waves, 473
  in 2D and 3D, 386–389, 456–459
  and Bloch transform, 477
  and electromagnetic analysis of periodic structures, 364
  and energy velocity, 382, 383
  and FLAME, 405, 410
  and Fourier analysis, 375–378
  Fourier harmonics of, 379
  and group velocity, 380–382, 460
  problem, 461–464
Boundary value problems
  1D, 39–45
  2D, 47–51
  and FEM, 283, 284
  implementation of, 46, 47
  singular, 213
Bravais lattice, 473
B-splines, cardinal, 267–270
Butcher tableau, 21, 38

C

Casimir forces, 325
Céa's theorem, 86–89, 126
Charge-to-grid interpolation, 262, 264
Cholesky decomposition, 113, 134, 397, 400

Cholesky factorization, 113–116, 134
Ciarlet-Raviart theory, 125
Cloaking, 3, 488
Collatz Mehrstellen Schemes
  in 2D, 51–54
  in 3D, 58–61
Collocation
  explained, 71–75
  and Multiple Multipole Method (MMP), 425
  shortcomings of, 89
  spline-based, 213, 214
Colloidal simulation, 4, 5
  electrostatics of, 228
Commercial software, examples of, 7
Computational analysis, 4, 5
Computer simulation, 6, 7
Condition number and approximation accuracy, 179, 180
Continuum models, 4
Correlations and mean-field theory, 313, 314
Crank-Nicolson scheme, 16–18, 27, 28, 33
Cylindrical harmonics, 292, 401
  applications in FLAME, 229, 230
  FLAME basis functions, 286, 287, 427
  in particle problem, 190
  Trefftz basis functions, 428
  used for verification and error analysis, 217, 218

D

Dark-field microscopy, 439–441
Debye length, 314, 332, 335, 336
Debye-Hückel parameter, 312, 324
Depine-Lakhtakia condition, 459, 460
Derjaguin-Landau-Verwey-Overbeek theory (DLVO)
  comparison with FLAME, 332–336
  treatment of electrostatic interactions in, 321–324
Dielectric circular cylinder, 217
Dielectric constant. *See* permittivity
Difference schemes
  consistency and convergence of, 59–63
  stability, 55, 64, 231
  by variational homogenization, 222
Diffraction limit, 434, 435
Dirac deltas, 71, 72
Direct solvers, 51, 113, 137, 138, 481, 486
Direct sum, 253, 255, 256
Dirichlet boundary conditions, 46, 47, 50, 78, 83, 84, 95
Discrete-Dipole method, 425, 426
Distributions. *See* generalized functions
Divergence and curl operators, 186, 187
Domain boundary conditions, 202, 203
Domain decomposition, 11, 51, 226, 445
Double-slit experiments, 2
Drude model, 418, 423

E

Edge elements
  and electromagnetic problems, 139–145
  historical notes, 146, 147
  tetrahedral, 146–148, 161, 431
Edge shape matrix, 173
Eigenvalue analysis, 161, 480
Eigenvalue solvers, 478–486
Electostatics of macromolecules, 228
Electromagnetic wave scattering, 300
Electrostatic energy DLVO theory, 321
Electrostatic forces
  computing, 329, 332
  formulas for, 276, 277
  grid-to-grid interpolation, 275
  long-range, 54
  relative rms error in, 295
Element shape
  and approximation accuracy, 158–160
  Jamet condition, 160
  maximum eigenvalue condition, 163
  most common, 105
  singular value condition, 171, 173, 177, 178
Energy density in dispersive media, 359
Energy velocity, 358, 359, 382
Entropy, 331, 338
Error of solution by collocation, 74
Euler scheme, 15–18, 26
  stability region, 27
Ewald formulas, 252–254

Ewald methods
grid-based, 256–262
Particle-Particle Particle-Mesh, 269–271
Ewald sum, conditional convergence, 245
Ewald summation
derivation of, 249–252
introduction to, 242–246
particle-mesh, 264–266
role of parameters in, 254–256
Exact finite-difference schemes, 227
Explicit models, 4, 5
Explicit schemes, 28, 29

F

Fast Multipole Method (FMM)
advantages of, 196, 197
alternative to Ewald techniques, 282, 283
and long-range particle interactions, 241, 242
for numerical simulation, 314, 315
FE-Galerkin formulation, 93, 94
Fermi velocity, 423
Field enhancement particle cascades, 431
Finite Difference (FD) analysis, 11
example of, 12–15
Finite difference schemes
derivation using constraints, 193, 199
implementation of, 40
Finite Difference Time Domain (FDTD) methods, 223, 224
Finite element analysis, 148, 389, 401, 451, 486
Finite element approximation, 127, 128
Finite element approximation vs. interpolation, 109, 124, 159–161, 163, 177, 180
Finite element matrices, sparsity, 94, 113–116
Finite element mesh, 190
in 2D, 90
in 3D, 92
Finite Element Method (FEM), 11, 69, 70
convergence of, 159, 177
posteriori error estimates, 148, 151
priori error estimates, 184
Finite-difference time-domain methods (FDTD), 12, 65, 233, 405, 430, 465
First-order elements
in 1D, 91–100
in 2D, 105–120
FLAME
adaptive, 300, 334, 402
band structure computation, 405–411
numerical bases, 297, 303
particles in dielectrics, 307
particles in solvents, 315–318
photonic crystals, 231, 401, 405
plasmonic particles, 426–429
Flexible Local Approximation MEthods. *See* FLAME
Flux-balance scheme
in 2D, 48–50
in 3D, 56, 57
errors, 45
explained, 42–44, 46
Formulations, of nanoscale problems, 3
Fourth Order 19-point Mehrstellen Scheme, 210
Fourth Order 9-point Mehrstellen Scheme, 209
Fréchet derivative, 237, 238
Free energy, 328–332, 340, 341
Helmholtz, 338
minimize total, 275

G

Galerkin method
discontinous, 222, 223
and eigenvalue problem, 482
and GFEM, 183, 220
test functions, 323
and weak formulation, 74–82
Galerkin solution, 78, 81–83, 86–89, 124
Gaussian elimination, 113, 130, 131
Gaussian factorization, 131
Gauss's Theorem, 48, 49, 56
Generalized curl, 143, 186, 187
Generalized divergence, 144, 186, 187, 346, 348
Generalized Finite Element Method (GFEM), 183, 191, 199
adaptive, 196

algorithmic complexity of, 220
described, 181–183
for numerical simulation, 314, 315
by partition of unity, 221, 222
performance, 194
Generalized functions (distributions), 186, 343–348, 430
Global approximation error, 183
Goal-oriented error estimation, 153, 154
Gouy-Chapman length, 311
Grid-to-charge interpolation, 157, 262, 269, 275
Group velocity, 157, 262, 269, 275

H
Hamiltonian systems, 33–38, 64
Harmonic oscillator, motion of, 36
Hat function, 92, 93
Heisenberg principle, 434, 435
Helmholtz equation, 388
1D, 206
2D, 384
and FD schemes, 227
solution, 428
Helmholtz free energy, 331, 338, 339, 341
Hertzian dipole, 433
Hierarchial bases
edge elements, 150
nodal elements, 150
Higher-order elements, 102, 104, 105
Hill's equation, 360, 363, 366
Homgenization, local in FDTD, 223, 224

I
Immersed surface methodology, 227, 228
Implicit models, 4
Implicit schemes, 29, 38
Integral conservation principle, 49
Integral equation methods, 283, 422, 425, 426, 428
Iterative solvers, 51, 138, 148, 149, 157, 483

J
Jamet's condition, 160, 173

K
Kohn-Sham equation, 5, 210
Kohn-Sham equation. *See* Schrödinger equation
Korringa-Kohn-Rostoker (KKR) method, 393
Kronecker-delta property, 104, 122

L
Ladyzhenskaya-Babuska-Brezzi (LBB) Condition, 89, 124, 126, 127
Lagrange multipliers, 53, 54
Laplace equation, 53, 54, 190, 198
2D and 3D, 208–210
Lattice cell size, 389, 471, 474, 476
Lax-Milgram theorem, 86–89, 126
LDU factorization, 133
Leapfrog scheme, 38
Lennard-Jones potential, 325
Linear multistep schemes, 24–27
Lorentz model, 418

M
Macromolecular simulation, 4
Mandelshtam's chain, 459, 460, 462–464
Mass matrix, 111
Matlab code, 95, 100–104, 114, 117–120
Matrix exponential, 65–67
Matrix sparsity structure, 129, 130
Maxwell stress tensor, 289, 326, 332
Maxwell's equations
and Bloch wave, 379, 382
and homogenization limit, 474
material parameters for, 423
modern vector form of, 343
and negative refraction, 459
phenomenological stochastic terms in, 325, 326
and plane wave, 457
review of, 349–353
time-harmonic, 384, 411–414
Mehrstellen (Mehrstellenverfahren) schemes, 40, 42, 54
Mesh generation, 98
Mesh quality, 179
Mesh-based methods, 6, 7
Meshless methods, 220, 224, 225
Metamaterials, 3, 389, 467, 471, 472, 474, 488

Mie theory, 401, 431
Minimum degree reordering, 116
Modeling errors, 6
Momentum conservation, in grid-based Ewald methods, 261
Multigrid methods
  and adaptive mesh refinement, 148–154
  algorithms, 154–156
  convergence of, 157, 158
Multiple Multipole Method (MMP), 89, 425, 445
Multiscale modeling, 3, 4
Multivalued approximation, 190, 193, 199, 224, 232

N

Nano-focusing, 418, 430, 434
Nano-lens, 433, 434
Negative refraction. *See also* backward waves
  historical notes, 446–451
  intrinsic and extrinsic conditions, 467
  metamaterials vs. photonic crystals, 465, 467, 468, 471
  negative-index materials, 451, 473
Nested Dissection (ND), 136–138
Neumann boundary condition, 86
Newton-Raphson method, 31, 204, 205, 319
Newton-Raphson-Kantorovich method. *See* Newton-Raphson method
Numerical errors for different one-step schemes, 19
Numerical solutions
  for different one-step schemes, 18, 20
  for the forward Euler scheme, 19

O

One-Way Dissection (1WD), 136, 137
Optical tips. *See* Scanning Near-Field Optical Microscopy (SNOM)

P

Particle-Mesh Ewald methods, 265
Particle-Particle Particle-Mesh Ewald Methods, 262, 269–271
Partition function-sum over states, 341
Patches (subdomains)
  FLAME, 232, 291, 300, 303
  GFEM and FLAME, 199
Perfect lens, 435, 451–456, 487
Permittivity
  bulk, 423
  changing, 376, 377
  complex, 428
  complex dielectric, 328
  constant, 322, 420
  dielectric, 42, 69, 78, 84, 117, 237, 291, 324, 352, 388, 389, 391
  inverse, 392
  nanoscale vs. bulk value, 423
  negative, 417–422
  relative, 300, 430, 452, 453
  relative dielectric, 386
  varying, 442
Phase velocity, 471
  and Bloch wave, 379, 461, 462, 464, 470
  and group velocity, 474
  value of, 463
  and wave form, 354, 355, 458, 459
  and wave propagation, 449, 450, 467
Photonic band structure
  FEM, 393–397, 399, 415
  FLAME, 405–411
  PWE, 375–391, 393, 397, 401, 405, 409, 410, 486
Photonic crystal, 2, 3
Photonic crystal waveguide, 402, 405
Photonic crystals, band structure, 230, 349, 399, 410
Phototonic bandgap, 373, 374
  in 2D, 386–389, 393
  in 3D, 411–415
  computation, 486
  historical notes, 416
  layered structure, 368
  problems, 481, 482
Plane-wave expansion (PWE), 10
  Fourier approximation, 391–400
  role of polarization, 390, 391
  solution by, 389, 390
Plasma frequency, 417, 418, 420
Plasmon resonances
  and diffraction, 434
  electrostatic approximation, 421–423

and FLAME, 426
and Multiple Multipole Method (MMP), 425
phenomena, 420, 421
and SNOM, 444
Poisson equation
in 1D, 39–41, 60, 61, 127, 128
in 2D, 51, 52
in 3D, 55
and boundary conditions, 85
for cloud potential, 272
and collodial simulation, 5
and Ewald formulae, 252, 253
and Finite Element Method, 69–71, 73–76
in Fourier space, 259, 270, 275
Taylor-based difference scheme, 47, 48
Poisson-Boltzmann equation (PBE). *See* Poisson-Boltzmann model
Poisson-Boltzmann model
in colloidal simulation, 196, 197, 205, 210
introduction to, 309–313
limitations of, 313, 314
linearized, 322
Newton-Raphson-Kantorovich procedure, 319
operator, 237
problem, 317, 318
solution of, 331, 332, 339, 340
Polynomial interpolation, 24
Poynting vector
and Bloch wave, 474, 475
cell-averaged, 473
characterization, 467
and Fourier harmonics, 380–384
and group velocity, 358–360, 463, 464
mechanical, 460
and phase velocity, 459, 461
and wave vector, 447–449, 462
Pseudospectral methods (PSM), 191, 219, 226, 227

Q

Quotient Minimum Degree (QMD) method, 136

R

Raman spectroscopy, 438, 439
Reciprocal lattice, 242, 243
Reciprocal space, 249, 251, 263, 272, 275, 414
Reciprocal sum, 255, 256
Recovery-based error estimators, 151–153
R-K method. *See* Runge-Kutta (R-K) method
Robin boundary conditions, 85
Runge-Kutta (R-K) method, 20–23, 37, 38
Rytov–Lifshitz theory, 325, 326

S

Scanning Near-Field Optical Microscopy (SNOM)
apertureless, 438–441
aperture-limited type, 437, 438
and diffraction limit, 433–436
electrostatic approximation in, 441–445
infrared, 349, 436–439, 441, 446
scattering-type, 445, 446
two main approaches, 436–439
Scanning Probe Microscopy (SPM), 436
Scanning Tunneling Microscope (STM), 436
Scattering SNOM, 445, 446
1D schemes, 46, 47
3D schemes, 57
Schrödinger equation
1D, 197, 210–213, 221
in band theory of solids, 393
Second-order elements, 102, 104, 117, 121
Simulation model, 3
Single velocity, 355, 357, 381, 382
Singular value condition
and eigenvalue problem, 160
geometric implications of, 171–173
minimum, 158, 159, 173–178
Smooth Particle-Mesh Ewald Methods, 267–269
Solution by collocation, 73
Special approximation techniques, 227, 228
Spectral convergence, 139, 142, 145

Spherical harmonics, 322
and electrostatic potential, 419
and FLAME, 285, 302, 305, 307, 315, 321
in flexible approximation, 195, 196, 228
integration of, 323, 324
and Mie theory, 401
and T-matrix methods, 420
Trefftz-FLAME basis functions, 303
vector, 424, 425
Spurious modes, and edge elements, 139, 142, 145, 184
Stability classification, 31, 32
Stiff systems, 27–29, 33, 34
Stiff-stability region, 33
Störmer-Verlet method, 38
Subwavelength structure, 2, 3
Sum over states. *See* partition function-sum over states
Suris-Sanz-Serna condition, 38
Symmetric positive definite (SPD) systems, 133–135
Symplecticness, 5, 33, 38, 39
Synge-Babuška-Aziz maximum angle condition, 161, 170, 177, 180, 181

T
Taylor approximation, 42, 162–164, 166, 170, 192, 193
Taylor derivation, 39, 40
Taylor expansion, 39–42, 47, 48, 50
in 3D, 55
Taylor-based schemes, 17
Tetrahedral elements, 122, 160, 167, 168
Theory of homogenization, 225, 226
Thermodynamic potential, 328–332
for electrostatics in solvents, 338–343
Three-point flux balance scheme, 44
T-matrix method, 401, 422, 424, 425
Trapezoidal scheme. *See* Crank-Nicolson scheme
Trefftz-FLAME, 190, 197, 198, 200
case-studies, 202–211
simulation, 426–428
Triangular mesh example, 106

U
Uncertainty principle, 326, 434–436

V
Van der Waals forces, 324–326, 436
Variational FLAME, 231–236
Variational methods, 6
Vector and matrix norms, 65
Veselago medium, 456, 457, 474

W
Wave analysis, 443
Wavefunction, 2
Whitney-Nédélec elements, 142. *See also* edge elements
Wigner-Seitz cell, 243

Y
Yee scheme, 223
York–Yang method, 271, 272, 274, 487
Yukawa potential, 313, 324
and FLAME basis functions, 321
and linearized PBE, 317
superposition of, 314, 332, 336

Z
Zlámal's minimum angle condition, 159, 167, 180, 181

Nanostructure Science and Technology

(Continued from p. ii)

Current volumes in this series:

Nanoparticles: Building Blocks for Nanotechnology
Edited by Vincent Rotello

Nanostructured Catalysts
Edited by Susannah L. Scott, Cathleen M. Crudden, and Christopher W. Jones

Self-Assembled Nanostructures
Jin Z. Zhang, Zhong-lin Wang, Jun Liu, Shaowei Chen, and Gang-yu Liu

Polyoxometalate Chemistry for Nano-Composite Design
Edited by Toshihiro Yamase and M.T. Pope

Computational Methods for Nanoscale Applications: Particles, Plasmons and Waves
Igor Tsukerman