Finite Element Analysis of Acoustic Scattering

Frank Ihlenburg

Springer

Applied Mathematical Sciences
Volume 132

Springer
New York
Berlin
Heidelberg
Barcelona
Budapest
Hong Kong
London
Milan
Paris
Singapore
Tokyo

Applied Mathematical Sciences

1. *John:* Partial Differential Equations, 4th ed.
2. *Sirovich:* Techniques of Asymptotic Analysis.
3. *Hale:* Theory of Functional Differential Equations, 2nd ed.
4. *Percus:* Combinatorial Methods.
5. *von Mises/Friedrichs:* Fluid Dynamics.
6. *Freiberger/Grenander:* A Short Course in Computational Probability and Statistics.
7. *Pipkin:* Lectures on Viscoelasticity Theory.
8. *Giacoglia:* Perturbation Methods in Non-linear Systems.
9. *Friedrichs:* Spectral Theory of Operators in Hilbert Space.
10. *Stroud:* Numerical Quadrature and Solution of Ordinary Differential Equations.
11. *Wolovich:* Linear Multivariable Systems.
12. *Berkovitz:* Optimal Control Theory.
13. *Bluman/Cole:* Similarity Methods for Differential Equations.
14. *Yoshizawa:* Stability Theory and the Existence of Periodic Solution and Almost Periodic Solutions.
15. *Braun:* Differential Equations and Their Applications, 3rd ed.
16. *Lefschetz:* Applications of Algebraic Topology.
17. *Collatz/Wetterling:* Optimization Problems.
18. *Grenander:* Pattern Synthesis: Lectures in Pattern Theory, Vol. I.
19. *Marsden/McCracken:* Hopf Bifurcation and Its Applications.
20. *Driver:* Ordinary and Delay Differential Equations.
21. *Courant/Friedrichs:* Supersonic Flow and Shock Waves.
22. *Rouche/Habets/Laloy:* Stability Theory by Liapunov's Direct Method.
23. *Lamperti:* Stochastic Processes: A Survey of the Mathematical Theory.
24. *Grenander:* Pattern Analysis: Lectures in Pattern Theory, Vol. II.
25. *Davies:* Integral Transforms and Their Applications, 2nd ed.
26. *Kushner/Clark:* Stochastic Approximation Methods for Constrained and Unconstrained Systems.
27. *de Boor:* A Practical Guide to Splines.
28. *Keilson:* Markov Chain Models—Rarity and Exponentiality.
29. *de Veubeke:* A Course in Elasticity.
30. *Shiatycki:* Geometric Quantization and Quantum Mechanics.
31. *Reid:* Sturmian Theory for Ordinary Differential Equations.
32. *Meis/Markowitz:* Numerical Solution of Partial Differential Equations.
33. *Grenander:* Regular Structures: Lectures in Pattern Theory, Vol. III.
34. *Kevorkian/Cole:* Perturbation Methods in Applied Mathematics.
35. *Carr:* Applications of Centre Manifold Theory.
36. *Bengtsson/Ghil/Källén:* Dynamic Meteorology: Data Assimilation Methods.
37. *Saperstone:* Semidynamical Systems in Infinite Dimensional Spaces.
38. *Lichtenberg/Lieberman:* Regular and Chaotic Dynamics, 2nd ed.
39. *Piccini/Stampacchia/Vidossich:* Ordinary Differential Equations in $\mathbf{R}^n$.
40. *Naylor/Sell:* Linear Operator Theory in Engineering and Science.
41. *Sparrow:* The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors.
42. *Guckenheimer/Holmes:* Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields.
43. *Ockendon/Taylor:* Inviscid Fluid Flows.
44. *Pazy:* Semigroups of Linear Operators and Applications to Partial Differential Equations.
45. *Glashoff/Gustafson:* Linear Operations and Approximation: An Introduction to the Theoretical Analysis and Numerical Treatment of Semi-Infinite Programs.
46. *Wilcox:* Scattering Theory for Diffraction Gratings.
47. *Hale et al:* An Introduction to Infinite Dimensional Dynamical Systems—Geometric Theory.
48. *Murray:* Asymptotic Analysis.
49. *Ladyzhenskaya:* The Boundary-Value Problems of Mathematical Physics.
50. *Wilcox:* Sound Propagation in Stratified Fluids.
51. *Golubitsky/Schaeffer:* Bifurcation and Groups in Bifurcation Theory, Vol. I.
52. *Chipot:* Variational Inequalities and Flow in Porous Media.
53. *Majda:* Compressible Fluid Flow and System of Conservation Laws in Several Space Variables.
54. *Wasow:* Linear Turning Point Theory.
55. *Yosida:* Operational Calculus: A Theory of Hyperfunctions.
56. *Chang/Howes:* Nonlinear Singular Perturbation Phenomena: Theory and Applications.
57. *Reinhardt:* Analysis of Approximation Methods for Differential and Integral Equations.
58. *Dwoyer/Hussaini/Voigt (eds):* Theoretical Approaches to Turbulence.
59. *Sanders/Verhulst:* Averaging Methods in Nonlinear Dynamical Systems.
60. *Ghil/Childress:* Topics in Geophysical Dynamics: Atmospheric Dynamics, Dynamo Theory and Climate Dynamics.

(continued following index)

Frank Ihlenburg

Finite Element Analysis of Acoustic Scattering

With 88 Illustrations

Springer

Frank Ihlenburg
Germanischer Lloyd
Vorsetzen 32
D-20459 Hamburg
Germany

Editors

J.E. Marsden
Control and Dynamical Systems, 107-81
California Institute of Technology
Pasadena, CA 91125
USA

L. Sirovich
Division of Applied Mathematics
Brown University
Providence, RI 02912
USA

Mathematics Subject Classification (1991): 65Nxx, 65Lxx, 76M10, 73Vxx

Library of Congress Cataloging-in-Publication Data
Ihlenburg, Frank.
Finite element analysis of acoustic scattering / Frank Ihlenburg.
p. cm. – (Applied mathematical sciences ; 132)
Includes index.
ISBN 0-387-98319-8 (alk. paper)
1. Wave equations. 2. Scattering (Physics)—Mathematical models. 3. Helmholtz equation—Numerical solutions. 4. Boundary value problems—Numerical solutions. 5. Finite element method. I. Title. II. Series: Applied mathematical sciences (Springer-Verlag New York Inc.) ; v. 132.
QA1.A647 vol. 132
[QC174.26.W28]
510 s—dc21
[531'.1133'01515353] 97-49493

ISBN 0-387-98319-8 Springer-Verlag New York Berlin Heidelberg SPIN 10636277

To Krystyna and Katja

Love's not Time's fool
— William Shakespeare, Sonnett 116

This page intentionally left blank

Preface

> Als überragende Gestalt ... tritt uns HELMHOLTZ entgegen ... Seine außerordentliche Stellung in der Geschichte der Naturwissenschaften beruht auf einer ungewöhnlich vielseitigen, eindringenden Begabung, innerhalb deren die mathematische Seite eine wichtige, für uns natürlich in erster Linie in Betracht kommende Rolle spielt. (Felix Klein, [84, p. 223])[1]

Waves are interesting physical phenomena with important practical applications. Physicists and engineers are interested in the reliable simulation of processes in which waves are scattered from obstacles (scattering problems). This book deals with some of the mathematical issues arising in the computational simulation of wave propagation and fluid–structure interaction.

The linear mathematical models for wave propagation and scattering are well-known. Assuming time-harmonic behavior, one deals with the Helmholtz equation $\Delta u + k^2 u = 0$, where the wave number k is a physical parameter. Our interest will be mainly in the numerical solution of exterior boundary value problems for the Helmholtz equation which we call Helmholtz problems for short.

The Helmholtz equation belongs to the classical equations of mathematical physics. The fundamental questions about existence and uniqueness

[1] In HELMHOLTZ we meet an overwhelming personality. His extraordinary position in the history of science is based on his unusually diverse and penetrating talents, among which the mathematical side, which for our present purpose is of primary importance, plays an important role.

of solutions to Helmholtz problems were solved by the end of the 1950s; cf., e.g., the monographs of Leis [87], Colton–Kress [39], and Sanchez Hubert–Sanchez Palencia [107]. Those results of mathematical analysis form the fundamental layer on which the numerical analysis in this book is built. The two main topics that are discussed here arise from the practical application of finite element methods (FEM) to Helmholtz problems.

First, FEM have been conceptually developed for the numerical discretization of problems on *bounded* domains. Their application to unbounded domains involves a domain decomposition by introducing an artificial boundary around the obstacle. At the artificial boundary, the finite element discretization can be coupled in various ways to some discrete representation of the analytical solution. We review some of the coupling approaches in Chapter 3, focusing on those methods that are based on the series representation of the exterior solution. In particular, we review localized Dirichlet-to-Neumann and other absorbing boundary conditions, as well as the recent perfectly matched layer method and infinite elements.

Second, when using discrete methods for the solution of Helmholtz problems, one soon is confronted with the significance of the parameter k. The wave number characterizes the oscillatory behavior of the exact solution. The larger the value of k, the stronger the oscillations. This feature has to be resolved by the numerical model. The "rule of thumb" is to resolve a wavelength by a certain fixed number of mesh points. It has been known from computational experience that this rule is not sufficient to obtain reliable results for large k. Looking at this problem from the viewpoint of numerical analysis, the reason for the defect can be found in the loss of operator stability at large wave numbers. We address this topic in Chapter 4, where we present new estimates that precisely characterize the error behavior in the range of engineering computations. We call these estimates preasymptotic in order to distinguish them from the well-known asymptotic error estimates for indefinite problems satisfying a Gårding inequality. In particular, we accentuate the advantages of the *hp*-version of the FEM, as opposed to piecewise linear approximation. We also touch upon generalized (stabilized) FEM and investigate *a posteriori* error estimation for Helmholtz problems. Our theoretical results are obtained mainly for one-dimensional model problems that display most of the essential features that matter in the true simulations.

In the introductory Chapters 1 and 2, we set the stage for the finite element analysis. We start with an outline of the governing equations. While our physical application is acoustic fluid–structure interaction, much of the mathematics in this book may be relevant also for numerical electrodynamics. We therefore include a short section on Maxwell's equations. In Chapter 2, we first (Section 2.1) review mathematical techniques for the analytical solution of exterior Helmholtz problems. Our focus is on the separation of variables and series representations of the solution (complementary to the integral methods and representations), as needed for the

outline of the coupling methods in Chapter 3. The second part (Sections 2.2–2.5) of Chapter 2 is a preparation of the finite element analysis in Chapter 4. We first briefly review some necessary definitions and theorems from functional analysis inasmuch as they are needed for the subsequent investigation. Then we consider the variational formulation of Helmholtz problems and discuss variational methods.

Computational results for three-dimensional scattering problems are reported in Chapter 5.

This text is addressed to mathematicians as well as to physicists and computational engineers working on scattering problems. Having a mixed audience in mind, we attempted to make the text self-contained and easily readable. This especially concerns Chapters 3 and 4. We hope that the illustration with many numerical examples makes for a better understanding of the theory. The material of the introductory chapters is presented in a more compact manner for the convenience of later reference. It is assumed that the reader is familiar with the basic physical and mathematical concepts of fluid–structure interaction and/or finite element analysis. References to various expositions of these topics are supplied.

Acknowledgments

> Es ist eben viel wichtiger, in welche geistige Umgebung ein Mensch hineinkommt, die ihn viel stärker beeinflußt als Tatsachen und konkretes Wissen, das ihm geboten wird. (Felix Klein, [84, p. 249])[2]

Much of this book is a report of my own cognitive journey towards the reliable simulation of scattering problems with finite element methods. My interest in numerical acoustics began while I was working as an associate of Ivo Babuška at the University of Maryland at College Park. Many results in this book stem from our joint work, and I have tried my best to put the spirit of our discussions down on paper. My gratitude goes to J. Tinsley Oden and Leszek Demkowicz, of the Texas Institute for Computational and Applied Mathematics (TICAM). Most of this monograph was written during my appointment as a TICAM Research Fellow, and TICAM's extraordinary working conditions and stimulating intellectual atmosphere were an essential ingredient for its shape and content. I gratefully acknowledge the financial support from the Deutscher Akademischer Austauschdienst and the Deutsche Forschungsgemeinschaft. Thanks to Professors Reißmann and Röhr from my home University of Rostock, Germany, for their support of my grant applications. With deep gratitude I acknowledge the close cooperation with Joseph Shirron of the Naval Research Laboratory (NRL) in Washington, D.C., who made his program SONAX available to me and was always ready to share his broad and solid experience in numerical acoustics.

[2]The intellectual environment a person enters is more significant and will be of much greater influence than the facts and concrete knowledge that are offered him.

He was also the first discerning reader of the manuscript. Many thanks also to Oliver Ernst (Freiberg/ College Park), Lothar Gaul (Stuttgart), Jens Markus Melenk (Zürich), and Guy Waryee (Brussels), who carefully read later versions of the text. Their remarks led to a considerable improvement in content and style. Thanks to Louise Couchman (NRL) for permission to use SONAX and to Brian Houston (NRL) for providing me with the results of his experiments and for the numerous explanations of the details of his studies. Many thanks go to Achi Dosanjh, David Kramer, and Victoria Evarretta from the Springer-Verlag New York, for their professional support in making the book. I am much obliged to Malcolm Leighton who thoroughly checked the language of the manuscript. Finally, I gladly use this opportunity to thank my parents Ingrid and Karl Heinz Ihlenburg for their encouragement of my interests.

Hamburg, Germany *Frank Ihlenburg*
February 1998

Contents

1

The Governing Equations of Time-Harmonic Wave Propagation

In this chapter, we outline some of the basic relations of linear wave physics, starting with acoustic waves, proceeding to elastic waves and fluid–solid interaction, and finally considering the Maxwell wave equation. Thus we deal with a scalar field (acoustics), a vector field (elastodynamics), the coupling of these fields in fluid–solid interaction, or two coupled vector fields (electrodynamics). While each class of problems has its distinctive features, there are underlying similarities in the mathematical models, leading to similar numerical effects in computational implementations.

We will be interested mostly in the time-harmonic case, assuming that all waves are steady-state with circular frequency ω. We introduce here the convention that the time variable in a time-dependent scalar field $F(\mathbf{x}, t)$ is separated as

$$F(\mathbf{x}, t) = f(\mathbf{x})e^{-i\omega t}\,, \tag{1.0.1}$$

where f is a stationary function. A similar convention holds for vector fields.

1.1 Acoustic Waves

Acoustic waves (sound) are small oscillations of pressure $P(\mathbf{x}, t)$ in a compressible ideal fluid (acoustic medium). These oscillations interact in such a way that energy is propagated through the medium. The governing equations are obtained from fundamental laws for compressible fluids.

1.1.1 Linearized Equations for Compressible Fluids

Conservation of Mass:

Consider the flow of fluid material with pressure $P(\mathbf{x},t)$, density $\rho(\mathbf{x},t)$, and particle velocity $\mathbf{V}(\mathbf{x},t)$. Let V be a volume element with boundary ∂V, and let $\mathbf{n}(\mathbf{x})$, $\mathbf{x}\in\partial V$, be the normal unit vector directed into the exterior of V, see Fig. 1.1. Then $\mathbf{V}(\mathbf{x},t)\cdot\mathbf{n}(\mathbf{x})$ is the velocity of normal flux

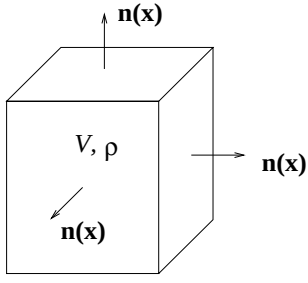


FIGURE 1.1. A volume element with definition of normal direction.

through ∂V. The conservation of mass in a unit time interval is expressed by the relation

$$-\frac{\partial}{\partial t}\int_V \rho dV = \oint_{\partial V} \rho\,(\mathbf{V}\cdot\mathbf{n})\,dS\,. \tag{1.1.1}$$

The surface integral on the right is transformed into a volume integral using the Gauss theorem,

$$\oint_{\partial V} (\rho\mathbf{V})\cdot\mathbf{n}\,dS = \int_V \mathrm{div}\,(\rho\mathbf{V})\,dV\,.$$

We thus obtain

$$\int_V \left(\frac{\partial\rho}{\partial t} + \mathrm{div}\,(\rho\mathbf{V})\right) dV = 0\,,$$

which leads to the continuity equation

$$\frac{\partial\rho}{\partial t} + \mathrm{div}\,(\rho\mathbf{V}) = 0\,. \tag{1.1.2}$$

Remark 1.1. The signs in (1.1.1) correspond to the *decrease* of mass due to *outflow* (namely, in the direction of the exterior normal) of material. A similar relation with opposite signs on both sides is obtained for inflow of material/increase of mass.

Equation of Motion:

Assume that the volume element V is subject to a hydrostatic pressure $P(\mathbf{x},t)$. The total force along ∂V is then $\mathbf{F} = -\oint P\mathbf{n}dS$, where again $\mathbf{n}$ denotes the outward unit normal vector along ∂V. The second Newtonian law $\mathbf{F} = m\mathbf{a}$ now gives

$$-\oint_{\partial V} P\mathbf{n}\, dS = \int_V \rho\frac{d\mathbf{V}}{dt}dV\,.$$

The total differential in the integral on the right is linearized as $d\mathbf{V}/dt \approx \partial\mathbf{V}/\partial t$ (see Remark 1.2 below). Further, from the Gauss theorem it follows that $-\oint_{\partial V} P\mathbf{n}\, dS = -\int_V \nabla P\, dV$, where $\nabla = \{\cdot_{,x}\,, \cdot_{,y}\,, \cdot_{,z}\,\}$ is the nabla operator (gradient) in spatial cartesian coordinates. Thus we arrive at the equation of motion (Euler equation)

$$\rho\frac{\partial\mathbf{V}}{\partial t} = -\nabla P\,. \tag{1.1.3}$$

Remark 1.2. Generally, the total differential $d\mathbf{V}/dt$ is expanded into the nonlinear expression (cf. Landau–Lifshitz [86, p. 3])

$$\frac{d\mathbf{V}}{dt} = \frac{\partial\mathbf{V}}{\partial t} + (\mathbf{V}\cdot\nabla)\mathbf{V}.$$

With the assumption of small oscillations, this relation is linearized in acoustics.

Using the time-harmonic assumption, we obtain the steady-state expression of the Euler equation,

$$i\omega\rho\mathbf{v} = \nabla p\,. \tag{1.1.4}$$

where we applied the separation of variables as in (1.0.1) to the scalar field $P(\mathbf{x},t)$ and the vector field $\mathbf{V}(\mathbf{x},t)$. Introducing the vector field $\mathbf{U}(\mathbf{x},\mathbf{t})$ of fluid particle displacements, the Euler equation is equivalently written as

$$\rho\frac{\partial^2\mathbf{U}}{\partial t^2} = -\nabla P \tag{1.1.5}$$

or, in stationary form,

$$\rho\omega^2\mathbf{u} = \nabla p\,. \tag{1.1.6}$$

1.1.2 Wave Equation and Helmholtz Equation

By definition, sound is a small perturbation (P,ρ) of a constant state (P_0,ρ_0) of a compressible, ideal fluid. At any field point $\mathbf{x}$, the functions $P(\mathbf{x},t), \rho(\mathbf{x},t)$ represent vibrations with a small amplitude. Using the

Euler equation the velocities are also small. Assuming a linear material law, we write

$$P = c^2 \rho, \tag{1.1.7}$$

where the material constant c is called the speed of sound. Then, using linearized versions of (1.1.2) and (1.1.3), we obtain

$$P_{,tt} = c^2 \rho_{,tt} = -c^2 \rho_0 \operatorname{div} \mathbf{V}_{,t} = c^2 \operatorname{div} (\nabla P) \,,$$

to arrive at the wave equation

$$\Delta P - \frac{1}{c^2} P_{,tt} = 0 \,, \tag{1.1.8}$$

where $\Delta = \nabla \cdot \nabla$ is the Laplacian in spatial coordinates. Throughout the book, we denote by $f_{,t}$, $f_{,x}$, etc. the partial derivatives $\frac{\partial f}{\partial t}$, $\frac{\partial f}{\partial x}$, etc. With the assumption of time-harmonic waves (1.0.1), we finally obtain the Helmholtz equation

$$\Delta p + k^2 p = 0 \,, \tag{1.1.9}$$

with

$$k := \frac{\omega}{c} \,. \tag{1.1.10}$$

The physical parameter k of dimension m^{-1} is called the wave number. This notion will become clear in the following paragraph.

One-Dimensional Wave Equation:

Let $x \in \mathbf{R}$. It can be easily checked that functions of the form $P(x,t) = f(kx - \omega t)$ are solutions of the one-dimensional wave equation. The value of the function f does not change if $d(kx - \omega t) = 0$ or, equivalently,

$$\frac{dx}{dt} = \frac{\omega}{k} \,. \tag{1.1.11}$$

The expression

$$v_{\mathrm{ph}} := \frac{dx}{dt} \tag{1.1.12}$$

is called the phase velocity of the solution $f(kx - \omega t)$. Comparing (1.1.11) with (1.1.10) we see that the phase velocity of the one-dimensional solution $f(kx - \omega t)$ is equal to the speed of sound in the acoustic medium and hence depends on material properties only. A wave whose phase velocity is independent of k or ω is called nondispersive. The phase velocity of dispersive waves depends on the wave number k.

To illustrate the meaning of the parameter k, consider steady-state solutions $P(x,t) = p(x)e^{-i\omega t}$. The stationary part satisfies the Helmholtz equation

$$p'' + k^2 p = 0 \,,$$

with the general solution $p(x) = Ae^{ikx} + Be^{-ikx}$. The solution is periodic; i.e., $p(x + \lambda) = p(x)$ holds for all x with

$$\lambda = \frac{2\pi}{k}.$$

The parameter λ is called the wavelength of the stationary wave p. Thus k is the number of waves per "unit" (2π) wavelength.

The corresponding time-dependent solution is

$$P(x,t) = Ae^{i(kx-\omega t)} + Be^{-i(kx+\omega t)}\,. \tag{1.1.13}$$

Computing the phase velocities, we see that the first term on the right-hand side of (1.1.13) represents an *outgoing* wave (traveling to the right with $v_{\text{ph}} = c$), whereas the second term is an *incoming* wave (traveling to the left with $v_{\text{ph}} = -c$). Further, applying at any point $x = x_0$ the boundary condition

$$cP_{,x}\,(x_0) + P_{,t}\,(x_0) = 0\,, \tag{1.1.14}$$

we eliminate the incoming wave. Condition (1.1.14) thus acts as a nonreflecting boundary condition (NRBC) at x_0. We summarize this in Table 1.1 as follows.

TABLE 1.1. Wave directions and nonreflecting boundary conditions.

Sign convention:		$\boxed{P(x,t) = p(x)e^{-i\omega t}}$	
Solutions:	$P_1 = e^{i(kx-\omega t)}$		$P_2 = e^{-i(kx+\omega t)}$
Wave direction:	$d(kx - \omega t) = 0$		$d(kx + \omega t) = 0$
	$v_{\text{ph}} = x_{,t} = c$ $\longrightarrow$ (outgoing)		$v_{\text{ph}} = x_{,t} = -c$ $\longleftarrow$ (incoming)
NRBC:	$cP_{,x} + P_{,t} = 0$	$\downarrow$ $P_{,t} = -i\omega P$	$cP_{,x} - P_{,t} = 0$
	$\swarrow \quad \searrow$ $\boxed{p_{,x} - ikp = 0}$		$\boxed{p_{,x} + ikp = 0}$

Plane Waves:

Important particular solutions of the 2-D and 3-D Helmholtz equations are the plane waves

$$p(\mathbf{x}) = e^{i(\mathbf{k}\cdot\mathbf{x})}$$

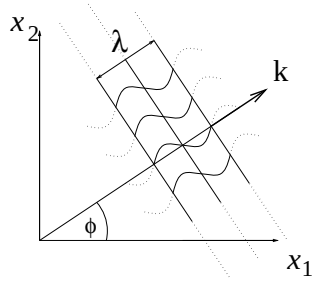


FIGURE 1.2. Plane wave in 2-D.

with $|\mathbf{k}| = k$. Writing (in 2-D) $\mathbf{k} = k\{\cos\phi, \sin\phi\}$, we have

$$p(\mathbf{x}) = e^{ik(x_1 \cos\phi + x_2 \sin\phi)},$$

describing a plane wave with wave number k moving in direction ϕ, see Fig. 1.2. The wave front is a plane (reducing to a straight line in two dimensions) through the point (x_1, x_2) with normal $\mathbf{n} = \mathbf{k}/k = \{\cos\phi, \sin\phi\}$. Along an axis x in direction $\mathbf{k}$, plane waves are one-dimensional waves e^{ikx}, as in the previous example. A nonreflecting boundary condition for a plane wave can thus be prescribed if its direction is known. In general, this is not possible. Instead, one can prescribe *absorbing* boundary conditions (ABC) as an approximation to nonreflecting conditions.

The impedance of a plane wave is constant over the wave front and equal to the *characteristic impedance*

$$z = \rho c\,. \tag{1.1.15}$$

Indeed, impedance is defined as the ratio of the force amplitude to the particle velocity in the normal direction. Multiplying the Euler equation (1.1.4) by the normal $\mathbf{n} = \mathbf{k}/k$ and denoting $v_n = \mathbf{v} \cdot \mathbf{n}$ we have $i\omega\rho v_n = (\mathbf{k}/k) \cdot \nabla p = ikp$, since $\nabla p = i\mathbf{k}p$ for plane waves, and (1.1.15) follows.

1.1.3 The Sommerfeld Condition

Considering wave propagation in free space (unbounded acoustic domain) we postulate that no waves are reflected from infinity. The mathematical expression for this far-field condition is obtained from the Helmholtz integral equation as follows. Let $u(\mathbf{r})$ be a solution of the homogeneous Helmholtz equation $-\Delta u - k^2 u = 0$ in an exterior domain $\Omega^+ = \mathbf{R}^3 \setminus \overline{\Omega}$, and let $g(\mathbf{r}, \mathbf{r}') = g(|\mathbf{r}' - \mathbf{r}|)$ be the free space Green's function (fundamental solution). This function is defined everywhere in Ω^+, relating an "observation point" $\mathbf{r} = (x_1, x_2, x_3)$ to a "source point" $\mathbf{r}' = (x_1', x_2', x_3')$. By definition, this Green's function satisfies the inhomogeneous Helmholtz equation

$$\Delta g(|\mathbf{r}' - \mathbf{r}|) + k^2 g(|\mathbf{r}' - \mathbf{r}|) = \delta(|\mathbf{r}' - \mathbf{r}|)$$

for all $\mathbf{r}, \mathbf{r}' \in \Omega^+$, where δ is the Dirac delta function. In three dimensions, the Green's function is given by

$$g(|\mathbf{r}' - \mathbf{r}|) = \frac{e^{ik|\mathbf{r}'-\mathbf{r}|}}{4\pi|\mathbf{r}' - \mathbf{r}|}\,. \tag{1.1.16}$$

Remark 1.3. In two dimensions, the function is

$$g(|\mathbf{r}' - \mathbf{r}|) = \frac{iH_0^{(1)}(k|\mathbf{r}' - \mathbf{r}|)}{4}\,, \tag{1.1.17}$$

where $H_0^{(1)}(x)$ is the cylindrical Hankel function of the first kind (see Chapter 2).

Using the free space Green's function, one can show (cf. Kress [85, p. 60]) that $u(\mathbf{r}),\ \mathbf{r} \in \Omega^+$, satisfies the integral equation

$$u(\mathbf{r}) = \int_{\partial\Omega} \left[u(\mathbf{r}') \frac{\partial}{\partial n'} g(\mathbf{r}, \mathbf{r}') - g(\mathbf{r}, \mathbf{r}') \frac{\partial}{\partial n'} u(\mathbf{r}') \right] dS(\mathbf{r}')\,. \tag{1.1.18}$$

We truncate Ω^+ at a fictitious far-field boundary in the form of a sphere S_R with large radius R that encloses Ω. The full region will then be recovered by letting $R \to \infty$. Thus $\partial\Omega^+ = \partial\Omega \cup S_R$. We demand that

$$\int_{S_R} \left[u(\mathbf{r}') \frac{\partial}{\partial n'} g(\mathbf{r}, \mathbf{r}') - g(\mathbf{r}, \mathbf{r}') \frac{\partial}{\partial n'} u(\mathbf{r}') \right] dS(\mathbf{r}') \to 0 \tag{1.1.19}$$

for $R \to \infty$. For any fixed $\mathbf{r} \in \Omega^+$, we may assume that the sphere S_R is sufficiently large that

$$R = |\mathbf{r}' - \mathbf{r}| \approx |\mathbf{r}'|\,,$$

and the normal derivatives $\partial/\partial n'$ can be identified with $\partial/\partial R$. Then (1.1.19) becomes

$$\int_{S_R} \frac{1}{R} \left(ik\, u - \frac{u}{R} - \frac{du}{dR} \right) \frac{e^{ikR}}{4\pi}\, dS\,.$$

Hence waves are absorbed at infinity if (note that $dS \sim R^2$)

$$u = O\left(R^{-1}\right)\,, \qquad ik\, u - \frac{du}{dR} = o\left(R^{-1}\right)\,, \quad R \to \infty\,. \tag{1.1.20}$$

Here, the notation $f(x) = o(g(x)),\ x \to \infty$, means that the ratio $f(x)/g(x)$ approaches zero as $x \to \infty$, while $f(x) = O(g(x))$ means that this ratio is bounded for all x. Equation (1.1.20) is known as the Sommerfeld condition. We will call solutions of exterior Helmholtz problems that satisfy the Sommerfeld condition radiating solutions.

Remark 1.4. Strictly speaking, the Sommerfeld condition consists of two equations, one of which characterizes the decay and the other the directional character of the stationary solution in the far field (cf. Dautray–Lions [41, p. 94] or Sanchez Hubert–Sanchez Palencia [107, p. 329]). It can be shown (Wilcox [119]; cf. also Colton–Kress [39, p. 18]) that any function that satisfies both the Helmholtz equation and the radiation condition (i.e., the second equation in the Sommerfeld condition) automatically satisfies the decay condition. Therefore, only the radiation condition is explicitly assumed in most references.

A similar consideration leads to the Sommerfeld condition in $\mathbf{R}^2$,

$$u = O(R^{-1/2}), \quad ik\,u - \frac{du}{dR} = o(R^{-1/2}), \quad R \to \infty\,. \tag{1.1.21}$$

In the one-dimensional case, the condition

$$u'(x) - iku(x) = 0 \tag{1.1.22}$$

selects the outgoing wave e^{ikx} from the set of solutions $\left\{e^{-ikx},\, e^{ikx}\right\}$. Unlike the higher-dimensional case, this condition can be imposed for finite x as a usual mixed boundary condition (also called a Robin condition).

A compact way of writing the radiation condition for any dimension d is

$$u = O\left(R^{-(d-1)/2}\right), \quad ik\,u - \frac{du}{dR} = o\left(R^{-(d-1)/2}\right),\ R \to \infty\,. \tag{1.1.23}$$

1.2 Elastic Waves

In an elastic medium, waves propagate in the form of small oscillations of the stress field. The dynamic equations of elasticity are obtained from the same basic relations of continuum mechanics as the hydrodynamic equations. We review the basic relations only very briefly; a detailed introduction can be found, e.g., in Bedford–Drumheller [21].

1.2.1 Dynamic Equations of Elasticity

Equilibrium:

The derivatives of the stress tensor components σ_{ij} and the components of the applied dynamic load $\mathbf{F}$ are in equilibrium with the components of the dynamic volume force $\rho_{\mathrm{s}}\mathbf{U}_{,tt}$ (here again $\mathbf{U}$ is the vector of displacements),

$$\sigma_{ij\,,j} + F_i = \rho_{\mathrm{s}} U_{i,tt} \tag{1.2.1}$$

for $i = 1, 2, 3$. Note that the summation convention has to be applied in the first term on the left; i.e., the sum is taken over the range of the index j.

Strain-Displacement Relation:

With the assumption of small deformations, the strains are related to the displacements by the linearized equations

$$e_{ij} = \frac{1}{2}(U_{i,j} + U_{j,i}), \quad i,j = 1,2,3\,. \tag{1.2.2}$$

Material Law:

The material is elastic:

$$\sigma_{ij} = \lambda e_{ll}\delta_{ij} + 2Ge_{ij}, \quad i,j = 1,2,3\,, \tag{1.2.3}$$

with

$$\lambda = \frac{E\nu}{(1+\nu)(1-2\nu)} \quad , \quad G = \frac{E}{2(1+\nu)}\,,$$

where E and ν, respectively, denote Young's modulus and the Poisson ratio of the solid, and the summation convention applies for the term e_{ll}.

From equations (1.2.1), (1.2.2), and (1.2.3), the equation of elastodynamic equilibrium is found to be

$$\Delta^*\mathbf{U} + \mathbf{F} = \rho_s \mathbf{U}_{,tt}\,, \tag{1.2.4}$$

where

$$\Delta^*\mathbf{U} = G\Delta\mathbf{U} + (\lambda + G)\nabla(\nabla\cdot\mathbf{U})\,. \tag{1.2.5}$$

Here $\Delta\mathbf{U}$ is the vector Laplacian of $\mathbf{U}$ given by $\Delta\mathbf{U} = \{\Delta U_1, \Delta U_2, \Delta U_3\}^T$, and $\nabla\cdot\mathbf{U}$ is the divergence of the vector field $\mathbf{U}$.

1.2.2 Vector Helmholtz Equations

Equation (1.2.4) can be further transformed, introducing the Helmholtz decomposition

$$\mathbf{U} = \nabla\Phi + \nabla\times\mathbf{\Psi}\,, \tag{1.2.6}$$

with scalar potential Φ and vector potential $\mathbf{\Psi}$. This leads to (in the absence of volume forces)

$$\nabla\left[\rho_s\Phi_{,tt} - (\lambda+2G)\Delta\Phi\right] + \nabla\times\left[\rho_s\mathbf{\Psi}_{,tt} - G\Delta\mathbf{\Psi}\right] = \mathbf{0}\,,$$

which is satisfied if the wave equations

$$\Phi_{,tt} - \alpha^2\Delta\Phi = 0\,, \tag{1.2.7}$$

$$\mathbf{\Psi}_{,tt} - \beta^2\Delta\mathbf{\Psi} = \mathbf{0} \tag{1.2.8}$$

hold, where the elastic speeds of sound are defined as

$$\alpha = \left(\frac{\lambda+2G}{\rho_s}\right)^{1/2}, \qquad \beta = \left(\frac{G}{\rho_s}\right)^{1/2}.$$

For $G = 0$, the equation (1.2.7) reduces to the acoustic wave equation. Since $\lambda = B - 2/3G$, where B is the bulk modulus[1] of the material, the speed of sound in the acoustic medium is

$$c = \sqrt{\frac{B}{\rho}}\,. \tag{1.2.9}$$

All waves in the acoustic medium are compressional. In addition to the compressional waves, an elastic medium also allows shear waves. For further details, see, e.g., [21].

With the assumption of time-harmonic fields, the elastic wave equations lead to the elastic Helmholtz equations

$$\Delta\Phi + k_\alpha^2\Phi = 0\,, \tag{1.2.10}$$
$$\Delta\boldsymbol{\Psi} + k_\beta^2\boldsymbol{\Psi} = \mathbf{0}\,, \tag{1.2.11}$$

with the elastic wave numbers

$$k_\alpha := \frac{\omega}{\alpha}, \qquad k_\beta := \frac{\omega}{\beta}\,.$$

Example 1.5. (In vacuo vibrations of a spherical shell). Consider an elastic spherical shell freely vibrating in vacuo. We denote by t the constant thickness of the shell and by a the radius of its midsurface, see Fig. 1.3. We

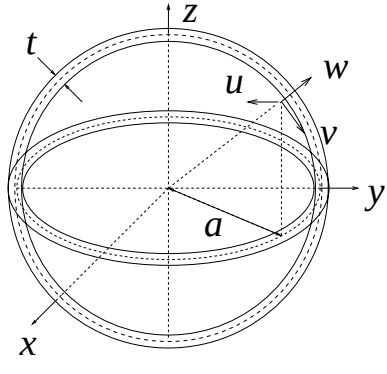


FIGURE 1.3. Spherical shell with coordinate system and deflection components.

are interested in the eigenfrequencies of the vibrations. The characteristic equations are obtained in the following steps:

[1]The bulk modulus is the material constant in the hydrostatic stress-strain relation $s = Be$, where $s = \sigma_{ll}/3$ is the hydrostatic pressure and $e = e_{ll}$ is the volume dilatation.

1. Solve the Helmholtz equations (1.2.10), (1.2.11) for the potentials Φ and $\mathbf{\Psi}$.

2. Compute the displacement field $\{u, v, w\}$ from the potentials using definition (1.2.6).

3. Compute the stress fields (in terms of displacements) using (1.2.2), (1.2.3).

4. Set the stress fields to zero at the inner and outer shell boundaries to obtain a homogeneous system of equations.

5. Compute the characteristic eigenvalues from the condition that the determinant of the system vanishes.

For a detailed outline of this procedure, see Chang and Demkowicz [36]. For the full 3–D solution, the homogeneous system consists of six equations. If the shell is subject to nontorsional axisymmetric motions only; i.e., $v = 0$ and $u = u(\theta), w = w(\theta)$, then the number of equations reduces to four. Chang and Demkowicz further show that for thin shells (up to $t/a = 0.05$) the first 25 eigenfrequencies are computed with sufficient accuracy from the Kirchhoff–Love shell theory. In this theory, eigenfrequencies Ω are found as the real solutions of the frequency equation (see also Junger–Feit [81, p. 231])

$$\Omega^4 - [2(1+\nu) + \lambda_n - (\beta^2(\lambda_n + 1) + 1)(1 - \nu - \lambda_n)]\Omega^2 \\ +(\lambda_n - 2)(1-\nu^2) + \beta^2\left[\lambda_n^3 - 4\lambda_n^2 + \lambda_n(5-\nu^2) - 2(1-\nu^2)\right] = 0$$

with

$$\Omega = \frac{a\omega}{c_p}; \quad \lambda_n = n(n+1); \quad \beta = \frac{t}{\sqrt{12}a}; \quad c_p = \left(\frac{E}{(1-\nu^2)\rho}\right)^{1/2}.$$

This equation is quadratic in Ω; there are two different resonant frequencies for each mode except for the case $n = 0$, which permits only one positive real solution.

1.3 Acoustic/Elastic Fluid–Solid Interaction

An acoustic wave that is incident on a rigid obstacle is totally reflected. This is called rigid scattering of sound. If the obstacle is elastic, a part of the incident energy is transmitted in the form of elastic vibrations. The acoustic pressure waves act as time-varying loads, causing forced elastic vibrations. In that case, we speak of elastic scattering. Conversely, if the

acoustic medium picks up elastic vibrations of an embedded body in the form of acoustic waves, we say that sound is radiated from the body.

In this section, we will obtain the equations for all these effects as special cases of the general equations of fluid–solid interaction.

1.3.1 Physical Assumptions

We consider elastic scattering, making the following assumptions:

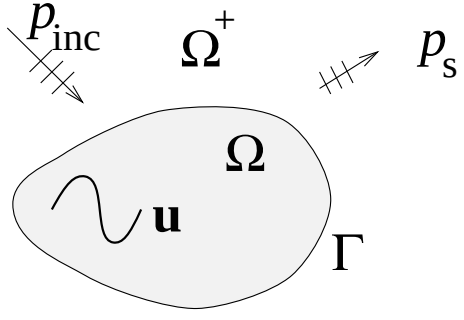


FIGURE 1.4. Fluid–solid interaction (schematic plot).

Space: Let $\Omega \subset \mathbf{R}^3$ be a bounded domain of a solid (called obstacle) with boundary Γ (called wet surface), which is enclosed by the unbounded fluid domain $\Omega^+ = \mathbf{R}^3 \setminus \overline{\Omega}$; see Fig. 1.4. It is assumed that outgoing waves are absorbed in the far field; i.e., no waves are reflected from infinity.

Material: The fluid is supposed to be ideal, compressible, and homogeneous with density ρ_{f} and speed of sound c. The solid obstacle is considered as rigid or linearly elastic with density ρ_{s}. The scalar pressure field in the fluid is denoted by $P(\mathbf{x}, t)$. The vector field of solid displacements is denoted by $\mathbf{U}(\mathbf{x}, t)$.

Time: All waves are steady-state (time-harmonic) with circular frequency ω, satisfying the separation convention (1.0.1).

Range of unknowns: The amplitudes of the oscillations both in the solid and in the fluid regions are supposed to be small.

Load: In the fluid region Ω^+, an incident acoustic field $P_{\text{inc}}(\mathbf{x}, t) = p_{\text{inc}}(\mathbf{x})e^{-i\omega t}$ is given and/or the solid region is subject to a time-harmonic driving force $\mathbf{F}(\mathbf{x}, t) = \mathbf{f}(\mathbf{x}) \exp(-i\omega t)$.

Coordinates: A cartesian coordinate system is fixed in $\mathbf{R}^3$ throughout all calculations (global Lagrangian approach).

The objective is to determine the stationary acoustic field of scattered pressure $p(\mathbf{x})$ for $\mathbf{x} \in \mathbf{\Omega}^+$. This solution is complex-valued. The solution of physical interest is then the real part of $P(\mathbf{x}, t) = p(\mathbf{x})e^{-i\omega t}$. We have

$$\operatorname{Re} P \quad = \quad \operatorname{Re}\left((\operatorname{Re} p + i \operatorname{Im} p)(\cos \omega t - i \sin \omega t)\right)$$

$$
\begin{aligned}
&= |p| \left(\frac{\operatorname{Re} p}{|p|} \cos \omega t + \frac{\operatorname{Im} p}{|p|} \sin \omega t \right) \\
&= |p| \sin(\phi + \omega t) \,,
\end{aligned}
$$

where we define

$$
\phi := \arctan \frac{\operatorname{Re} p}{\operatorname{Im} p} \,. \tag{1.3.1}
$$

Hence, in the standard terminology for harmonic motion (see, e.g., Inman [77, p. 12]), the absolute value of the stationary solution is the amplitude of the physical solution, whereas ϕ is its phase.

1.3.2 Governing Equations and Special Cases

Transmission Conditions:

Consider an arbitrary point $P \in \Gamma$ of the wet surface. Let $\mathbf{n}$ be a unit vector in the outward normal direction and let $\{\mathbf{t_1}, \mathbf{t_2}, \mathbf{n}\}$ be a local orthonormal basis at P; see Fig. 1.5. We then can formulate two conditions of static equilibrium at the point P.

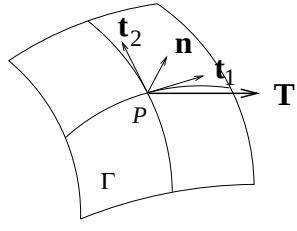


FIGURE 1.5. Local orthonormal coordinate system at the point $P \in \Gamma$.

The pressure is in static equilibrium with the traction normal to the solid boundary:

$$
-(p + p_{\text{inc}}) = T_n \,, \tag{1.3.2}
$$

where p_{inc} is the known incident pressure.

Since the fluid is supposed to be ideal, no tangential traction occurs at the boundary:

$$
T_{t_1} = T_{t_2} = 0 \,. \tag{1.3.3}
$$

The local components of the traction vector $\mathbf{T} = \{T_{t_1}, T_{t_2}, T_n\}^T$ are related by an orthonormal transform to the global cartesian components $\mathbf{T} = \{T_x, T_y, T_z\}^T$. These components are computed by

$$
T_i = \sigma_{ij} n_j \,,
$$

where n_j are the cartesian components of the normal vector $\mathbf{n}$. In local coordinates, we have $\mathbf{T} = \{0, 0, T_n\}^T$, whence

$$T_n = \mathbf{T} \cdot \mathbf{n} = T_i n_i = \sigma_{ij} n_i n_j \,,$$

and the equilibrium conditions are equivalently written as

$$\sigma_{ij} n_i n_j = -(p + p_{\text{inc}}) \,. \tag{1.3.4}$$

Another equation is obtained from the compatability condition that the normal displacements of solid and fluid are equal at the wet surface. We multiply the Euler equation (1.1.6) by the normal vector $\mathbf{n}$ and interpret the resulting normal displacements as solid displacements. This leads to

$$\rho_{\text{f}} \omega^2 \mathbf{u} \cdot \mathbf{n} = \frac{\partial (p + p_{\text{inc}})}{\partial n} \,. \tag{1.3.5}$$

Using the time-harmonic velocity form (1.1.4) of the Euler equation, we can write alternatively

$$-i\omega \rho_{\text{f}} v_n = \frac{\partial (p + p_{\text{inc}})}{\partial n} \,,$$

relating the normal particle velocity to the normal pressure derivative.

Governing Equations:

The transmission conditions, together with the elastodynamic and the acoustic equations, form the following general system of equations for fluid–solid interaction:

$$\begin{aligned}
\Delta^* \mathbf{u} + \rho_{\text{s}} \omega^2 \mathbf{u} &= -\mathbf{f} && \text{in } \Omega_{\text{s}} \,, && (1.3.6)\\
\rho_{\text{f}} \omega^2 \mathbf{u} \cdot \mathbf{n} - \frac{\partial p}{\partial n} &= \frac{\partial p_{\text{inc}}}{\partial n} && \text{on } \Gamma \,, && (1.3.7)\\
\sigma_{ij}(\mathbf{u}) n_i n_j + p &= -p_{\text{inc}} && \text{on } \Gamma \,, && (1.3.8)\\
\Delta p + k^2 p &= 0 && \text{in } \Omega^+ \,, && (1.3.9)\\
p &= O\left(R^{-(d-1)/2}\right) \,, && && \\
\frac{dp}{dR} - ik\, p &= o\left(R^{-(d-1)/2}\right) \,, \quad R \to \infty \,, && && (1.3.10)
\end{aligned}$$

where Δ^* in the first equation is the elasticity operator; see (1.2.5). In equations (1.3.9), (1.3.10), $k = \omega/c$ is the wave number in the fluid.

Special Cases:

From the general equations above, we can deduce several important special cases:

1. $\mathbf{u} = 0$
   Rigid scattering: (1.3.7), (1.3.9), (1.3.10).
2. $T_n = 0$
   Soft scattering: (1.3.8), (1.3.9), (1.3.10).
3. $\mathbf{u} = 0\,,\ p_{\mathrm{inc}} = 0$
   Wave propagation in free space: (1.3.9), (1.3.10).
4. $p_{\mathrm{inc}} = 0$
   Radiation of sound from an elastic body vibrating in fluid: all equations.
5. $p_{\mathrm{inc}} = 0\,,\ p = 0$
   Vibrations of an elastic body in vacuo: (1.3.6), (1.3.8).

Since the model is linear, the general case can be interpreted as a superposition of special cases as follows. We write the solution for elastic scattering $p = p_{\mathrm{se}}$ formally as $p_{\mathrm{se}} = p_{\mathrm{r}} + p_{\mathrm{s}\infty}$, where $p_{\mathrm{s}\infty}$ is the solution for rigid scattering. Since $\partial p_{\mathrm{s}\infty}/\partial n = -\partial p_{\mathrm{inc}}/\partial n$, the transmission condition (1.3.7) reduces to

$$\rho_{\mathrm{f}}\omega^2 \mathbf{u}\cdot\mathbf{n} - \frac{\partial p_{\mathrm{r}}}{\partial n} = 0.$$

The equilibrium condition reads

$$T_n = p_{\mathrm{inc}} + p_{\mathrm{s}\infty} + p_{\mathrm{r}}.$$

Referring to the appropriate special cases, we see that p_{r} can be interpreted as the pressure radiated from the elastic body.

We conclude with a simple example for solid–fluid interaction.

Example 1.6. The system of one-dimensional equations

$$\begin{aligned}
-E\frac{d^2u}{dx^2} - \rho_{\mathrm{s}}\omega^2 u &= f\,, & 0 \le x \le l\\
u &= 0\,, & x = 0\,,\\
\rho_{\mathrm{f}}\omega^2 u - \frac{dp}{dx} &= 0\,, & x = l\,,\\
E\frac{du}{dx} + p &= 0\,, & x = l\,,\\
-\frac{d^2p}{dx^2} - k^2 p &= 0\,, & x \ge l\,,\\
\frac{dp}{dx} - ikp &= 0\,, & x \to \infty\,,
\end{aligned}$$

can be physically interpreted as the forced longitudinal vibrations of a slab of length l that is coupled to an infinite fluid layer. Recall that in one dimension, the radiation condition can be prescribed at any field point $x > l$. Such a condition can even be imposed directly on the slab. Indeed, the solution in the fluid is $p = p_0 \exp(ikx)$, where the amplitude p_0 is determined by the boundary and transmission conditions. Inserting this result into the transmission conditions at $x = l$, we can eliminate p_0 to obtain the boundary condition

$$E\frac{du}{dx} - i\omega\sqrt{\rho_{\mathrm{f}}B}u = 0 \qquad \text{at} \quad x = l\,, \tag{1.3.11}$$

or, equivalently,

$$a\frac{du}{dx} - iku = 0 \qquad \text{at} \quad x = l\,, \tag{1.3.12}$$

with the nondimensional parameter $a = E/\rho_{\mathrm{f}}c^2$. Thus the solution of the coupled problem can be reduced to the solid domain; cf. Demkowicz [42].

A second example (elastic scattering from a thin spherical shell) will be given in Chapter 2.

1.4 Electromagnetic Waves

The electromagnetic wave equations follow from Maxwell's electrodynamic equations. Maxwell's equations describe the interaction between two time-varying force fields: the electric field and the magnetic field. Electric fields are generated by charges. In conducting media, electric fields enforce currents (flow of free electric charges). The interaction of currents generates the magnetic force field. If the magnetic field does not change in time, it does not influence the electric field. The static electric and magnetic fields are coupled only implicitly via their relation to the steady current. However, any change of a magnetic field produces an electric field. Thus the fields are directly coupled in the dynamic case. A simple transformation of Maxwell's equations shows that both the electric and the magnetic fields satisfy a vector wave equation, showing that in both fields energy is transmitted in the form of waves.

1.4.1 Electric Fields

By Coulomb's law, the force exerted by a charge Q on a test charge q is

$$\mathbf{F}_Q = \frac{qQ}{4\pi\epsilon_0 R^2}\mathbf{a}_R\,,$$

where ϵ_0 is a material constant (permittivity), R is the distance between the charges, and $\mathbf{a}_R$ is a unit vector pointing from Q in the direction of q.

Generalizing to a volume distribution of charges with charge density ρ_V, the force exerted on q is

$$\mathbf{F} = \int_V \frac{q\rho_V}{4\pi\epsilon_0 R^2}\mathbf{a}_R dV\,,$$

which leads directly to the definition of electric field intensity

$$\mathbf{E} = \frac{\mathbf{F}}{q} = \int_V \frac{\rho_V}{4\pi\epsilon_0 R^2}\mathbf{a}_R dV\,.$$

Gauss's law states that the total charge in a volume enclosed by a closed surface S is equal to the surface integral of the electric flux density $\epsilon_0\mathbf{E}$:

$$\oint_S \epsilon_0\mathbf{E}\cdot d\mathbf{s} = \int_V \rho\, dV\,.$$

Here $d\mathbf{s} := \mathbf{n}dS$ is the outward differential surface vector, and ρ is the charge per unit volume.

The electric field produces a flow of free charges in a conducting medium. By Ohm's law, the current in a conductor is proportional to the applied electrical field:

$$\mathbf{J} = \sigma\mathbf{E}\,,$$

where σ is the conductivity of the material.

Similarly to the conservation of mass in continuum mechanics, the principle of conservation of charge states that the current passing through any closed surface S is equal to the decrease of charge in the volume enclosed by S:

$$\oint_S \mathbf{J}\cdot d\mathbf{s} = -\frac{\partial}{\partial t}\int_V \sigma dV\,.$$

1.4.2 Magnetic Fields

The magnetic field intensity $\mathbf{H}$ characterizes the force field that is exerted by a current element on another current element in its vicinity. By Ampère's law, the integral of the (static) magnetic field intensity around any closed path is equal to the total current enclosed by that path,

$$\oint_c \mathbf{H}\cdot d\mathbf{l} = \int_S \mathbf{J}\cdot d\mathbf{s}\,.$$

Analogously to electric flux, the magnetic flux density $\mathbf{B}$ is related to the field intensity by the material law

$$\mathbf{B} = \mu\mathbf{H}\,,$$

where the material parameter μ is called the permeability. The integral of the magnetic flux over any closed surface is zero, since magnetic point sources (that would be similar to electric charges) do not exist:

$$\oint_S \mathbf{B}\cdot d\mathbf{s} = 0\,.$$

1.4.3 Maxwell's Equations

Time-varying magnetic fields produce electric fields. The interaction is described by Faraday's law,

$$\oint_c \mathbf{E} \cdot d\mathbf{l} = -\frac{\partial}{\partial t} \int_S \mathbf{B} \cdot d\mathbf{s}\,,$$

and the dynamic Ampère's law reads

$$\oint_c \mathbf{H} \cdot d\mathbf{l} = \int_S \mathbf{J} \cdot d\mathbf{s} + \int_S \frac{\partial \mathbf{D}}{\partial t} \cdot d\mathbf{s}\,.$$

Assuming linear material laws with constant material parameters, we collect the full system of Maxwell's equations in integral form:

$$\begin{aligned}
\oint_c \mathbf{E} \cdot d\mathbf{l} &= -\mu \frac{\partial}{\partial t} \int_S \mathbf{H} \cdot d\mathbf{s}\,,\\
\oint_S \mathbf{E} \cdot d\mathbf{s} &= \frac{1}{\epsilon_0} \int_V \rho dV\,,\\
\oint_c \mathbf{H} \cdot d\mathbf{l} &= \sigma \int_S \mathbf{E} \cdot d\mathbf{s} + \epsilon_0 \frac{\partial}{\partial t} \int_S \mathbf{E} \cdot d\mathbf{s}\,,\\
\oint_S \mathbf{H} \cdot d\mathbf{s} &= 0\,,
\end{aligned}$$

The corresponding differential equations

$$\begin{aligned}
\boldsymbol{\nabla} \times \mathbf{E} &= -\mu \frac{\partial}{\partial t} \mathbf{H}\,,\\
\boldsymbol{\nabla} \cdot \mathbf{E} &= \frac{\rho}{\epsilon_0}\,,\\
\boldsymbol{\nabla} \times \mathbf{H} &= \sigma \mathbf{E} + \epsilon_0 \frac{\partial}{\partial t} \mathbf{E}\,,\\
\boldsymbol{\nabla} \cdot \mathbf{H} &= 0
\end{aligned}$$

are obtained from the integral equalities by applying the divergence theorem

$$\oint_S \mathbf{F} \cdot d\mathbf{s} = \int_V \boldsymbol{\nabla} \cdot \mathbf{F} dV$$

or Stokes's theorem

$$\oint_c \mathbf{F} \cdot d\mathbf{l} = \int_V (\boldsymbol{\nabla} \times \mathbf{F}) \cdot d\mathbf{s}\,,$$

respectively. If the electric field is free of charges ($\rho \equiv 0$), then the system of Maxwell's equations reduces to the coupled equations

$$\begin{aligned}
\boldsymbol{\nabla} \times \mathbf{E} + \mu \frac{\partial}{\partial t} \mathbf{H} &= \mathbf{0}\,,\\
\boldsymbol{\nabla} \times \mathbf{H} - \sigma \mathbf{E} - \epsilon_0 \frac{\partial}{\partial t} \mathbf{E} &= \mathbf{0}\,,
\end{aligned}$$

with the condition that the vector fields be divergence-free. This system can be transformed to the vector wave equation

$$\nabla^2 \mathbf{A} = \mu\sigma \frac{\partial \mathbf{A}}{\partial t} + \mu\epsilon_0 \frac{\partial^2 \mathbf{A}}{\partial t^2} ,$$

where $\mathbf{A} = \mathbf{E}$ or $\mathbf{H}$. The stationary form of the Maxwell wave equation is

$$\nabla^2 \mathbf{A} + \gamma^2 \mathbf{A} = \mathbf{0} , \tag{1.4.1}$$

where

$$\gamma^2 = \omega^2 \mu\epsilon_0 + i\omega\mu\sigma .$$

For exterior radiation and scattering problems, the Maxwell wave equation is associated with the Silver–Müller radiation condition [119]

$$\lim_{R\to\infty} R \left(\mathbf{R}_0 \times (\boldsymbol{\nabla} \times \mathbf{A}) + i\gamma \mathbf{A}\right) = \mathbf{0}; \tag{1.4.2}$$

cf. the Sommerfeld radiation condition in the scalar case. Here $\mathbf{R}_0$ is a unit vector in the direction of $\mathbf{R}$, and $R = |\mathbf{R}|$. Solutions of the Maxwell wave equation satisfying (1.4.2) are called radiating solutions. The radiating solutions automatically satisfy the decay condition $|\mathbf{A}| = O\left(R^{-1}\right)$; see the discussion on the Sommerfeld condition in Section 1.1.3.

If the medium is lossless ($\sigma = 0$), then the wave number is real, and the Maxwell wave equation transforms to the vector Helmholtz equation

$$\nabla^2 \mathbf{A} + k^2 \mathbf{A} = \mathbf{0} ,$$

with

$$k^2 = \omega^2 \mu\epsilon_0 .$$

1.5 Summary

The physical effect of wave propagation is described by the scalar wave equation (acoustics) or the vector wave equation (elastic waves) or a system of two vector wave equations (electrodynamics). In the time-harmonic case, the wave equations transform to the corresponding Helmholtz equations. The Helmholtz equations are characterized by a physical parameter—the wave number. If no loss of energy occurs in the medium, the wave number is real. For exterior problems, we impose the Sommerfeld condition (or the Silver–Müller condition for Maxwell's equations) in the far field. This condition, which prevents the reflection of outgoing waves from infinity, introduces the radiation damping into the physical model. Consequently, standing waves cannot occur in exterior problems.

The components of vector Helmholtz equations satisfy scalar Helmholtz equations. Thus the investigation of scalar (acoustic) Helmholtz equations is also relevant for the physical problems of elastic and electrodynamic wave propagation.

1.6 Bibliographical Remarks

The material of this chapter can be found in many textbooks and monographs. We have used mostly the corresponding volumes of the treatise by Landau and Lifshitz [86] as well as the textbooks by Nettel [99] and by Bedford and Drumheller [21]. The monograph by Junger and Feit [81] is a standard reference for formulations and methods in acoustic scattering and fluid–solid interaction. A very precise and detailed outline of the mathematical models is given by Dautray and Lions [41]. For instance, we have introduced here the assumption of time-harmonic behavior without further ado. This assumption has to be understood in its asymptotic meaning, as given in [41, p. 92, Remark 4].

2

Analytical and Variational Solutions of Helmholtz Problems

In this chapter we review essential facts concerning the analytical ("strong") and variational ("weak") solutions of exterior Helmholtz problems. This is in preparation for the following chapters, where we will treat the solution of exterior Helmholtz problems with finite element methods. The finite element discretization of the exterior is carried out in a small annular domain enclosing the scatterer. The solution behavior in the exterior of that domain must be modeled and coupled appropriately to the degrees of freedom of the finite element model. Thus information from the analytical solution is used in the exterior region, leading to a semianalytical numerical model. In particular, one has to make sure that essential characteristics of the mathematical formulation, such as the radiation damping, are carried over to the numerical approximation.

As a practical matter, solutions to partial differential equations can be found in the form of an integral representation or by separation of variables (which, in general, leads to a series representation).[1] Integral representations are used in the boundary integral method. For our purpose of finite element analysis and simulation, we will be mostly interested in the coupling of finite elements with various discretizations of the separated solution. Further, we review variational formulations of exterior Helmholtz problems and discuss variational methods. In particular, we explore the conditions

[1] cf. Morse and Feshbach [95, pp. 493-494]: "Aside from a few cases where solutions are guessed and then verified to be solutions, only two generally practicable methods of solution are known, the *integral solution* and the *separated solution*."

for well-posedness of variational problems with indefinite variational forms and discuss the convergence of variational methods in this case.

2.1 Separation of Variables

We illustrate the technique on Helmholtz problems in cartesian, spherical, and cylindrical coordinates. The solutions will be used in the semianalytical discretization methods described in Chapter 3.

2.1.1 Cartesian Coordinates

Consider the Helmholtz equation $\Delta u + k^2 u = 0$ in $\mathbf{R}^3$. Looking for nontrivial solutions of the form $u = X(x)Y(y)Z(z)$, we get

$$X''YZ + XY''Z + XYZ'' + k^2 XYZ = 0\,,$$

which we can rewrite as

$$-\frac{X''}{X} = \frac{Y''}{Y} + \frac{Z''}{Z} + k^2\,.$$

Since the right-hand side of the latter equation does not depend on x, equality can hold only if both sides are equal to a constant, say λ. Thus we find that the two equations

$$\begin{aligned} X'' + \lambda X &= 0\,, \\ \frac{Y''}{Y} + \frac{Z''}{Z} + k^2 - \lambda &= 0 \end{aligned}$$

must hold simultaneously. Repeating now the argument for the second equation, we see that functions X, Y, and Z satisfy

$$\begin{aligned} X'' + \lambda X &= 0\,, \\ Y'' + \nu Y &= 0\,, \\ Z'' + (k^2 - \lambda - \nu)Z &= 0\,, \end{aligned} \tag{2.1.1}$$

for independent constants λ, ν. Since we are interested in propagating solutions, we consider only positive real values of these constants, letting $\lambda := \alpha^2, \nu := \beta^2$ with $\alpha, \beta \in \mathbf{R}$. Then the last equation can be rewritten as

$$Z'' + \gamma^2 Z = 0\,,$$

with

$$\gamma := \sqrt{k^2 - \alpha^2 - \beta^2}\,.$$

For $k^2 \geq \alpha^2 + \beta^2$, the parameter γ is also real, and we have obtained solutions in the form of plane waves

$$u(x, y, z) = e^{i(\alpha x + \beta y + \gamma z)} ,$$

where the parameters α, β, γ satisfy the dispersion relation

$$\alpha^2 + \beta^2 + \gamma^2 = k^2 . \tag{2.1.2}$$

For $\alpha^2 + \beta^2 > k^2$, γ is complex, and hence the solution is decaying in the z-direction. Such solutions are called evanescent waves. Similarly, one obtains evanescent waves by allowing α or β to be a complex number.

Consider the function

$$u(x, y, z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U(\alpha, \beta) e^{i(\alpha x + \beta y + \gamma z)} d\alpha d\beta , \tag{2.1.3}$$

where α, β, γ have to satisfy the dispersion relation (2.1.2), and $U(\alpha, \beta)$ is an amplitude function such that the integral is defined. This general form of the separated solution is called a "packet" of plane waves with wave vector $\mathbf{k} = \{\alpha, \beta, \gamma\}$. Choosing $\gamma = +\sqrt{k^2 - \alpha^2 - \beta^2}$ defines a wave packet traveling to the right on the z-axis (outgoing in the z-direction), whereas the choice $\gamma = -\sqrt{k^2 - \alpha^2 - \beta^2}$ defines a wave packet traveling to the left.

Example 2.1. Wave Guide with a Square Cross-Section. In Fig. 2.1, a wave guide with a square cross-section $0 \leq x, y \leq \pi$ is depicted. Assume that the pressure-release condition $p = 0$ is given at the walls of the wave guide. Looking for a solution of the Helmholtz equation $\Delta p + k^2 p = 0$ by

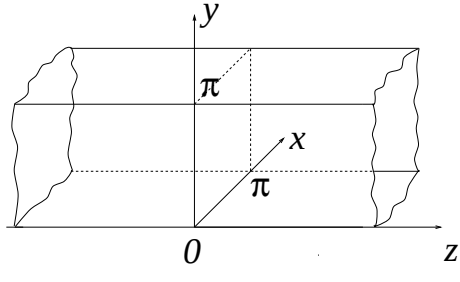


FIGURE 2.1. Wave guide with a square cross-section.

separation of variables, we arrive at the system (2.1.1). From the pressure-release condition, we find that the separated boundary conditions for the first two equations are $X(0) = X(\pi) = 0, Y(0) = Y(\pi) = 0$. The corresponding boundary value problems have eigenfunctions $X = \sin nx, Y = \sin my$

for the integer eigenvalues $\lambda = n^2, \nu = m^2$, respectively. The solution p hence can be written as

$$p(x,y,z) = \sum_{m,n=1}^{\infty} p_{mn} Z_{mn}(z) \sin nx \sin my\,, \tag{2.1.4}$$

where the functions $Z_{mn}(z)$ satisfy

$$Z''_{mn} + \gamma^2_{mn} Z_{mn} = 0\,,$$

with

$$\gamma_{mn} := \sqrt{k^2 - n^2 - m^2}\,.$$

The elementary solutions of this equation are

$$Z^{(1)}_{mn} = e^{i\gamma_{mn} z}\,, \quad Z^{(2)}_{mn} = e^{-i\gamma_{mn} z}\,,$$

which are propagating if $k^2 \geq m^2 + n^2$. With the convention of Table 1.1, we find that the functions $Z^{(1)}_{mn}$ represent waves that travel in the positive z-direction (outgoing solutions), whereas functions $Z^{(2)}_{mn}$ represent incoming solutions.

2.1.2 Spherical Coordinates

Spherical Solutions:

Consider the Neumann problem in the exterior of a sphere of radius a. We seek a function $u(r,\phi,\theta)$ satisfying

$$\Delta u + k^2 u = 0\,, \qquad r > a\,, \tag{2.1.5}$$

$$\frac{\partial u}{\partial r} = w\,, \qquad r = a\,, \tag{2.1.6}$$

$$\frac{\partial u}{\partial r} - iku = o(r^{-1})\,, \qquad r \to \infty\,, \tag{2.1.7}$$

where $r = \sqrt{x^2+y^2+z^2}$, $\theta = \arctan(\sqrt{x^2+y^2}/z)$, $\phi = \arctan(y/x)$ are the spherical coordinates; see Fig. 2.2. We recall that solutions of exterior Helmholtz problems that satisfy the Sommerfeld condition are called radiating solutions.

The Laplacian in spherical coordinates has the form

$$\Delta u(r,\theta,\phi) = \frac{1}{r^2}\left[\frac{\partial}{\partial r}\left(r^2 \frac{\partial u}{\partial r}\right) + \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta \frac{\partial u}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2 u}{\partial\phi^2}\right].$$

Looking for a solution $u = f(r)g(\theta)h(\phi)$, we obtain the separated ordinary differential equations

$$\frac{d}{dr}\left(r^2 \frac{df(r)}{dr}\right) + (k^2 r^2 - \lambda) f(r) = 0\,, \tag{2.1.8}$$

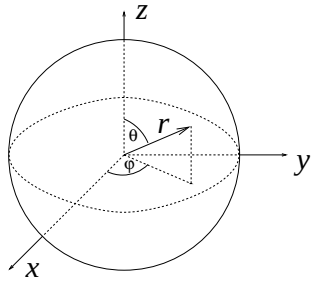


FIGURE 2.2. Spherical coordinates.

$$\sin\theta \frac{d}{d\theta}\left(\sin\theta \frac{dg(\theta)}{d\theta}\right) + \left(\lambda \sin^2\theta - \nu\right) g(\theta) = 0\,, \tag{2.1.9}$$

$$\frac{d^2h(\phi)}{d\phi^2} + \nu h(\phi) = 0, \tag{2.1.10}$$

where λ and ν are constants. The function f is defined in the exterior $r > a$, whereas the functions g, h are defined on S. Moreover, the function f satisfies the Sommerfeld condition.

Since S is a closed surface, the function h is subject to the periodicity condition $h(0) = h(2\pi)$. Equation (2.1.10) hence admits the solutions $\sin mx, \cos mx$ for $\nu = m^2, m = 0, 1, 2, \ldots$

Equation (2.1.9) can be transformed via $t := \cos\theta$ to

$$(1-t^2)\frac{d^2g}{dt^2} - 2t\frac{dg}{dt} + \left(\lambda - \frac{m^2}{1-t^2}\right) g(t) = 0\,.$$

For $\lambda = n(n+1), n = 0, 1, \ldots$, this is Legendre's equation. The solutions are the associated Legendre functions

$$g_{mn}(\theta) = P_n^m(\cos\theta), \quad 0 \le m \le n\,.$$

These functions are defined from the Legendre polynomials by

$$P_n^m(t) := (1-t^2)^{m/2} \frac{d^m P_n(t)}{dt^m}, \tag{2.1.11}$$

where the Legendre polynomials are defined by the recurrence relation

$$\begin{aligned} (n+1)P_{n+1}(t) &= (2n+1)tP_n(t) - nP_{n-1}(t)\,, \\ P_1(t) &= t\,, \\ P_0(t) &= 1\,. \end{aligned} \tag{2.1.12}$$

Finally, for $\lambda = n(n+1)$, equation (2.1.8) is Bessel's differential equation. For each n, this equation has the independent solutions

$$h_n^{(1)}(kr) = i^{-n-1}\frac{e^{ikr}}{kr}\sum_{j=0}^{n}\left(n+\frac{1}{2}, j\right)(-2ikr)^{-j}\,,$$

$$h_n^{(2)}(kr) \quad = \quad i^{n+1}\frac{e^{-ikr}}{kr}\sum_{j=0}^{n}\left(n+\frac{1}{2},j\right)(2ikr)^{-j}\,,$$

called spherical Hankel functions [1, 10.1.16–17]. We have used the notation

$$\left(n+\frac{1}{2},j\right) = \frac{(n+j)!}{j!(n-j)!}$$

for integer $n, j \geq 0$ [1, 10.1.9]. It is easy to see that in the far field (large r) the Hankel functions depend on r as

$$h_n^{(1)}(kr) \sim \frac{e^{ikr}}{r}\,, \quad h_n^{(2)}(kr) \sim \frac{e^{-ikr}}{r}\,,$$

and hence the Hankel functions of the second kind represent incoming waves, which are eliminated by the Sommerfeld condition. In the following, we will omit the superscript where no confusion can arise, writing $h_n(z)$ for the Hankel functions of the first kind $h_n^{(1)}(z)$.

Collecting results, we can write the solution u as

$$u(r,\theta,\phi) = \sum_{n=0}^{\infty}\sum_{m=0}^{n} h_n(kr)P_n^m(\cos\theta)(A_{nm}\cos(m\phi)+B_{nm}\sin(m\phi))\,. \tag{2.1.13}$$

If u is a radiating solution of the Helmholtz equation in the domain exterior to the spherical surface $|\mathbf{x}| = a$, then the series (2.1.13) converges absolutely and uniformly in every closed and bounded domain that is contained in $|\mathbf{x}| > a$ (cf. Colton–Kress [39, p. 33]).

Spherical Harmonics:

Using the de Moivre identity, (2.1.13) is more compactly written as

$$u(r,\theta,\phi) = \sum_{n=0}^{\infty} h_n(kr)\sum_{m=-n}^{n} c_{mn}y_{mn}(\theta,\phi)\,, \tag{2.1.14}$$

where c_{mn} are complex coefficients and

$$y_{mn}(\theta,\phi) := P_n^{|m|}(\cos\theta)e^{im\phi}, \qquad -n \leq m \leq n\,,$$

are the *spherical harmonics*. From the construction outlined above, it is clear that the spherical harmonics are eigenfunctions of the Laplace operator for $r \equiv$ constant. The harmonics

$$y_{0n} = P_n(\cos\theta)$$

do not depend on ϕ and thus represent the axisymmetric (with respect to the z-axis) modes. Some important properties of the spherical harmonics are listed below. For the proof, see, e.g., Colton–Kress [39, pp. 20–26].

(1) For each $n \geq 0$, there exist exactly $2n+1$ linearly independent spherical harmonics y_{mn}, $m = -n, \ldots, n$.

(2) The spherical harmonics are orthogonal with respect to the inner product $(u, v)_0 = \int_S u\bar{v}dS$, where S is the surface of the unit sphere. More precisely, we have

$$\int_S y_{mn}\bar{y}_{m'n'}dS = 0 \qquad \text{if } m \neq m' \text{ or } n \neq n' ,$$

and

$$\int_S |y_{mn}|^2 dS = \int_0^{2\pi} d\phi \int_0^{\pi} |y_{mn}|^2 \sin\theta d\theta = \frac{4\pi}{(2n+1)} \frac{(n+|m|)!}{(n-|m|)!} =: \alpha_{mn}^2 .$$

To have an orthonormal system, we redefine

$$Y_{mn} := \frac{y_{mn}}{\alpha_{mn}} .$$

The orthonormalized harmonics Y_{mn} satisfy

$$\int_S Y_{mn}\overline{Y_{m'n'}}dS = \delta_{mm'}\delta_{nn'} .$$

(3) Each square-integrable function $f(\theta, \phi)$ (i.e. the integral $\int_S |f|^2 dS$ exists and is finite) can be expanded into a series of spherical polynomials

$$f(\theta, \phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} f_{mn} Y_{mn}(\theta, \phi) , \tag{2.1.15}$$

where

$$f_{mn} = \int_S f(\theta', \phi')\overline{Y_{mn}}(\theta', \phi')dS' . \tag{2.1.16}$$

Solution of Boundary Value Problems:

Properties (2) and (3) are effectively used to determine the unknown coefficients c_{mn} in (2.1.14) from the boundary condition (2.1.6). By (3), the data can be expanded as

$$w(\theta, \phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} w_{mn} Y_{mn}(\theta, \phi) ,$$

leading to the equality

$$\frac{\partial u}{\partial r}|_{r=a} = \sum_{n=0}^{\infty} kh_n'(ka) \sum_{m=-n}^{n} c_{mn} Y_{mn}(\theta, \phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} w_{mn} Y_{mn}(\theta, \phi) .$$

Multiplying this equation by $\overline{Y}_{mn}$ and integrating over S, we get, by orthogonality,

$$c_{mn} = \frac{w_{mn}}{kh_n'(ka)}\,.$$

Similarly, the coefficients for the exterior Dirichlet problem are

$$c_{mn} = \frac{w_{mn}}{h_n(ka)}\,.$$

Example 2.2. (Rigid Scattering from a Sphere) Assume that a plane wave with amplitude P_0 is incident in the z-direction on a sphere with radius a. The incident pressure is written in spherical coordinates as

$$p_{\rm inc}(r,\theta) = P_0 e^{ikr\cos\theta}. \tag{2.1.17}$$

The exponential function in (2.1.17) can be expanded as [1, 10.1.47]

$$e^{ikr\cos\theta} = \sum_{n=0}^{\infty}(2n+1)i^n P_n(\cos\theta) j_n(kr)\,.$$

Here $j_n(z)$ are the spherical Bessel functions of the first kind, which can be defined by an ascending series as [1, 10.1.2]

$$j_n(z) = \frac{z^n}{1\cdot 3\cdot\cdots\cdot(2n+1)}\left(1 - \frac{\frac{1}{2}z^2}{1!(2n+3)} + \frac{\left(\frac{1}{2}z^2\right)^2}{2!(2n+3)(2n+5)} - \cdots\right).$$

The problem is symmetric with respect to the angular coordinate ϕ, and hence only coefficients $w_n := w_{n0}$ need be computed in the expansion for $\partial u/\partial r$. We compute the boundary data for the Neumann condition (2.1.6) as

$$w(\theta,\phi) = -\frac{dp_{\rm inc}}{dr}\bigg|_{r=a} = -kP_0\sum_{n=0}^{\infty}(2n+1)i^n P_n(\cos\theta) j_n'(ka)\,.$$

Hence the coefficients in the data expansion are $w_n = -kP_0(2n+1)i^n j_n'(ka)$, and we readily compute

$$c_{n0} =: c_n = \frac{w_n}{kh_n'(ka)} = -P_0 i^n(2n+1)\frac{j_n'(ka)}{h_n'(ka)}\,.$$

The scattered wave is thus

$$p_{s\infty}(r,\theta) = -P_0\sum_{n=0}^{\infty} i^n(2n+1)\frac{j_n'(ka)}{h_n'(ka)}P_n(\cos\theta)h_n(kr)\,. \tag{2.1.18}$$

Example 2.3. (**Elastic Scattering from a Thin Spherical Shell**). As in the case of vibrations in vacuo (Example 1.4), the stress field and displacement field are computed from the potentials Φ and $\mathbf{\Psi}$. The stress field again satisfies homogeneous boundary conditions at the interior boundary. A nonhomogeneous condition (cf. (1.3.8)) is given for the stress resultant on the wet surface. Applying separation of variables and expansion into spherical harmonics, the exact solution for the scattered pressure in terms of the Kirchhoff–Love shell theory is found to be

$$p_{\mathrm{r}}(r,\theta) = P_0 \frac{\rho_{\mathrm{f}} c_{\mathrm{f}}}{(kr_{\mathrm{o}})^2} \sum_{n=0}^{\infty} i^n \frac{(2n+1)P_n(\cos\theta)h_n(kr)}{(h_n'(kr_{\mathrm{o}}))^2(Z_n + z_n)}\,, \tag{2.1.19}$$

where we used notations $c_{\mathrm{f}}, \rho_{\mathrm{f}}$, respectively, for the speed of sound and density in the fluid, and r_{o} for the outer radius of the shell (marking the wet surface). As before, h_n are the spherical Hankel functions of the first kind.

The expression Z_n is the *modal mechanical impedance*, defined as the ratio of the nth pressure mode to the corresponding shell velocity mode,

$$Z_n := \frac{\begin{vmatrix} \Omega^2 - (1+\beta^2)(\nu+\lambda_n-1) & -\beta^2(\nu+\lambda_n-1) - (1+\nu) \\ -\lambda_n(\beta^2(\nu+\lambda_n-1) + (1+\nu)) & \Omega^2 - 2(1+\nu) - \beta^2\lambda_n(\nu+\lambda_n-1) \end{vmatrix}}{-i\omega \begin{vmatrix} \Omega^2 - (1+\beta^2)(\nu+\lambda_n-1) & 0 \\ -\lambda_n(\beta^2(\nu+\lambda_n-1) + (1+\nu)) & \dfrac{-a^2(1-\nu^2)}{Et} \end{vmatrix}}\,, \tag{2.1.20}$$

where, as before, $\lambda_n = n(n+1)$ and t denotes the thickness of the sphere.

Further, the *modal specific acoustic impedance*, defined as the ratio of pressure to normal fluid particle velocity, is

$$z_n = i\rho_{\mathrm{f}} c_{\mathrm{f}} \frac{h_n(kr_{\mathrm{o}})}{h_n'(kr_{\mathrm{o}})}\,. \tag{2.1.21}$$

The details of the solution procedure can be found, for example, in Junger–Feit [81] or in Chang and Demkowicz [37].

Comparing to the case of vibrations in vacuo, we see that the modal mechanical impedance is zero at the eigenfrequencies of the elastic shell. Still, no resonance of the coupled system fluid–solid occurs due to the imaginary term z_n. Thus the acoustic impedance represents the fluid damping of the coupled system. The behavior of the system is close to resonance for small z_n. Note that $z_n = 0$ for $\rho_{\mathrm{f}} = 0$. This is the case of forced vibrations in vacuo where resonance occurs at the eigenfrequencies of free vibration.

2.1.3 Cylindrical Coordinates

Certain problems of scattering from infinite cylinders can be reduced to two-dimensional scattering from a "circle." We then seek radiating solutions

to the Helmholtz equation in polar coordinates. Separation of variables leads to

$$u(r,\phi) = \sum_{n=0}^{\infty} H_n(kr)(A_n \cos(n\phi) + B_n \sin(n\phi)) \,,$$

where $H_n(z) =: H_n^{(1)}(z)$ are the cylindrical Hankel functions of the first kind. These functions are defined as

$$H_n(z) = J_n(z) + iY_n(z) \,,$$

where

$$J_n(z) = \left(\frac{z}{2}\right)^n \sum_{j=1}^{\infty} \frac{\left(-\frac{z^2}{4}\right)^j}{j!(n+j)!}$$

are the cylindrical Bessel functions of the first kind and $Y_n(z)$ are the cylindrical Bessel functions of the second kind.[2] By identifying $H_{-n} = H_n, n = 1, 2, \ldots$, this can formally also be written as

$$u(r,\phi) = \sum_{n=-\infty}^{\infty} u_n H_n(kr) e^{in\phi} \,. \tag{2.1.22}$$

For the solution of boundary value problems, the same procedure as in three dimensions applies, but the expansion into spherical harmonics is replaced by Fourier expansion around the unit circle. Hence,

$$u_n = \frac{1}{2\pi} \int_0^{2\pi} u(\phi) e^{-in\phi} \, d\phi \,.$$

Example 2.4. (Rigid Scattering of a Plane Wave). We assume an incident plane wave as in (2.1.17). Using an addition theorem [1, 9.1.44–45], we decompose

$$\begin{aligned} p_{\text{inc}}(r,\theta) &= P_0 e^{ikr\cos(\theta)} \\ &= 2P_0 {\sum_{n=0}^{\infty}}' i^n J_n(kr)\cos(n\theta) \,, \end{aligned} \tag{2.1.23}$$

where the $'$ after the sum symbolizes the rule that the term corresponding to $n = 0$ is multiplied by the factor $1/2$. The scattered pressure is then computed as

$$p_{\text{s}\infty}(r,\theta) = -2P_0 {\sum_{n=0}^{\infty}}' i^n \frac{J_n'(ka)H_n(kr)}{H_n'(ka)} \cos(n\theta) \,, \tag{2.1.24}$$

[2]The definition of these functions is beyond the scope of this book. For a detailed introduction to Bessel functions, see [29].

and the total pressure field on the wet surface is

$$p(a,\theta) = p_{\text{inc}} + p_{\text{s}\infty} = 2P_0 {\sum_{n=0}^{\infty}}' i^n \left(J_n(ka) - \frac{J_n'(ka) H_n(ka)}{H_n'(ka)} \right) \cos(n\theta) \,.$$

2.1.4 Atkinson–Wilcox Expansion

Using the integral representation of radiating solutions, Wilcox [119] shows that any vector field $\mathbf{A}(\mathbf{r})$ satisfying for $r = |\mathbf{r}| > c$ the vector Helmholtz equation (1.4.1) and the Silver–Müller radiation condition (1.4.2) can be expanded as a function of r in the series

$$\mathbf{A}(\mathbf{r}) = \frac{e^{ikr}}{r} \sum_{n=0}^{\infty} \frac{\mathbf{A}_n(\theta,\phi)}{r^n} \,, \tag{2.1.25}$$

where r, θ, ϕ are the spherical coordinates of $\mathbf{r}$. The series converges absolutely and uniformly in the parameters r, θ, ϕ in any region $r \geq c + \varepsilon > c$. The series is differentiable with respect to all coordinates any number of times, and the derivatives have the same convergence properties as the original expansion.

For a scalar field satisfying the Helmholtz equation and the Sommerfeld radiation condition, it is shown that similarly

$$u(\mathbf{r}) = \frac{e^{ikr}}{r} \sum_{n=0}^{\infty} \frac{u_n(\theta,\phi)}{r^n} \,, \tag{2.1.26}$$

with the same convergence properties as in the vector case.

The expansion theorems are proven without an assumption on the decay character of the solution. Hence the fact that

$$|u| = O(r^{-1})$$

asymptotically for large r *follows* from the theorem, which assumes only that u is a radiating solution (cf. the discussion in Section 1.1.3).

For the two-dimensional case in polar coordinates, similar expansions were given by Karp [82]. He proves that for each outgoing wave function the expansion

$$u(k\mathbf{r}) = H_0^{(1)}(kr) \sum_{n=0}^{\infty} \frac{f_n(\theta)}{r^n} + H_1^{(1)}(kr) \sum_{n=0}^{\infty} \frac{g_n(\theta)}{r^n} \tag{2.1.27}$$

converges absolutely and uniformly for $r > a > 0$. Also, this representation can be differentiated any number of times to yield again a convergent series.

2.1.5 Far-Field Pattern

For large r, the Atkinson–Wilcox expansion leads to the asymptotic equality

$$u(\mathbf{r}) \simeq \frac{e^{ikr}}{r} F(\theta, \phi)\,,$$

disregarding terms of order $O(r^2)$. The function $F(\theta, \phi)$, which describes the angular behavior of u at a large distance from the origin, is called the *far-field pattern* of u. It can be equivalently defined as

$$\boxed{F\left(\frac{\mathbf{r}}{r}\right) \simeq u(\mathbf{r})\, r\, e^{-ikr}\,, \quad r \to \infty\,.} \tag{2.1.28}$$

The integral representation can be used to compute the far-field pattern of a radiating solution. The radius $R = |\mathbf{r} - \mathbf{r}'|$ can for $r \gg r'$ be approximated as

$$\begin{aligned} R &= r\left(1 + \left(\frac{r'}{r}\right)^2 - 2\,\frac{r'}{r}\cos(r, r')\right)^{1/2} \\ &\simeq r\,(1 - \frac{r'}{r}\,\cos(r, r')\,) = r - \frac{\mathbf{r}\cdot\mathbf{r}'}{r}\,, \end{aligned}$$

where we have neglected the square of r'/r and introduced a first order approximation for the square root. Inserting this approximation into (1.1.16), we obtain

$$g(\mathbf{r}, \mathbf{r}') \simeq \frac{1}{4\pi}\frac{e^{ikr}}{r}\,\exp\left(-ik\frac{\mathbf{r}\cdot\mathbf{r}'}{r}\right)\,.$$

We can assume without loss of generality that $\mathbf{r}'$ lies on the unit sphere S_0. Hence we write $\mathbf{r}' = \mathbf{n}'$ to obtain the asymptotic expression

$$g(\mathbf{r}, \mathbf{n}') \simeq \frac{1}{4\pi}\frac{e^{ikr}}{r}\,\exp\left(-ik(\mathbf{n}\cdot\mathbf{n}')\right)\,, \tag{2.1.29}$$

where $\mathbf{n} = \mathbf{r}/r$. Introducing this expression into the Helmholtz integral equation (1.1.18), we arrive at the asymptotic (for large r) formula for the computation of the far-field pattern:

$$F\left(\frac{\mathbf{r}}{r}\right) \simeq -\frac{1}{4\pi}\int_{S_0}\left(ik\,(\mathbf{n}\cdot\mathbf{n}')u(\mathbf{n}') + \frac{\partial}{\partial n'}u(\mathbf{n}')\right)\,e^{-ik(\mathbf{n}\cdot\mathbf{n}')}\,dS\,. \tag{2.1.30}$$

2.1.6 Computational Aspects

The series in expansions (2.1.15) or (2.1.22), respectively, must be truncated for numerical evaluation. Thus, instead of the exact solution u_{ex}, one really computes a truncated solution $u_N = \sum_{n=0}^{N} \cdots$. Computational experience

shows that there is a connection between the critical truncation parameter N (i.e., N such that some norm of the error $u - u_N$ is smaller than some tolerance ε) and the argument ka of the Hankel functions. Let us illustrate this issue with computational experiments for cylindrical scattering.

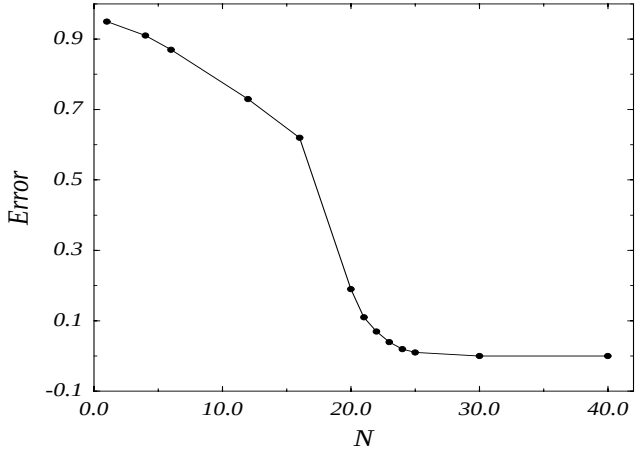


FIGURE 2.3. Approximation of plane wave by truncated series along circle $r \equiv 1$. Truncation error for $k = 20$.

We evaluate first the series expansion of the incident signal (2.1.23). The implementation uses the routine **bessjy** from [100, Section 6.7]. We compute the exact plane wave $e^{ikx} = e^{ik(r\cos\theta)} =: p(kr, \theta)$ and its truncated series

$$p_N := \sum_{n=0}^{N}{}' i^n J_n(kr)\cos(n\theta)\,. \tag{2.1.31}$$

We evaluate both functions at n_{res} points (parameter of graphical resolution), first on the interval $\theta \in [0, \pi]$ with $r \equiv a = 1$ (wet surface). The error is computed in the discrete l^2-norm

$$e(N,k) = \frac{1}{n}\left(\sum_{j=1}^{n} |[p - p_N](kr_j, \theta_j)|^2\right)^{1/2}$$

at $n = 50$ uniformly distributed control points.

In Fig. 2.3, we show the error in the l^2-norm as a function of N, for $k = 20$. We observe a dropping of the error around $k = N$. Similar observations can be made if one computes the error along a radial line $\theta \equiv$ constant. In Fig. 2.4, we show the error of the total solution $p = p_{\text{inc}} + p_{\text{s}\infty}$. The values for the "exact solution" are obtained here with truncation $N = 100$.

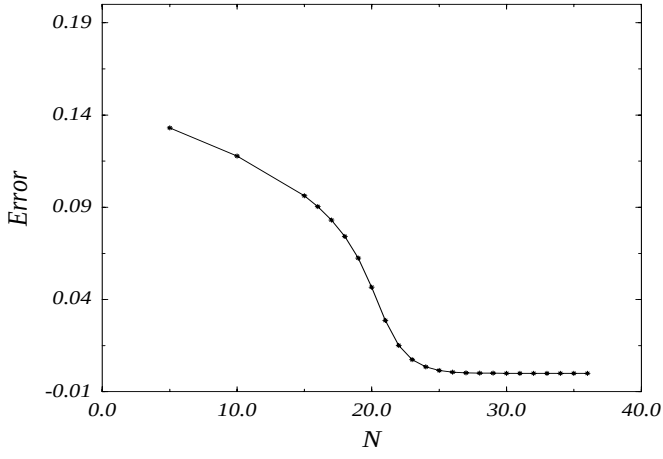


FIGURE 2.4. Approximation of the total solution by a truncated series along the wet surface $r \equiv 1$: Truncation error as a function of N for $k = 20$.

The above results are representative of a number of similar experiments, from which we conclude the truncation rule

$$N \approx 2k\,. \tag{2.1.32}$$

In the following, we give a heuristic outline for the decay of Bessel functions $J_n(x)$ as the index n grows for fixed x. We will see that the result corresponds well to numerical observations.

Analytical Outline of a Truncation Rule:

Returning to the expansion (2.1.23), we are interested in a truncation rule for the sum (2.1.31). Since $|\cos(n\theta)|$ is bounded independently of n, we seek a bound for $|J_n(kr)|$ as a function of $x = kr$ and n.

Consider *Kapteyn's inequality* (cf. Olver [103, p. 426]; we thank Markus Melenk for suggesting this reference),

$$|J_n(x)| \le \left(\frac{x}{n}\right)^n \left| \frac{\exp(\sqrt{n^2 - x^2})}{\left(1 + \sqrt{1 - \frac{x}{n}^2}\right)^n} \right| . \tag{2.1.33}$$

Assuming $n \ge x > 0$, the roots are real, and the absolute value signs on the right can be omitted. Using further $n^2 - x^2 \le n^2$, we get

$$|J_n(x)| \le \left(\frac{e\left(\frac{x}{n}\right)}{1 + \sqrt{1 - \frac{x}{n}^2}} \right)^n .$$

Hence, for fixed x the functions $J_n(x)$ are small if the expression within the parantheses is smaller than 1. It is easy to see that this expression decreases as (xn^{-1}) decreases. Solving the inequality

$$\frac{e\left(\frac{x}{n}\right)}{1+\sqrt{1-\frac{x}{n}^2}} < 1$$

for (xn^{-1}) yields

$$n > \frac{e^2+1}{2e}x \approx 1.543x\,.$$

Replacing x with kr, we conclude that for fixed values of kr, the contributions of the $J_n(kr)$ decay quickly with increasing n once

$$n > 1.6kr\,. \tag{2.1.34}$$

2.2 References from Functional Analysis

Modern numerical methods are based on the variational or weak formulations of boundary value problems. The natural function spaces for weak forms of differential operators are the Sobolev spaces. If the operators are linear, these spaces are also Hilbert spaces. We give the basic definitions in this section.

2.2.1 Norm and Scalar Product

Let V be a complex linear space. V is called a normed space if any element $v \in V$ is uniquely mapped to a real number $\|v\| \geq 0$ with the properties

$$\begin{aligned} \|v\| &= 0 \;\Rightarrow\; v = 0\,, \\ \|u+v\| &\leq \|u\| + \|v\|\,, \qquad \forall u, v \in V\,, \\ \|\alpha v\| &= |\alpha|\|v\|\,, \qquad \forall v \in V, \alpha \in \mathbf{C}\,. \end{aligned} \tag{2.2.1}$$

A space V is equipped with a scalar product if a map $V \times V \to \mathbf{C}$ with the properties

$$\begin{aligned} (v,v) &\geq 0\,, \qquad (v,v) = 0 \;\Rightarrow\; v = 0\,, \\ (\alpha u + v, w) &= \alpha(u,w) + (v,w)\,, \qquad \forall u, v, w \in V,\, \alpha \in \mathbf{C}\,, \\ (u,v) &= \overline{(v,u)}\,, \qquad \forall u, v \in V \end{aligned} \tag{2.2.2}$$

is defined on V. If a scalar product is defined on a linear space V, then V is normed by the induced norm $\|\cdot\|_V = (\cdot,\cdot)_V^{1/2}$.

A sequence $\{v_n\} \subset V$ in a normed linear space V is called a Cauchy sequence if

$$\sup_{m,n\geq k} \|v_n - v_m\|_V \to 0 \quad \text{for } k \to \infty\,.$$

A normed linear space V is called complete if any Cauchy sequence $\{v_n\} \subset V$ converges to an element $v \in V$; i.e., if there exists an element $v \in V$ such that $\lim_{n\to\infty} \|v - v_n\|_V = 0$.

We will frequently use the Cauchy–Schwarz inequality. Let V be a linear space equipped with a scalar product. Then any $u, v \in V$ satisfy

$$|(u,v)| \leq \|u\|\|v\|\,, \tag{2.2.3}$$

where $\|\cdot\|$ is the norm that is induced by the scalar product. Indeed, consider

$$\begin{aligned} 0 &\leq (\alpha u + \beta v, \alpha u + \beta v) \\ &= |\alpha|^2\|u\|^2 + \alpha\bar{\beta}(u,v) + \beta\bar{\alpha}(v,u) + |\beta|^2\|v\|^2 \end{aligned}$$

and take

$$\alpha = -\frac{\overline{(u,v)}}{\|u\|}\,, \quad \beta = \|u\|\,,$$

assuming that $u \neq 0$ (the statement is straightforward for $u = 0$). Thus

$$0 \leq -|(u,v)|^2 + \|u\|^2\|v\|^2\,,$$

which is equivalent to (2.2.3).

2.2.2 Hilbert Spaces

A linear space V is called a Hilbert space if it is equipped with a scalar product $(\cdot,\cdot)_V$ and is complete with respect to the induced norm $\|\cdot\|_V$.

Example 2.5. Consider the interval $(0,1) \subset \mathbf{R}$ and define the space

$$L^2(0,1) := \{f : (0,1) \to \mathbf{C},\ \int_0^1 |f(x)|^2 dx < \infty\} \tag{2.2.4}$$

of square-integrable functions. For example, the function $f = x$ lies in $L^2(0,1)$, whereas $g = x^{-1}$ does not. The operation

$$(f,g) := \int_0^1 f(x)\overline{g(x)}dx$$

defines a scalar product. In fact, $L^2(0,1)$ is a Hilbert space[3] with the norm

$$\|f\| = \left(\int_0^1 |f(x)|^2\, dx\right)^{1/2}\,.$$

[3]Completeness is proven in the Fiszer–Riesz theorem; cf. [105, 28].

For integer $m > 0$, define the subspace

$$H^m(0,1) := \left\{ f \in L^2(0,1) : \partial^i f \in L^2(0,1),\, i = 0,1,\ldots,m \right\}, \qquad (2.2.5)$$

where $\partial^i f$ are the weak derivatives of the function f. A function $f' = \partial f$ is called the weak derivative of f if

$$\int_{-\infty}^{\infty} f'(x)\varphi(x)dx = \int_{-\infty}^{\infty} f(x)\varphi'(x)dx$$

for all test functions φ that are differentiable in the classical sense and vanish, together with all their derivatives, at $\pm\infty$. The higher derivatives $\partial^i, i \geq 2$, are defined similarly, the derivative for $i = 0$ is formally identified with the function f itself. One easily verifies that a scalar product on H^m is defined by

$$(f,g)_m = \sum_{i=0}^{m} \int_0^1 \partial^i f(x)\overline{\partial^i g(x)}dx\,,$$

inducing the norm $\|f\|_m = (f,f)_m^{1/2}$. The subspaces $H^m(0,1) \subset L^2(0,1)$ are again Hilbert spaces. They are also Sobolev spaces, namely, the special case $p = 2$ of the general Sobolev spaces $W^{m,p}$. In particular, the Sobolev space $H^0(0,1)$ is identical with $L^2(0,1)$.

For $f \in H^m(0,1)$, we will also work with the seminorm

$$|f|_m := \left(\int_0^1 |\partial^m f(x)|^2 dx \right)^{1/2}.$$

Seminorms are linear maps that satisfy $(2.2.1)_{2,3}$ but do not, in general, satisfy $(2.2.1)_1$. For example, consider $H^1(0,1) \ni f \equiv 1$. Then $|f|_1 = 0$ but $f \neq 0$.

Remark 2.6. The definitions that have been given here for a one-dimensional interval can be generalized to domains $\Omega \subset \mathbf{R}^n$. This generalization requires a detailed discussion of the regularity of the domain, which depends on the smoothness of its boundary. For a rigorous and systematic treatment of these points, we refer to Hackbusch [62, Section 6.2]. In our present applications, we deal with regular domains only.[4] In our review of the theory, we will frequently use the expression "sufficiently regular domain", which means that the assumptions of the fundamental trace theorem (see the reference at the end of this section) are satisfied. Convex domains with piecewise smooth boundaries (curvilinear polygons) are an example. A domain Ω is called convex if for any $x_1, x_2 \in \Omega$ also $\Omega \ni x(t) = x_1 + t(x_2 - x_1)$ for all $t \in (0,1)$.

[4] The behavior of solutions to the Maxwell equation in domains with singularities has been discussed recently in [40].

2.2.3 Sesquilinear Forms and Linear Operators

We will write the variational (weak) formulation of a boundary value problem in the form

$$\begin{cases} \text{Find } u \in V_1 : \\ \qquad b(u,v) = f(v)\,, \qquad \forall v \in V_2\,, \end{cases} \tag{2.2.6}$$

where $\forall$ means "for all," V_1, V_2 are normed linear spaces (called the trial and the test space, respectively), b is a bilinear (or sesquilinear) form $V_1 \times V_2 \to \mathbf{C}$, and f is a linear (or antilinear) functional $V_2 \to \mathbf{C}$. We give the precise definitions in the remainder of this section.

The map $b(\cdot,\cdot)$ is called a bilinear form if it is linear in both arguments. If the form $b(\cdot,\cdot)$ is linear in the first and antilinear in the second argument, namely if

$$\begin{aligned} b(\alpha(u_1+u_2),v) &= \alpha(b(u_1,v)+b(u_2,v))\,, \\ b(u,\alpha(v_1+v_2) &= \bar{\alpha}(b(u,v_1)+b(u,v_2))\,, \end{aligned}$$

then it is called sesquilinear. The adjoint form b^* of a sesquilinear form $b : V \times V \to \mathbf{C}$ is defined as

$$b^*(u,v) := \overline{b(v,u)} \qquad \forall u, v \in V\,. \tag{2.2.7}$$

The form $b : V \times V \to \mathbf{C}$ is called self-adjoint if $b(\cdot,\cdot) = b^*(\cdot,\cdot)$.

A sesquilinear form $b : V_1 \times V_2 \to \mathbf{C}$ is called bounded if there exists a constant M such that

$$|b(u,v)| \le M\|u\|_{V_1}\|v\|_{V_2}$$

for all $\{u,v\} \in V_1 \times V_2$.

Let V_1, V_2 be normed linear spaces. A map $A : V_1 \to V_2$ is called a linear operator if

$$A(\alpha u + \beta v) = \alpha Au + \beta Av, \quad \forall u, v \in V_1,\ \alpha, \beta \in \mathbf{C}\,.$$

The linear operators $V_1 \to V_2$ form the linear space $\mathcal{L}(V_1, V_2)$. The operator A is bounded if there exists a real constant M such that

$$\|A(u)\|_{V_2} \le M\|u\|_{V_1} \tag{2.2.8}$$

for all $u \in V_1$. The smallest of the possible bounds is called the norm of the operator A. The bounded operators form a normed linear space $\mathcal{B}(V_1, V_2) \subset \mathcal{L}(V_1, V_2)$. A linear operator is bounded if and only if it is continuous; hence both notions can be used equivalently.

In the special case that V_2 is the real or complex number space, the operators are called functionals. The set of bounded linear functionals on a normed space V forms the dual space V' equipped with the norm

$$\|f\|_{V'} := \sup_{\|u\|=1} |f(u)| = \sup_{0 \ne u \in V} \frac{|f(u)|}{\|u\|_V}\,. \tag{2.2.9}$$

In our applications, we will work with the space V^* of antilinear functionals with the property $f(\alpha u) = \bar{\alpha} f(u)$. The norm is defined as in (2.2.9). The complex conjugation $f \to \bar{f}$ establishes a one-to-one map $V' \leftrightarrow V^*$. Let $f \in V^*$, $u \in V$. The complex number $f(u)$ can be equivalently written in the form of the dual product $\langle f, u \rangle_{V^* \times V}$. For example, the scalar product in a complex-valued Hilbert space is a sesquilinear form, and for any fixed $u \in V$ the map $v \to (u, v)$ defines an antilinear functional $f_u \in V^*$. Conversely, any functional $f \in V^*$ can be uniquely mapped to some $u_f \in V$. By the Riesz representation theorem (cf., e.g., Yosida [120, p. 90]), there exists for any Hilbert space V a one-to-one correspondence $V^* \leftrightarrow V$ mapping each functional $f \in V^*$ to a $u_f \in V$ such that $f(v) = (u_f, v)$, $v \in V$. The Riesz map $R : f \to u_f$ is a linear operator.

The dual spaces $(H^m)^*$ to the Sobolev spaces H^m are also denoted by H^{-m}. Since $H^m \subset L^2$, any functional that is defined on L^2 is automatically defined also on H^m for $m \geq 0$ (and hence lies in H^{-m}). Therefore we have

$$H^m(\Omega) \subset L^2(\Omega) \equiv \left(L^2(\Omega)\right)^* \subset H^{-m}(\Omega)\,,$$

where we have identified $L^2(\Omega) \equiv \left(L^2(\Omega)\right)^*$ by the Riesz map. This relation is called a Gelfand triple (cf. Hackbusch [62, Section 6.3.3.]). One can show that the embeddings are continuous and dense and that the inner product $(\cdot, \cdot)_{L^2}$ is a continuous extension of the dual product $\langle \cdot, \cdot \rangle_{H^{-m} \times H^m}$ [62, Section 6.3]. Therefore, in a Gelfand triple, instead of the dual pairing $f(v) = \langle f, v \rangle_{H^{-m} \times H^m}$ we can write equivalently $(f, v)_{L^2}$.

2.2.4 Trace of a Function

For the precise treatment of boundary value problems in Hilbert spaces one needs the notion of the trace of a function. Let Ω be a bounded domain with boundary Γ. If a function u is continuous on the closed domain $\overline{\Omega} = \Omega \cup \Gamma$ then the restriction $u|_\Gamma$ (called the trace of u on Γ) is well-defined. However, the Sobolev spaces are defined on open domains and the functions in these spaces are, in general, not continuous. In this case, the trace of a function is defined as a linear map from the Sobolev space on the domain Ω to a Sobolev space on the boundary Γ.

Considering the Sobolev space $H^m(\Omega)$, $m \geq 1$, we define the trace γu of $u \in H^m(\Omega)$ on Γ as a linear operator

$$\gamma : H^m(\Omega) \to H^{m-1/2}(\Gamma)\,.$$

This operator is defined in such a way that $\gamma u = u|_\Gamma$ for all $u \in H^m(\Omega) \cap C^0(\overline{\Omega})$. Note that $m - 1/2$ is not an integer. For the definition of Sobolev spaces H^s with index $s \in \mathbf{R}$, see Hackbusch [62, Section 6.2.4]. The trace theorem (see Hackbusch [62, Theorem 6.2.40]) states that, provided certain assumptions on the regularity of the domain Ω are satisfied, γ is a bounded

operator. This is expressed by the *trace inequality*

$$\|\gamma u\|_{H^{m-1/2}(\Gamma)} \leq C_\gamma \|u\|_{H^m(\Omega)}\,, \tag{2.2.10}$$

where (cf. (2.2.8))

$$C_\gamma = \|\gamma\|_{\mathcal{B}(H^m(\Omega),H^{m-1/2}(\Gamma))}\,.$$

Furthermore, the operator γ is surjective. Any $w \in H^{m-1/2}(\Gamma)$ can be extended onto the domain Ω, and the extension operator ("lifting") is also bounded.

2.3 Variational Formulation of Helmholtz Problems

2.3.1 Helmholtz Problems on Bounded Domains

For Helmholtz problems given on a bounded domain Ω, the natural trial and test spaces are the Sobolev spaces

$$H^1(\Omega) = \{v \mid \|\nabla v\|^2 + \|v\|^2 < \infty\}\,,$$

where $\|\cdot\|$ is the L^2-norm. As an example, consider the mixed boundary value problem

$$\begin{array}{rcll} -\Delta u - k^2 u &=& 0 & \text{in} \quad \Omega\,, \\ \partial_\nu u + \beta u &=& g & \text{on} \quad \Gamma\,, \end{array} \tag{2.3.1}$$

where β is a complex constant and ∂_ν denotes the exterior normal derivative. We derive a weak formulation by the method of weighted residuals. For some function $u \in H^1(\Omega)$, we call $H^{-1}(\Omega) \ni r_\Omega(u) := -\Delta u - k^2 u$ the domain residual and $H^{-1/2}(\Gamma) \ni r_\Gamma(u) := \partial_\nu u + \beta u - g$ (here u is understood in the sense of trace) the boundary residual. Let $v \in H^1(\Omega)$ be an arbitrary test function and $w \in H^{1/2}(\Gamma)$ its trace. We demand that the sum of the weighted residuals vanish:

$$\langle r_\Omega(u), v\rangle_{H^{-1}\times H^1} + \langle r_\Gamma(u), w\rangle_{H^{-1/2}\times H^{1/2}} = 0\,.$$

Identifying the dual pairings with the L^2 inner products and integrating by parts, we arrive at the variational problem:

$$\left\{ \begin{array}{l} \text{Find } u \in H^1(\Omega): \\ \qquad b(u,v) = (g,v)_{L^2(\Gamma)}\,, \qquad \forall v \in H^1(\Omega)\,, \end{array} \right. \tag{2.3.2}$$

with

$$b(u,v) = \int_\Omega (\nabla u \nabla \bar{v} - k^2 u \bar{v})\, dV + \beta \int_\Gamma u \bar{v}\, dS \tag{2.3.3}$$

and

$$(g,v)_{L^2(\Gamma)} = \int_\Gamma g \bar{v}\, dS\,. \tag{2.3.4}$$

Alternatively, one can use the test function $\bar{v}$ to obtain the bilinear form

$$\tilde{b}(u,v) = \int_\Omega (\nabla u \nabla v - k^2 uv) dV + \beta \int_\Gamma uv \, dS \,.$$

If a solution of the variational problem (2.3.2) exists, we say that the problem is weakly solvable and call the solution u a variational or weak solution of the boundary value problem (2.3.1).

2.3.2 Helmholtz Problems on Unbounded Domains

For the exterior problem, the appropriate choice of test and trial spaces in the weak formulation is less obvious since integration is carried out over the unbounded domain Ω^+ (exterior of the scatterer Ω). Assume for simplicity that we are computing the rigid scattering of a wave from the unit sphere. Then the normal derivative of the pressure is prescribed on the boundary $\Gamma = \{r \equiv 1, \theta, \phi\}$: we seek radiating solutions $p(r,\theta,\phi)$ of the Helmholtz equation satisfying a Neumann boundary condition

$$\frac{\partial p}{\partial r}(1,\theta,\phi) = g(\theta,\phi) \,.$$

The Sommerfeld condition is given on a sphere S_R with radius $R \to \infty$. Multiplying the Helmholtz equation in the bounded domain $\Omega_S := \{1 \leq r \leq R, \theta, \phi\}$ with a test function q, we obtain as before

$$b(p,q) := \int_{\Omega_S} (\nabla p \nabla \bar{q} - k^2 p \bar{q}) dV - ik \int_{S_R} p \bar{q} dS = \int_\Gamma g \bar{q} dS \,.$$

From the Atkinson–Wilcox expansion, we expect that the solution depends asymptotically on r as

$$p \simeq \frac{e^{ikr}}{r} := f(r) \,.$$

Hence one naturally requires that the function f be in the trial space. However, the L^2 inner product

$$(f,f) = \int_{\Omega^+} \frac{e^{ikr}}{r} \frac{e^{-ikr}}{r} dV = 4\pi \int_1^\infty \frac{1}{r^2} r^2 dr$$

is not finite.

One way to circumvent this problem is to measure the trial functions in weighted norms. Defining the weighted inner product

$$(p,q)_w = \int_{\Omega^+} w \, p \bar{q} \, dV \,, \qquad w := r^{-2} \,, \tag{2.3.5}$$

one easily confirms that both $(f,f)_w$ and $(f_{,r}\,, f_{,r}\,)_w$ exist. Accordingly, we demand that the trial functions satisfy

$$\|p\|_{1,w} := \left((p,p)_w + (\nabla p, \nabla p)_w\right)^{1/2} < \infty \,. \tag{2.3.6}$$

Now, if the trial functions satisfy (2.3.6), then the integrals $\int p\bar{q}$ and $\int \nabla p \nabla \bar{q}$ are well defined if test functions q are such that

$$(q,q)_{w*},\ (\nabla q, \nabla q)_{w^*} < \infty$$

with the inner product

$$(p,q)_{w*} := \int_{\Omega^+} w^* p\bar{q}\, dV\,, \qquad w^* := r^2\,. \tag{2.3.7}$$

Indeed, by the Cauchy–Schwarz inequality,

$$\left|\int_{\Omega^+} p\bar{q}\, dV\right| = \left|\int_{\Omega^+} (w^{\frac{1}{2}} p)\,(w^{*\frac{1}{2}}\bar{q})\, dV\right| \le (p,p)_w^{1/2}(q,q)_{w^*}^{1/2} < \infty\,.$$

Thus the trial and test functions, respectively, lie in the weighted Sobolev spaces

$$H^1_w(\Omega^+) = \{u\,:\,\|u\|_{1,w} < \infty\}\,, \quad H^1_{w*}(\Omega^+) = \{u\,:\,\|u\|_{1,w*} < \infty\}\,,$$

where the norm $\|\cdot\|_{1,w*}$ is defined using the inner product $(\cdot,\cdot)_{w*}$ in (2.3.6). Note that the trial and test spaces are not identical. The functions in the trial space are of order r^{-1} (or lower), whereas the functions in the test space are at most of order r^{-3}.

This property of the test functions does, however, prohibit inclusion of the Sommerfeld condition into the variational formulation. Indeed, one easily confirms that for arbitrary $p \in H^1_w(\Omega^+)$, $q \in H^1_{w*}(\Omega^+)$, the integral $\int_{S_R} p\bar{q} dS$ approaches zero as $R \to \infty$. Hence, in the present setting, one cannot test whether or not a trial function satisfies the Sommerfeld condition. However, the far-field term in the variational formulation is crucial for well-posedness of the weak formulation (cf. Example 2.23 below). This obstacle is overcome by including the Sommerfeld condition into the definition of the trial space, thus excluding the eigenfunctions as possible solutions. Following Leis [87, Section 4.4], one redefines the trial space as

$$H^{1+}_w(\Omega^+) = \left\{p\,:\,\|p\|^+_{1,w} < \infty\right\}\,, \tag{2.3.8}$$

with the norm

$$\|p\|^+_{1,w} := \left(\|p\|^2_{1,w} + \int_{\Omega^+}\left|\frac{\partial p}{\partial r} - ikp\right|^2 dV\right)^{1/2}. \tag{2.3.9}$$

Remark 2.7. The precise definition of the spaces introduced by Leis involves the notion of completion, which we have not introduced here. For instance, the space $H^{1+}_w(\Omega^+)$ is really the completion of the space $H^1_w(\Omega^+)$ in the norm $\|\cdot\|^+_{1,w}$.

With these preliminaries, weak radiating solutions of the exterior Neumann problem are sought from the variational problem:

$$\begin{cases} \text{Find } p \in H^{1+}_{w}(\Omega^+) : \\ \qquad \displaystyle\int_{\Omega^+} (\nabla p \cdot \nabla \overline{q} - k^2\, p\, \overline{q})\; dV = (g, q)\,, \qquad \forall q \in H^1_{w^*}(\Omega^+)\,, \end{cases} \tag{2.3.10}$$

where (g, q) denotes the L^2 inner product on the wet surface Γ. The variational problem (2.3.10) is of the form (2.2.6), where $V_1 = V_\rho$ and $V_2 = V_{\rho^*}$ are different weighted Sobolev spaces with weights ρ, ρ^*, respectively.

Remark 2.8. Due to the different weighting of test and trial functions, numerical methods based on Leis' variational formulation lead to non-symmetric systems of algebraic equations. For the purpose of numerical computations, it is also possible to state *on finite-dimensional spaces* variational principles with equal weights in such a way that integration in the corresponding variational forms is well-defined on Ω^+. We will return to this topic in the context of infinite elements in Section 3.5.

2.3.3 Weak Formulation for Solid–Fluid Interaction

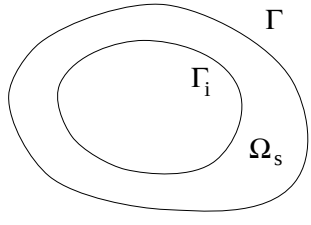


FIGURE 2.5. Solid domain.

Assume that the obstacle is a solid $\Omega_{\rm s}$, possibly with a traction-free interior boundary $\Gamma_{\rm i}$, see Fig. 2.5. To obtain a variational expression in the solid, we test equations (1.2.1) with the complex conjugate of a vector-function $\mathbf{v}$ and integrate to get

$$-\int_{\Omega_{\rm s}} (\nabla \cdot \sigma + \rho_{\rm s}\omega^2 \mathbf{u}) \cdot \bar{\mathbf{v}} dV = \int_{\Omega_{\rm s}} \mathbf{f} \cdot \bar{\mathbf{v}} dV\,. \tag{2.3.11}$$

In expanded form, the integral on the left reads

$$\begin{aligned} -\int_{\Omega_{\rm s}} \big(&(\sigma_{xx,x} + \sigma_{xy,y} + \sigma_{xz,z})\, \bar{v}_x + (\sigma_{yx,x} + \sigma_{yy,y} + \sigma_{yz,z})\, \bar{v}_y \\ &+ \quad (\sigma_{zx,x} + \sigma_{zy,y} + \sigma_{zz,z})\bar{v}_z + \rho_{\rm s}\omega^2 \left(u_x \bar{v}_x + u_y \bar{v}_y + u_z \bar{v}_z\right)\big)\, dV\,. \end{aligned}$$

To the components in the first part we apply the identity (with the summation convention, cf. Section 1.2.)

$$(\mathrm{div}\sigma_i)\bar{v}_i = \mathrm{div}(\sigma_i\bar{v}_i) - \sigma_i \cdot \nabla\bar{v}_i \,,$$

where the components of the stress tensor are denoted by

$$\sigma_i = \{\sigma_{ix}, \sigma_{iy}, \sigma_{iz}\} \,, \quad i = x, y, z \,.$$

By the Gauss theorem,

$$\int_{\Omega_s} \mathrm{div}(\sigma_i\bar{v}_i) = \int_\Gamma (\sigma_i\bar{v}_i)\cdot\mathbf{n}\, dS = \int_\Gamma (\sigma_i\cdot\mathbf{n})\bar{v}_i dS \,.$$

For equilibrium on Γ, the stress resultants $\sigma_i \cdot \mathbf{n}$ must be equal to the components of the exterior traction vector $\mathbf{T}$. Inserting the results into (2.3.11) yields

$$\int_{\Omega_s} \left(\sigma_i \cdot \nabla\bar{v}_i - \rho_s\omega^2\mathbf{u}\cdot\bar{\mathbf{v}}\right)\, dV = \int_\Gamma \mathbf{T}\cdot\bar{\mathbf{v}}\, dS + \int_{\Omega_s} \mathbf{f}\cdot\bar{\mathbf{v}} dV \,. \qquad (2.3.12)$$

The scalar product $\mathbf{T}\cdot\bar{\mathbf{v}}$ is invariant with respect to orthogonal transforms, and hence

$$\int_\Gamma (T_x\bar{v}_x + T_y\bar{v}_y + T_z\bar{v}_z)\, dS = \int_\Gamma (T_n\bar{v}_n + T_{t_1}\bar{v}_{t_1} + T_z\bar{v}_{t_2})\, dS \,, \qquad (2.3.13)$$

where $\{n, t_1, t_2\}$ is the local coordinate system shown in Fig. 1.5. From the equilibrium conditions (1.3.2) and (1.3.3), we have $\mathbf{T}\cdot\bar{\mathbf{v}} = -(p + p_{\mathrm{inc}})v_n$. We thus finally arrive at

$$\begin{aligned} &\int_{\Omega_s} \left(\sigma_i \cdot \nabla\bar{v}_i - \rho_s\omega^2\mathbf{u}\cdot\bar{\mathbf{v}}\right)\, dV + \int_\Gamma p(\bar{\mathbf{v}}\cdot\mathbf{n})dS = \\ &- \int_\Gamma p_{\mathrm{inc}}(\bar{\mathbf{v}}\cdot\mathbf{n})dS + \int_{\Omega_s} \mathbf{f}\cdot\bar{\mathbf{v}}\, dV \,. \end{aligned} \qquad (2.3.14)$$

Similarly, multiplying the fluid equation (1.3.9) by a test function $\bar{q}$ and integrating by parts, we obtain

$$0 = -\int_{\Omega^+} (\Delta p + k^2 p)\bar{q} dV = \int_{\Omega^+} (\nabla p\nabla\bar{q} - k^2 p\bar{q}) dV + \int_\Gamma \frac{\partial p}{\partial n}\bar{q} dS \,,$$

where we have taken into account that the exterior solid normal points *into* the fluid region. Inserting the coupling condition (1.3.7) now leads to

$$\int_{\Omega^+} (\nabla p\nabla\bar{q} - k^2 p\bar{q})\, dV + \rho_f\omega^2 \int_\Gamma (\mathbf{u}\cdot\mathbf{n})\bar{q} dS = \int_\Gamma \frac{\partial p_{\mathrm{inc}}}{\partial n}\bar{q} dS \,. \qquad (2.3.15)$$

Finally, equations (2.3.14) and (2.3.15) can be added up (after scaling (2.3.14) for symmetry) to give the coupled equation

$$\begin{aligned} & \int_{\Omega^+} (\nabla p \cdot \nabla \bar{q} - k^2\, p\, \bar{q})\, dV \\ + \quad & \rho_{\mathrm{f}}\omega^2 \left(\int_\Gamma (\mathbf{u}\cdot\mathbf{n})\, \bar{q}\, dS + \int_\Gamma p(\bar{\mathbf{v}}\cdot\mathbf{n})\, dS \right) \\ + \quad & \rho_{\mathrm{f}}\omega^2 \int_{\Omega_{\mathrm{s}}} \left[\sigma_i \cdot \nabla \bar{v}_i - \rho_{\mathrm{s}}\omega^2 \mathbf{u}\cdot\bar{\mathbf{v}} \right] dV \\ = \quad & \rho_{\mathrm{f}}\omega^2 \left(-\int_\Gamma p_{\mathrm{inc}}(\bar{\mathbf{v}}\cdot\mathbf{n})\, dS + \int_{\Omega_{\mathrm{s}}} \mathbf{f}\cdot\bar{\mathbf{v}} dV \right) \\ + \quad & \int_\Gamma \frac{\partial p_{\mathrm{inc}}}{\partial n} \bar{q} dS\,. \end{aligned} \tag{2.3.16}$$

We now define the Sobolev spaces $\mathcal{H} = H^1(\Omega_{\mathrm{s}}) \times H^{1+}_w(\Omega^+)$, $\mathcal{H}^* = H^1(\Omega_{\mathrm{s}}) \times H^1_{w^*}(\Omega^+)$. Then we pose the variational problem

$$\left\{ \begin{array}{l} \text{Find } \mathcal{U} := (\mathbf{u}, p) \in \mathcal{H} : \\ \qquad \mathcal{B}(\mathcal{U}, \mathcal{V}) = (\mathcal{F}, \mathcal{V})\,, \qquad \forall \mathcal{V} := (\mathbf{v}, q) \in \mathcal{H}^*\,, \end{array} \right. \tag{2.3.17}$$

where the form $\mathcal{B}$ and the right-hand side are given by (2.3.16).

Example 2.9. Let us outline the weak formulation for the one-dimensional problem introduced in Example 1.6. Assuming that the radiation condition is imposed at $x_0 > l$, we have $\Omega_{\mathrm{s}} = (0, l)$ and $\Omega^+ = (l, x_0)$. Then one solves $b(u, v) = \rho_{\mathrm{f}}\omega^2 (f, v)$ with

$$\begin{aligned} b(\{u,p\},\{v,q\}) = \rho_{\mathrm{f}}\omega^2 & \int_0^l \left(E \frac{du}{dx}\frac{d\bar{v}}{dx} - \rho_{\mathrm{s}}\omega^2 u\bar{v} \right) dx \\ & + \rho_{\mathrm{f}}\omega^2 \left(p(l)\bar{v}(l) + u(l)\bar{q}(l) \right) \\ & + \int_l^{x_0} \left(\frac{dp}{dx}\frac{d\bar{q}}{dx} - k^2 p\bar{q} \right) dx - ikp(x_0)\bar{q}(x_0)\,, \end{aligned}$$

and

$$(f, v) = \int_0^l f\bar{v} dx\,.$$

Equivalently, one may just solve $b_0(u, v) = (f, v)$ with

$$b_0(u,v) = \int_0^l \left(E\frac{du}{dx}\frac{d\bar{v}}{dx} - \rho_{\mathrm{s}}\omega^2 u\bar{v} \right) dx - ik\rho_{\mathrm{f}}c^2 u(l)\bar{v}(l)\,,$$

imposing weakly the Robin boundary condition (1.3.12).

2.4 Well-Posedness of Variational Problems

A boundary value problem is called well-posed if, for a given class of data, the solution exists, is unique, and depends continuously on the data (i.e., if it is stable). Since the FEM is based on the variational formulation of boundary value problems, it is of fundamental importance to know whether the problem is weakly solvable and if the solution is unique. Stability is crucial for the numerical solution of a variational problem. For if a small error in the data can cause a large error in the solution, a numerical method may converge very slowly or not at all.

We address these fundamental questions for the variational forms that arise from the Helmholtz equation. These forms are, in general, not positive definite. Existence and uniqueness can be concluded alternatively from two generalizations of the positive definite case, namely, the existence theory for forms that satisfy an inf–sup condition or the theory for forms that satisfy a Gårding inequality. We therefore first discuss the positive definite (V-elliptic) case and then proceed to the generalization for indefinite forms.

2.4.1 Positive Definite Forms

A sesquilinear form $a : V \times V \to \mathbf{C}$ on a Hilbert space V is called V-elliptic (positive definite) if there exists $\alpha > 0$ such that

$$|a(u,u)| \geq \alpha \|u\|_V^2$$

for all $u \in V$. Existence and uniqueness of solutions for positive definite problems is established by the Lax–Milgram theorem.

Theorem 2.10.(Lax–Milgram). *Assume that a sesquilinear form $a : V \times V \to \mathbf{C}$, defined on a Hilbert space V, satisfies*

1. *Continuity:*

$$\exists\, M > 0 : \qquad |a(u,v)| \leq M \|u\|_V \|v\|_V, \qquad \forall u, v \in V\,, \tag{2.4.1}$$

2. *V-Ellipticity:*

$$\exists\, \alpha > 0 : \qquad \alpha \|u\|_V^2 \leq |a(u,u)|, \qquad \forall u \in V\,, \tag{2.4.2}$$

and let f be a bounded linear functional defined on V. Then there exists a unique element $u_0 \in V$ such that

$$a(v, u_0) = (v, f), \qquad \forall v \in V\,. \tag{2.4.3}$$

For the proof, see Yosida [120, p. 92].

Remark 2.11. Note that now f is associated with a linear functional $(\cdot, f)$. Equivalently (just take the conjugate on both sides of (2.4.3)), one can state that the dual problem

$$a^*(u, v) = (f, v)$$

has a unique solution. The problem $a(u, v) = (f, v)$ has a unique solution for all antilinear functionals $f \in V^*$.

The ellipticity condition (2.4.2) implies immediately that the solution u_0 is bounded by the data f, measured in the norm of the dual space V^*. Indeed,

$$\|f\|_{V^*} = \sup_{0 \neq v \in V} \frac{|f(v)|}{\|v\|_V} = \sup_{0 \neq v \in V} \frac{|a(u_0, v)|}{\|v\|_V} \geq \frac{|a(u_0, u_0)|}{\|u_0\|_V} \geq \alpha \|u_0\|_V \,,$$

and hence

$$\|u_0\|_V \leq \frac{1}{\alpha} \|f\|_{V^*} \,. \tag{2.4.4}$$

Further, letting $v = u$ in the continuity condition, we get

$$\alpha \|u\|_V^2 \leq |a(u, u)| \leq M \|u\|_V^2 \qquad \forall u \in V \,. \tag{2.4.5}$$

Thus the *energy norm*

$$|\!|\!| u |\!|\!| := (|a(u, u)|)^{1/2} \tag{2.4.6}$$

induced by the V-elliptic form $a(\cdot, \cdot)$ is equivalent to the norm $\| \cdot \|_V$.

Remark 2.12. If the form a is self-adjoint, then $a(u, u)$ is real, and the absolute values in (2.4.2) and (2.4.5) can be omitted.

Example 2.13. (Poisson equation). Let Ω be a bounded convex domain with a piecewise smooth boundary Γ and consider the Poisson equation with Dirichlet boundary conditions

$$\begin{aligned} \Delta u &= f \qquad \text{on } \Omega \,, \\ u &= 0 \qquad \text{on } \Gamma \,, \end{aligned}$$

where f is a complex-valued function. The corresponding sesquilinear form

$$a(u, v) = \int_\Omega \nabla u \nabla \bar{v} dV$$

is defined on the Hilbert subspace $V = H_0^1(\Omega) \subset H^1(\Omega)$ containing all H^1-functions that vanish on Γ. For these functions, one can show the *Poincaré* inequality

$$\|u\| \leq C \|\nabla u\| \,, \tag{2.4.7}$$

where C is a positive constant and $\|\cdot\|$ denotes the L^2-norm. V-ellipticity then easily follows with

$$a(u,u) \geq \frac{1}{1+C^2}\|u\|_1^2\,,$$

where $\|u\|_1$ is the norm of the Sobolev space $H^1(\Omega)$.

Example 2.14. (Elasticity). Let Ω be as in the previous example and consider the bilinear form, cf. (2.3.12),

$$a(\mathbf{u},\mathbf{v}) = \int_\Omega \sigma_i \cdot \nabla v_i dV\,.$$

Introducing the material law (1.2.3) in the general form

$$\sigma_{ij} = E_{ijkl}e_{kl}$$

and using the strain–displacement relations (1.2.2), we can write equivalently

$$a(\mathbf{u},\mathbf{v}) = \int_\Omega E_{ijkl}e_{ij}(\mathbf{u})e_{kl}(\mathbf{v})dV = \int_\Omega E_{ijkl}\frac{\partial u_i}{\partial x_j}\frac{\partial v_k}{x_l}\,dV\,.$$

The form $a(\cdot,\cdot)$ is V-elliptic on the subspace $V \subset \left[H^1(\Omega)\right]^d$ of all functions satisfying a Dirichlet condition on a nonempty part $\Gamma_D \subset \Gamma$,

$$V = \left\{\mathbf{u} \in \left[H^1(\Omega)\right]^d\,,\ \mathbf{u} = 0 \quad \text{on}\,\Gamma_D \subset \Gamma\right\}\,.$$

The proof of the corresponding relation

$$|a(\mathbf{u},\mathbf{u})| \geq \alpha\|\mathbf{u}\|^2_{H^1(\Omega)} \qquad \forall \mathbf{u} \in V,$$

for some $\alpha > 0$, is based on the Korn inequality

$$\int_\Omega e_{ij}(\mathbf{u})e_{ij}(\mathbf{u})dV + \int_\Omega u_i u_i dV \geq \gamma\|\mathbf{u}\|^2_{H^1(\Omega)}\,,$$

with $\gamma > 0$. For details see, e.g., Sanchez Hubert–Sanchez Palencia [107].

2.4.2 The inf–sup Condition

Generally, the Helmholtz problem for large k is indefinite. Consider, as an example, the one-dimensional case with Dirichlet conditions. For $u, v \in H_0^1(0,1)$, the form

$$b(u,v) = \int_0^1 \left(u'v' - k^2uv\right)\,dx \tag{2.4.8}$$

is H_0^1-elliptic only if k is smaller than the minimal positive eigenvalue of the Laplace equation with Dirichlet boundary conditions, i.e., if $k < \pi$.

The following generalization of the Lax–Milgram theorem was shown by Babuška [7, p. 112].

Theorem 2.15. (Babuška). *Assume that a sesquilinear form $b : V_1 \times V_2 \to \mathbf{C}$ on Hilbert spaces V_1, V_2 satisfies*

1. *Continuity:*

$$\exists\, M > 0 : \qquad |b(u,v)| \le M \|u\|_{V_1} \|v\|_{V_2}, \qquad \forall u \in V_1,\, v \in V_2\,, \tag{2.4.9}$$

2. *inf–sup Condition:*

$$\exists\, \beta > 0 : \qquad \beta \le \sup_{0 \neq v \in V_2} \frac{|b(u,v)|}{\|u\|_{V_1} \|v\|_{V_2}}\,, \qquad \forall 0 \neq u \in V_1\,, \tag{2.4.10}$$

3. *"Transposed" inf–sup Condition:*

$$\sup_{0 \neq u \in V_1} |b(u,v)| > 0, \qquad \forall 0 \neq v \in V_2\,, \tag{2.4.11}$$

and let $f : V_2 \to \mathbf{C}$ be an antilinear bounded functional defined on V_2. Then there exists a unique element $u_0 \in V_1$ such that

$$b(u_0, v) = f(v), \qquad \forall v \in V_2\,.$$

The solution u_0 satisfies the bound

$$\|u_0\|_{V_1} \le \frac{1}{\beta} \|f\|_{V_2^*}\,. \tag{2.4.12}$$

Remark 2.16. Condition (2.4.10) is also called the Babuška–Brezzi condition. It is fundamental for a large class of indefinite problems, especially saddle-point problems, arising in mixed or hybrid FEM; see, e.g., Braess [30], Carey and Oden [33]. The real number

$$\beta = \inf_{0 \neq u \in V_1} \sup_{0 \neq v \in V_2} \frac{|b(u,v)|}{\|u\| \|v\|}$$

is called the inf–sup constant.

Remark 2.17. The Babuška theorem covers the case $V_1 \neq V_2$ and does not assume definiteness. If $V_1 = V_2$, we can compare the inf–sup constant β to the ellipticity constant α of the Lax–Milgram theorem by writing

$$\alpha = \inf_{u \in V} \frac{|a(u,u)|}{\|u\|^2}\,. \tag{2.4.13}$$

Obviously, since α is computed by taking the infimum also over the second argument, whereas for β the supremum is taken over this argument, we have always $\alpha \leq \beta$. This is also seen from

$$\alpha\|u\| \leq \frac{|a(u,u)|}{\|u\|} \leq \sup_{v\in V} \frac{|a(u,v)|}{\|v\|}\,, \qquad \forall u\,.$$

Since this holds for all $u \in V$, we have

$$\alpha \leq \inf_{u\in V} \sup_{v\in V} \frac{|b(u,v)|}{\|u\|\|v\|} = \beta\,.$$

In particular, it can occur that $\alpha = 0$ but still $\beta > 0$, i.e., a form may be not elliptic but satisfy the inf–sup condition. In fact, we will show in Chapter 4 that the Helmholtz variational forms of one-dimensional model problems satisfy the inf–sup condition with $\beta = Ck^{-1} > 0$.

Remark 2.18. Both α and β depend on the norm of the space V. For instance, considering a V-elliptic sesquilinear form $a(\cdot,\cdot)$, trivially $\alpha = 1$ in the energy norm $|\!|\!|u|\!|\!| = \sqrt{a(u,u)}$. Also, $\beta = 1$ in this norm. Indeed, applying the Cauchy–Schwarz inequality, we have

$$|a(u,v)| \leq \sqrt{|a(u,u))|}\sqrt{|(a(v,v)|}\,,$$

whence

$$\sup_{0\neq v\in V} \frac{|a(u,v)|}{|\!|\!|v|\!|\!|} \leq |\!|\!|u|\!|\!|\,.$$

Since equality holds for $v = u$, the supremum is exactly $|\!|\!|u|\!|\!|$. Thus

$$\beta = \inf_{u\in V} \frac{|\!|\!|u|\!|\!|}{|\!|\!|u|\!|\!|} = 1\,.$$

Remark 2.19. As already mentioned, the question of existence and uniqueness is always related to certain function spaces. In (conforming) FEM, the solution is sought on a finite-dimensional subspace $V_h \subset V$. Let us note here an important difference between definite and indefinite forms: The ellipticity property carries over from V to V_h, *whereas the inf–sup property does not*. Indeed, the infimum in condition (2.4.13) cannot decrease if it is taken on a subspace, whereas the supremum involved in the inf–sup constant, in general, decreases on a subspace. Consequently, the inf–sup condition may not be satisfied on the subspace.

Remark 2.20. Similar to the Lax–Milgram theorem, the Babuška theorem implies stability, and hence well-posedness. Conversely, requiring stable dependence

$$\|u\|_{V_1} \leq C\|f\|_{V_2^*} \tag{2.4.14}$$

leads in a natural way to the inf–sup condition as follows. Assume that for some fixed $f \in V_2^*$ we have obtained a solution $u_0 \neq 0$. Then, by the definition of the dual norm, (2.4.14) can be written as

$$\frac{1}{C} =: \beta \leq \frac{\|f\|_{V_2^*}}{\|u_0\|_{V_1}} = \sup_{0 \neq v \in V_2} \frac{|f(v)|}{\|u_0\|_{V_1}\|v\|_{V_2}} = \sup_{0 \neq v \in V_2} \frac{|a(u_0, v)|}{\|u_0\|_{V_1}\|v\|_{V_2}} .$$

This leads directly to the inf–sup condition if we now require that (2.4.14), and hence the above inequality, hold for all possible data $f \in V_2^*$ (or, equivalently, for all $u \in V_1$).

2.4.3 Coercive Forms

Let Ω be a bounded domain and consider the Hilbert space $V = H^1(\Omega)$. A sesquilinear form $b : V \times V \to \mathbf{C}$ is called V-coercive if it satisfies for all $u \in V$ the Gårding inequality

$$\Big| b(u,u) + C\|u\|^2_{L^2(\Omega)} \Big| \geq \alpha \|u\|^2_{H^1(\Omega)} \tag{2.4.15}$$

with positive constants C, α.

Remark 2.21. We specify here the general definition of V-coercivity for a Gelfand triple $V \subset H \subset V'$ (see Hackbusch [62, Section 6.5.13]) to the special case $H^1(\Omega) \subset L^2(\Omega) \subset H^{-1}(\Omega)$.

Remark 2.22. We can interpret (2.4.15) as a V-ellipticity property of the form $a(u,v) := b(u,v) + C(u,v)_{L^2(\Omega)}$.

Consider the general variational problem (2.2.6) with $V_1 = V_2 = H^1(\Omega)$ and assume that the sesquilinear form $b(u,v)$ is H^1-coercive. For a sufficiently regular domain[5] Ω, the embedding $H^1(\Omega) \subset L^2(\Omega)$ is compact. Then it can be shown (cf. Hackbusch [62, Theorem 6.5.15]) that the problem satisfies the Fredholm alternative: either equation (2.2.6) has a solution $u \in H^1(\Omega)$ for all f or there exists a nontrivial solution of the homogeneous problem (with $f \equiv 0$). Hence the existence of the solution follows if we can show uniqueness. We illustrate this in the following example of a one-dimensional Helmholtz problem.

Example 2.23. Let $\Omega = (0,1)$ and let $f \in L^2(0,1)$ be given. Consider the boundary value problem[6]

$$-u'' - k^2 u \;=\; f \qquad \text{on } \Omega = (0,1)\,,$$

[5] see Remark 2.6.

[6] to be precise, this problem is well-defined in the classical sense for functions $f \in C^0[0,1]$, $u \in C^2(0,1) \cap C^0[0,1]$.

$$\begin{aligned} u(0) &= 0\,, \\ u'(1) - i\alpha(k)u(1) &= 0\,, \end{aligned} \tag{2.4.16}$$

with some $\mathbf{R} \ni \alpha(k) > 0$. The corresponding sesquilinear form is

$$b(u,v) = \int_0^1 \left(u'\bar{v}' - k^2 u\bar{v}\right)\, dx - i\alpha u(1)\bar{v}(1)\,. \tag{2.4.17}$$

The test and trial spaces are $V_1 = V_2 = H^1_{(0}(0,1)$, where

$$H^1_{(0}(0,1) := \left\{u \in H^1(0,1);\, u(0) = 0\right\} \tag{2.4.18}$$

is the subspace of all functions that satisfy the Dirichlet condition $u(0) = 0$.

Obviously, the real part of the Helmholtz variational form (2.4.17) with $v = u$ satisfies the Gårding inequality (with $C = k^2$). To show the uniqueness of the solution, suppose that there are two solutions u_1, u_2. Then their difference $w = u_1 - u_2$ satisfies the homogeneous equality $b(w,v) = 0$ for all v. Taking $v = w$, we get

$$\int (|w'|^2 - k^2|w|^2) = i\alpha|w(1)|^2.$$

The left-hand side of this equation is real, and the right-hand side is imaginary. Hence for $\alpha \neq 0$, equality can hold only if $w(1) = 0$. Then, by $(2.4.16)_3$ also[7] $w'(1) = 0$, and we are given a homogeneous initial value problem that has the unique solution $w \equiv 0$. Hence $u_1 = u_2$, showing uniqueness. Existence then follows from the Fredholm alternative.

Note that the nonvanishing imaginary boundary term was essential for the proof of uniqueness. In the case that $\alpha = 0$, the problem (2.4.17) reduces to the equation for forced vibrations of a slab. The homogeneous equation then has nontrivial solutions (the eigenmodes) for $k = \pi/2 + n\pi$. At these frequencies, the inhomogeneous problem is no longer uniquely solvable.

Similarly, in two and three dimensions, interior problems for the Helmholtz equation generally have eigenvalues at which there is no unique solution. However, the interior problem has, in general, a unique solution if the vibrations are damped. Since damping is proportional to the velocity, the wave number in that case is complex.

Exterior (Dirichlet or Neumann) problems with a radiation condition are uniquely solvable for all real wave numbers k; cf. Sanchez Hubert–Sanchez Palencia [107, Chapter VIII, Sections 3–4].

---

[7] Here we assume that the variational formulation and the original boundary value problem are equivalent, i.e., u has a second derivative in the weak sense. This is true for $f \in L^2(0,1)$. A more general argument for $f \in H^{-1}(0,1)$ is given in [72].

2.4.4 Regularity and Stability

We have seen (cf. (2.4.12) that stable dependence of the solution on the data follows, in general, from the inf–sup condition. Consider again the one-dimensional model problem (2.4.16) and assume for simplicity that $\alpha = k$. We can show (cf. Theorem 4.1.) that $\beta = O(k^{-1})$. Thus the problem to find $u \in V$ such that $b(u, v) = (f, v)$, $v \in V$, has a unique solution u that satisfies the stability estimate (cf. (2.4.12))

$$\|u\|_{H^1} \leq Ck \|f\|_{H^{-1}} . \tag{2.4.19}$$

The dual space in this case is the space of distributions $H^{-1}(0, 1)$. However, the data f may lie in the space of square-integrable functions $L^2(0, 1)$ or be even smoother. Hence the stability estimates are closely related to the regularity of the solution. The question is, if a boundary value problem is weakly solvable, in what space is the solution contained. In the context of Sobolev spaces $H^s(\Omega)$, we seek $\max\{s : u \in H^s(\Omega)\}$. The answer depends not only on the data and the properties of the differential operator, but also on the regularity of the domain Ω. Here, we consider only the case that Ω is a convex domain[8]. Then the following proposition holds.

Proposition 2.24. *Let $\Omega \in \mathbf{R}^n$, $n = 2, 3$, be a convex domain and assume that u is a solution of the variational problem*

$$\begin{cases} \text{Find } u \in H_0^1(\Omega) : \\ \qquad \displaystyle\int_\Omega (\nabla u \nabla \bar{v} - k^2 u \bar{v})\, dV = (f, v)\,, \qquad \forall v \in H_0^1(\Omega)\,. \end{cases}$$

Then $u \in H^2(\Omega) \cap H_0^1(\Omega)$ if $f \in L^2(\Omega)$.

Since the differential operator of the Helmholtz equation has constant coefficients, and the corresponding sesquilinear form is $H_0^1(\Omega)$-coercive, this proposition is a corollary from a well-known regularity theorem; cf. Hackbusch [62, Theorem 9.1.22].

2.5 Variational Methods

2.5.1 Galerkin Method and Ritz Method

Let V_1, V_2 be Hilbert spaces. Consider the general variational problem (2.2.6) with a sesquilinear form $b : V_1 \times V_2 \to \mathbf{C}$ and an antilinear functional $f \in V_2^*$, and assume that there exists a unique solution $u \in V_1$. We seek an

[8] cf. Remark 2.6.

approximate solution u^N of the form

$$u^N = \sum_{j=1}^{N} u^j \phi_j \,,$$

where the ϕ_j are linearly independent elements of the space V_1, and u^j are unknown complex coefficients. The linear span of the basis functions ϕ_j forms the finite-dimensional subspace $V_1^N \subset V_1$. In general, the space V_1 is infinite-dimensional, and V_1^N is a proper subspace. Similarly, we define a test space $V_2^N \subset V_2$ as the linear span of basis functions $\psi_i \in V_2$ for $i = 1, \ldots, N$.

A function $u^N \in V_1^N$ is called the *Galerkin solution* of the variational problem (2.2.6) if it satisfies the variational equality for all test functions $v \in V_2^N$. Equivalently, one requires that the variational equality hold for each of the basis functions of the test space V_2^N,

$$b(u^N, \psi_i) = (f, \psi_i) \,, \qquad i = 1, \ldots, N \,. \tag{2.5.1}$$

The coefficients u^j are thus determined from the linear system of equations

$$\mathbf{A}\,\mathbf{u} = \mathbf{f} \,, \tag{2.5.2}$$

where

$$[\mathbf{A}]_{ij} = b(\phi_j, \psi_i), \qquad [\mathbf{f}]_i = (f, \psi_i), \qquad i, j = 1, \ldots, N \,. \tag{2.5.3}$$

This general case of the Galerkin method, where $V_1^N \neq V_2^N$, is also called the *Petrov–Galerkin* method. If $V_1^N = V_2^N$, one speaks of the *Bubnov–Galerkin* method.

A similar system of equations is obtained from the *Ritz method*. Let V be a Hilbert space and consider the variational equality

$$a(u, v) = (f, v) \,, \quad \forall v \in V \,, \tag{2.5.4}$$

for a V-elliptic self-adjoint form $a : V \times V \to \mathbf{C}$. By the Lax–Milgram theorem, the form a can be uniquely associated with a self-adjoint positive definite operator $A : V \to V$ such that

$$a(u, v) = (Au, v)_V$$

for all $v \in V$, where $(\cdot, \cdot)_V$ denotes the scalar product on the Hilbert space V. Let $V^N \subset V$ be a finite-dimensional subspace. With the Ritz method, one seeks the function $u^N \in V^N \subset V$ that minimizes the energy functional

$$J(u) := (Au, u) - 2(f, u)$$

in V^N.

The Ritz method is a special case of the Galerkin method. Indeed, since u^N is minimizing,

$$\begin{aligned}(Au^N, u^N) - 2(f, u^N) &\le (A(u^N + \tau v), u^N + \tau v) - 2(f, u^N + \tau v) \\ &= (Au^N, u^N) - 2(f, u^N) + 2\tau \left[(Au^N, v) - (f, v) \right] + \tau^2 (Av, v)\end{aligned}$$

must hold for arbitrary $v \in V^N$ and $\tau \in \mathbf{R}$, and hence the term in the square brackets must vanish. Replacing again $(Au^N, v) = a(u^N, v)$, we obtain (2.5.4).

2.5.2 Convergence Results

Convergence of the Ritz method is assured if the basis functions ϕ_j are *complete* in V with respect to the energy norm $|\!|\!| u |\!|\!| := |a(u,u)|^{1/2}$. The system $\{\phi_j\}_1^\infty$ is called complete if for any function $u \in V$ and arbitrarily small $\varepsilon > 0$ one can find a number N and coefficients α^j such that

$$|\!|\!| u - \sum_{j=1}^{N} a^j \phi_j |\!|\!| \le \varepsilon \,.$$

This condition is satisfied if the system $\{\phi_j\}_1^\infty$ represents the union of the basis functions of an infinite sequence

$$V^1 \subset V^2 \subset \cdots \subset V^N \subset \cdots \subset V \tag{2.5.5}$$

of finite-dimensional subspaces with the property

$$\inf_{v \in V^N} |\!|\!| u - v |\!|\!| \to 0 \quad \text{for } N \to \infty \,, \qquad \forall u \in V \,. \tag{2.5.6}$$

For fixed N, the infimum in (2.5.6) is called the error of best approximation (of an element $u \in V$ by elements $v \in V^N$) in the energy norm. If the form a is V-elliptic, the Ritz solution $u^N \in V^N$ is the best approximation (in the energy norm) of the exact solution $u \in V$.

Indeed, $a(u, v) = a(u^N, v) = (f, v)$ holds for all $v \in V^N$. Thus the error $u - u^N$ is orthogonal to V^N: $a(u - u^N, v) = 0,\ v \in V^N$. Hence

$$|\!|\!| u - u^N |\!|\!|^2 = a(u - u^N, u - u^N) = a(u - u^N, u - v) \le |\!|\!| u - u^N |\!|\!| \, |\!|\!| u - v |\!|\!|$$

for all $v \in V^N$ (we applied the Cauchy–Schwarz inequality in the last step). Canceling the error norm on both sides, we have proved the claim. Therefore, the rate of convergence of the Ritz method on the subspaces (2.5.5) is completely determined by the convergence rate in (2.5.6).

We recall that the property of V-ellipticity carries over to the subspaces in (2.5.5). Furthermore, the energy norm is equivalent to the norm of the space V. Thus an error estimate in the norm $\|\cdot\|_V$ follows directly from

the equivalence relation (2.4.5). We have shown the following proposition.

Theorem 2.25. *Let $a : V \times V \to \mathbf{C}$ be a V-elliptic continuous sesquilinear form. For some $f \in V^*$, let $u \in V$ be the exact solution and let $u^N \in V^N \subset V$ be the Ritz solution of the variational equality (2.5.4). Then*

$$\|u - u^N\|_V \leq \frac{M}{\alpha} \inf_{v \in V^N} \|u - v\|_V \,, \tag{2.5.7}$$

where α, M are the ellipticity and continuity constants of the form a, respectively.

This theorem is known under the name Céa's lemma. As a direct corollary, we see that condition (2.5.6) is sufficient for convergence of the Ritz solutions on the sequence of subspaces V^N.

In the general case of an indefinite form, problem (2.2.6) is uniquely solvable if the continuous form $b(\cdot, \cdot)$ satisfies the "discrete" inf–sup condition

$$\exists\, \beta_N > 0 : \qquad \beta_N \leq \sup_{0 \neq v \in V_2^N} \frac{|b(u, v)|}{\|u\|_{V_1} \|v\|_{V_2}} \,, \qquad \forall 0 \neq u \in V_1^N \,, \tag{2.5.8}$$

and the discrete transposed condition

$$\sup_{u \in V_1^N} |b(u, v)| > 0, \qquad \forall 0 \neq v \in V_2^N \,. \tag{2.5.9}$$

Note that the discrete inf–sup condition does not follow from the "continuous condition" (2.4.10) on $V_1 \times V_2$. Hence, in general, (2.5.8) has also to be proven separately if the continuous condition (2.4.10) is known to hold.

Let us estimate the error $u - u^N$. Trivially, for any $v \in V_1^N$, $\|u - u^N\|_{V_1} \leq \|u - v\|_{V_1} + \|u^N - v\|_{V_1}$ holds by the triangle inequality. We apply the discrete inf–sup condition to the second term on the right:

$$\begin{aligned} \|u - u^N\|_{V_1} &\leq \|u - v\|_{V_1} + \frac{1}{\beta_N} \sup_{0 \neq w \in V_2^N} \frac{|b(u^N - v, w)|}{\|w\|_{V_2}} \\ &\leq \|u - v\|_{V_1} + \frac{1}{\beta_N} \sup_{0 \neq w \in V_2^N} \frac{|b(u - v, w)|}{\|w\|_{V_2}} \,. \end{aligned}$$

The second line is obtained from the first by adding and subtracting u in $b(u^N - v, w)$ and using the orthogonality relation $b(u - u^N, w) = 0$, $w \in V_2^N$. Applying the continuity condition $|b(u - v, w)| \leq M \|u - v\|_{V_1} \|w\|_{V_2}$ and taking the infimum over $v \in V_1^N$, we have shown the following theorem; cf. Babuška [7, Chapter 6].

Theorem 2.26. *Let V_1, V_2 be Hilbert spaces. Consider a sesquilinear form $b : V_1 \times V_2 \to \mathbf{C}$ that satisfies the assumptions of Theorem 2.15 as well*

as (2.5.8) and (2.5.9). Let $u \in V_1$ and $u^N \in V_1^N \subset V_1$ be the exact and approximate solutions, respectively, of the variational problem (2.2.6). Then the error $u - u^N$ satisfies

$$\|u - u^N\|_{V_1} \leq \left(1 + \frac{M}{\beta_N}\right) \inf_{v \in V_1^N} \|u - v\|_{V_1} . \tag{2.5.10}$$

With the notations introduced in (2.5.5) and (2.5.6) we can also formulate a sufficient condition of convergence for the Bubnov-Galerkin method for H^1-coercive forms. We consider the problem (2.2.6) with $V_1 = V_2 = H^1(\Omega)$. Then the following theorem holds.

Theorem 2.27. *Let Ω be a bounded domain and let $b(\cdot, \cdot)$ be a V-coercive sesquilinear form on $V = H^1(\Omega)$, i.e., b satisfies the Gårding inequality (2.4.15). Consider a sequence of subspaces satisfiying (2.5.5) and (2.5.6).*

There exists a number N_0 such that the variational problem (2.2.6) has unique solutions $u^N \in V^N$ for all $N \geq N_0$ and $\|u^N - u\|_1 \to 0$ as $N \to \infty$.

The proof can be found in Hackbusch [62, Section 8.2.2]. Comparing the statement of the theorem to the convergence statement for Ritz solutions, we may predict a different convergence behavior as follows. Ritz solutions (of positive definite problems) will typically begin to converge from the start (with $N = 1$), whereas Galerkin solutions for indefinite but coercive forms may behave erratically for small N and start to display a convergence pattern only if the number of degrees of freedom in the discrete model has passed the critical number N_0.

2.5.3 Conclusions for Helmholtz Problems

We consider the Helmholtz variational form on a bounded domain Ω,

$$B(u, v) = \int_\Omega (\nabla u \nabla \bar{v} - k^2 u \bar{v}) dV ,$$

neglecting the boundary terms in the present discussion. The form B is, in general, not H^1-elliptic, and hence the Ritz method cannot be applied. However, $B(u, v)$ is H^1-coercive. Therefore Galerkin methods can be employed for the numerical solution.

Finite element methods are Galerkin methods with piecewise polynomials as trial and test functions. Thus the result of Theorem 2.27 can be used to establish the asymptotic convergence of finite element solutions to Helmholtz problems. In applied computations with large wave number k, the crucial point is the relation between the wave number and the critical dimension N_0 of the finite element subspaces. We will return to this topic in Chapter 4.

2.6 Summary

Radiating solutions of the Helmholtz equation can be written in the form of an infinite series (obtained by separation of variables) or as an integral representation (using the free space Green's function). One can say that the Green's function "maps" the analytical information from the exterior (in particular, the far-field behavior) onto the wet surface. The integral representation is more general than the series representation, since it can be written for an arbitrary boundary. It forms the analytical basis for boundary element methods. In connection with FEM, we will construct absorbing boundary conditions or infinite elements that are based on the series representation of the exact solution. The FEM is applied to the variational formulation of the governing equations. We outline this formulation and review the theory on existence and uniqueness of solutions to variational problems, as well as the theory of variational numerical methods. The sesquilinear forms for Helmholtz problems are, in general, indefinite, but they satisfy a Gårding inequality.

2.7 Bibliographical Remarks

The method of separation of variables is described in textbooks on solution methods for PDE. The peculiar part in the application of this method to the Helmholtz equation is the proof that the separation constant λ must be of the form $\lambda = n(n+1)$ with integer n. This is often given just as a fact. A rigorous justification can be found, e.g., in the treatise of Morse and Feshbach [95, Chapters 5,10].

On the topic of integral solution representations, as well as on methods for integral equations, see Colton and Kress [39]. The book on integral equations of Kress [85], in particular, Chapters 1–4, is also recommended for further reading on the topics of functional analysis that have been only briefly introduced in this chapter. A broader introduction to functional analysis with an emphasis on applications is given in the recent book by Oden and Demkowicz [102]. For an introduction to the variational formulation of PDE and corresponding solution methods, see, e.g., Michlin [94] or Hackbusch [62]. The first volume of the treatise by Dautray and Lions [41] is an encyclopedic reference on the mathematical modeling of applied problems that lead to the wave and the Helmholtz equations. The methods of solution and many fundamental theoretical results are also covered. The essential results of the mathematical analysis of the wave and Helmholtz equations in exterior domains, especially in the nonclassical weak formulation, are reviewed by Leis [87]. A compact outline of the mathematical theory for exterior Helmholtz problems, including a spectral analysis of the Laplace operator in exterior domains and the investigation of scatte-

ring frequencies, is given by Sanchez Hubert and Sanchez Palencia [107, Chapter VIII]. Roughly speaking, the scattering frequencies are resonant frequencies of Helmholtz problems with radiation damping. Resonance can occur only if $\operatorname{Im} k < 0$ [107, p. 348]. We do not elaborate here on this issue since we assumed that $k \in \mathbf{R}$.

This page intentionally left blank

3

Discretization Methods for Exterior Helmholtz Problems

The finite element method (FEM) is a well-established method for the numerical solution of boundary value problems on bounded domains. The application of this method on unbounded exterior domains usually involves a decomposition of the exterior. In scattering problems, the obstacle is enclosed by an artificial boundary. The FEM is used for discretization of the bounded domain between the obstacle and the artificial boundary. On the artificial boundary, one can prescribe absorbing boundary conditions (ABC) that incorporate (exactly or approximately) the far-field behavior into the finite element model. In this chapter, we review some popular ABC, namely, the truncated Dirichlet-to-Neumann (DtN) map after Feng and Keller–Givoli, the recursion in the Atkinson–Wilcox expansion after Bayliss et al., the localization of a pseudodifferential operator by Padé approximation of its dispersion relation after Enquist and Majda, and the recent perfectly matched layer approach after Bérenger. We start with the outline of the domain decomposition approach and the spherical DtN operator, which exactly maps a radiating exterior solution to the radial derivative of its trace on a coupling sphere (Sections 3.1 and 3.2, respectively). The various ABC are reviewed in Section 3.3. In Section 3.4, we describe the finite element discretization of the near field. The discretization of the far field with infinite elements after Burnett and Demkowicz–Gerdes is investigated in Section 3.5.

3.1 Decomposition of Exterior Domains

A first natural domain decomposition for the coupled fluid–structure problem is suggested by the presence of two media. The elastic structure is discretized with finite elements. Let us assume for convenience that the elastic response is given as a Neumann boundary condition for the exterior problem. We will now describe a decomposition of the exterior into a near field and a far field by the introduction of an artificial boundary enclosing the obstacle.

3.1.1 Introduction of an Artificial Boundary

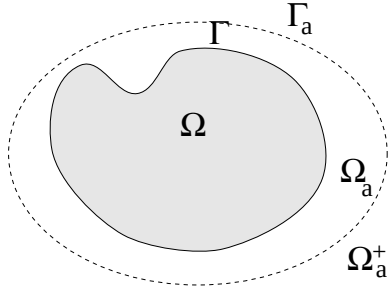


FIGURE 3.1. Scatterer and artificial boundary.

As before, let $\Omega \subset \mathbf{R}^3$ be a sufficiently regular domain[1] and denote by $\Omega^+ := \mathbf{R}^3 \setminus \overline{\Omega}$ the exterior. Consider the Neumann boundary value problem

$$\begin{aligned} -\Delta u - k^2 u &= 0 \quad \text{in } \Omega^+ , \\ \partial_n u &= g \quad \text{on } \Gamma := \partial\Omega , \\ \frac{\partial u}{\partial R} - iku &= o(R^{-1}), \quad R \to \infty , \end{aligned} \tag{3.1.1}$$

where ∂_n denotes the derivative in exterior normal direction on Γ. A natural way of discretizing this exterior problem is to introduce boundary elements on Γ. These boundary elements are generally based on an integral representation of the exact solution in the exterior.[2] If the free space Green's function is used in the kernels, the method of boundary elements (BEM) is theoretically applicable for an arbitrary shape of the obstacle. The Sommerfeld condition is automatically satisfied. It is, in general, not necessary to introduce an artificial boundary $\Gamma_a \subset \Omega^+$ along which finite elements are coupled with boundary elements. Still, a decomposition as depicted in Fig. 3.1 is also an option here (for example, it could be motivated

[1] cf. Remark 2.6.

[2] Some authors also refer to Dirichlet-to-Neumann maps with non-singular kernels as BEM.

by a complicated shape of the wet surface Γ); cf. Hsiao [69]. The BEM for Helmholtz equations is well-researched, and results are available in several monographs. Here we focus on different methods of numerical far-field resolution that are based on the solution by separation of variables. These methods require the introduction of an artificial boundary that is defined on coordinate lines. For instance, in cartesian coordinates the boundary Γ_a is a rectangle given by equations $x = \pm a, y = \pm b$, whereas Γ_a is a spherical surface in spherical coordinates.

3.1.2 Dirichlet-to-Neumann Operators

Assuming that the obstacle Ω is enclosed by a smooth artificial boundary $\Gamma_a \subset \Omega^+$, let Ω_a denote the annular domain between Γ and Γ_a, and let Ω_a^+ denote the reduced exterior domain; cf. Fig. 3.1. Problem (3.1.1) is then equivalently replaced by the coupled problem (cf. Johnson and Nedelec [78])

$$\begin{aligned} -\Delta u_- - k^2 u_- &= 0 && \text{in } \Omega_a\,, \\ \partial_\nu u_- &= g && \text{on } \Gamma\,, \\ u_- &= u_+ && \text{on } \Gamma_a\,, \\ \partial_\nu u_- &= \partial_\nu u_+ && \text{on } \Gamma_a\,, \\ -\Delta u_+ - k^2 u_+ &= 0 && \text{in } \Omega_a^+\,, \\ \frac{\partial u_+}{\partial R} - iku_+ &= o(R^{-1}), && R \to \infty\,. \end{aligned} \tag{3.1.2}$$

Suppose $u_- = u_+$ is given on Γ_a and we can solve analytically the exterior Dirichlet problem for u_+ in Ω_a^+. Having u_+, we can compute $\partial_\nu u_+$ on Γ_a. Thus we have constructed a mapping

$$G : u_+|_{\Gamma_a} \to \partial_\nu u_+|_{\Gamma_a}\,. \tag{3.1.3}$$

Since all the problems are linear, G is a linear operator $G : H^{1/2}(\Gamma_a) \to H^{-1/2}(\Gamma_a)$. This operator is called the *Dirichlet-to-Neumann* (DtN) operator. By $(3.1.2)_{3,4}$, the operator G equivalently maps $u_-|_{\Gamma_a} \to \partial_\nu u_-|_{\Gamma_a}$.

Remark 3.1. One can use both the integral or the series representation of the exact solution for the construction of the DtN operator. In the first case, the operator is also known as the Poincaré–Steklov operator. We will focus on the second case and consider only DtN operators from the series representation.

Using the DtN operator (3.1.3), the coupled problem (3.1.2) is equivalently replaced by the reduced problem

$$-\Delta u - k^2 u = 0 \qquad \text{on } \Omega_a\,,$$

$$\begin{aligned} \partial_\nu u &= g && \text{on } \Gamma\,, \\ \partial_\nu u &= Gu && \text{on } \Gamma_{\mathrm{a}}\,. \end{aligned} \tag{3.1.4}$$

Since we are now solving a problem on the bounded domain Ω_{a}, the question of well-posedness arises naturally.

3.1.3 Well-Posedness

Let us establish a sufficient condition for well-posedness of the reduced boundary value problem (3.1.4). We consider the weak formulation

$$\left\{ \begin{array}{l} \text{Find } u \in H^1(\Omega_{\mathrm{a}}) : \\ \qquad b(u,v) := B(u,v) - \langle Gu, v\rangle_{\Gamma_{\mathrm{a}}} = \langle g, v\rangle_\Gamma\,, \qquad \forall v \in H^1(\Omega_{\mathrm{a}})\,, \end{array} \right. \tag{3.1.5}$$

with

$$B(u,v) = \int_{\Omega_{\mathrm{a}}} \left(\nabla u \nabla \overline{v} - k^2 u \overline{v}\right)\, dV$$

and (we again identify the dual pairings with the L^2 inner products)

$$\langle Gu, v\rangle_{\Gamma_{\mathrm{a}}} = \int_{\Gamma_{\mathrm{a}}} Gu\overline{v}\, dS$$

and

$$\langle g, v\rangle_\Gamma = \int_\Gamma g\overline{v}\, dS\,.$$

From here on, we omit the subscripts in the notation of the dual pairings. We assume that the form $b(\cdot,\cdot)$ satisfies a Gårding inequality, hence existence of a solution to (3.1.5) follows from uniqueness of the adjoint problem. The following proposition was shown by Grote and Keller [58].

Theorem 3.2. *A solution of (3.1.5) is unique if*

$$\operatorname{Im}\langle Gu, u\rangle < 0 \quad (\text{or} > 0) \tag{3.1.6}$$

holds for all $u \in H^{1/2}(\Gamma_{\mathrm{a}}), u \neq 0$.

Indeed, let u be a solution of (3.1.5) with $g \equiv 0$. Then $b(u,u) = B(u,u) - \langle Gu, u\rangle = 0$. Note that $B(u,u)$ is real; hence $\operatorname{Im} b(u,u) = \operatorname{Im}\langle Gu, u\rangle = 0$. By (3.1.6), u vanishes on Γ_{a}. Then also[3] $\partial_\nu u \equiv 0$ on Γ. We continue function u by zero into the exterior to get a function $\hat{u} \equiv 0$ that is "locally H^2"; i.e., $u \in H^2(D)$ for each closed and bounded subdomain $D \subset \Omega_a^+$. From the regularity theory for the Helmholtz operator (cf. Proposition 2.24 and Hackbusch [62, Chapter 9]) it follows that $\hat{u}$ is analytic. Then we can apply

[3] This follows strongly just from $(3.1.4)_3$, but it can also be shown weakly.

the analytic continuation principle to find $0 \equiv \hat{u} \equiv u$ in Ω_a. This proves the theorem.

For the numerical solution, the operator G must be replaced by a truncated DtN operator G_N and, instead of (3.1.4), one solves

$$\begin{aligned} -\Delta u^N - k^2 u^N &= 0 && \text{on } \Omega_a\,, \\ \partial_\nu u^N &= g && \text{on } \Gamma\,, \\ \partial_\nu u^N &= G_N u^N && \text{on } \Gamma_a\,. \end{aligned} \tag{3.1.7}$$

In general, $u^N \neq u$ is some approximation of u. To establish well-posedness, one uses Theorem 3.2, replacing the exact operator G with the approximate operator G_N.

Problem (3.1.7) can be solved by the finite element method (FEM) in the standard way. If u_h^N denotes the finite element solution of (3.1.7), we have, by the triangle inequality

$$\|u - u_h^N\| \leq \|u - u^N\| + \|u^N - u_h^N\|\,. \tag{3.1.8}$$

In general, both errors occur, but (3.1.8) allows us to analyze the convergence of the FEM and the error of the DtN operator separately. The analysis of the finite element error will be the topic of Chapter 4. In the present chapter, we review various methods for the discretization of exterior Helmholtz problems in Ω_a^+

3.2 The Dirichlet-to-Neumann Operator and Numerical Applications

We assume in the following that the coupling surface Γ_a is a sphere of radius $r = a$. Though a spherical domain may not be the best choice in applications (e.g., the scattering from elongated obstacles), we prefer this simple case for the presentation of the principal ideas.

3.2.1 The Exact DtN Operator

The general construction principle for a DtN operator was described in Section 3.1. Suppose that the Dirichlet datum u_- is given on the sphere. We expand u_- into spherical harmonics (cf. Section 2.1.2) as

$$u_-(\theta, \phi) = a^2 \sum_{n=0}^{\infty} \sum_{m=-n}^{n} u_{mn} Y_{mn}(\theta, \phi)$$

with

$$u_{mn} = \int_{S_0} u(a, \theta', \phi') \overline{Y_{mn}}(\theta', \phi')\, dS'\,,$$

where S_0 is the unit sphere. The radiating solution $u_+ = u$ is given on Ω_a^+ by (2.1.14). In our present case, the constants c_{mn} are found from the Dirichlet condition to be

$$c_{mn} = \frac{u_{mn}}{h_n(ka)}\,.$$

Thus

$$u(r,\theta,\phi) = \sum_{n=0}^{\infty} \frac{h_n(kr)}{h_n(ka)} \sum_{m=-n}^{n} u_{mn} Y_{mn}(\theta,\phi)\,. \tag{3.2.1}$$

Differentiating in the radial direction and setting $r = a$ finally leads to

$$\begin{aligned} Gu(\theta,\phi) &:= -\frac{\partial u}{\partial r}(a,\theta,\phi) \\ &= -\sum_{n=0}^{\infty} k\frac{h_n'(ka)}{h_n(ka)} \sum_{m=-n}^{n} u_{mn} Y_{mn}(\theta,\phi)\,. \end{aligned} \tag{3.2.2}$$

Here the negative sign is taken since the outward normal of the exterior region Ω_a^+ points in the negative radial direction. A more compact notation for (3.2.2) is

$$Gu = \sum_{n=0}^{\infty} \alpha_n u_n\,,$$

where

$$\alpha_n = -k\frac{h_n'(ka)}{h_n(ka)} \tag{3.2.3}$$

and

$$u_n(\theta,\phi) = \sum_{m=-n}^{n} u_{mn} Y_{mn}(\theta,\phi)\,.$$

By the orthogonality of the spherical harmonics,

$$\langle Gu, u\rangle = \sum_{n=0}^{\infty} \alpha_n \sum_{m=-n}^{n} |u_{mn}|^2$$

for $u \in L^2(\Gamma_a)$. We will characterize the spectral properties of operator G in the next paragraph.

The DtN operator defines an exact nonreflecting condition on the artificial boundary; i.e., there are no spurious reflections introduced at Γ_a. The near field of the original exterior problem can be then exactly computed from the reduced problem (3.1.4). The operator G is nonlocal since one integrates over the whole surface to compute the coefficients u_{mn}. Thus the coupling matrix is dense, similar to the FEM–BEM coupling.

In cylindrical coordinates, one obtains by a similar procedure the DtN-condition

$$-\frac{\partial u}{\partial r}(a,\theta,\phi) = -\frac{1}{2\pi} \sum_{n=-\infty}^{\infty} k\frac{H_n'(ka)}{H_n(ka)} e^{in\phi} \int_0^{2\pi} u(a,\phi') e^{-in\phi'}\, d\phi'\,. \tag{3.2.4}$$

3.2.2 Spectral Characterization of the DtN-Operator

The following proposition is shown in [45].

Lemma 3.3. *Let $\alpha_n, n = 0, 1, 2, \ldots$, be the complex sequence defined in (3.2.3). Then, for all n,*

$$\operatorname{Im}\alpha_n < 0; \qquad \operatorname{Im}\alpha_n \to -0 \quad \text{as } n \to \infty \tag{3.2.5}$$

and

$$\max(\frac{1}{a}, \frac{n+1}{a} - k) \le \operatorname{Re}\alpha_n \le \frac{n+1}{a} + k\,. \tag{3.2.6}$$

That is, the imaginary parts of all the α_n are negative whereas the real parts are bounded from below and asymptotically behave like $(n+1)a^{-1}$.

Let us first show (3.2.5). The spherical Hankel function of the first kind can be written as a complex sum of Bessel functions and Weber functions, cf. [1, 10.1.1]. Thus

$$\alpha_n := -k\frac{j_n'(ka) + iy_n'(ka)}{j_n(ka) + iy_n(ka)}$$

Hence, with $ka := x$,

$$\operatorname{Im}\alpha_n = -\frac{k}{j_n^2 + y_n^2} \begin{vmatrix} j_n'(x) & y_n'(x) \\ j_n(x) & y_n(x) \end{vmatrix}.$$

By [1, 10.1.6], the Wronskian is x^{-2}, hence

$$\operatorname{Im}\alpha_n = -\frac{k}{|h_n(x)|^2 x^2}\,.$$

This shows that $\operatorname{Im}\alpha_n < 0$ for all n. Now we replace the modulus of the Hankel functions, using [1, 10.1.27],

$$|h_n(x)|^2 = x^{-2} S_n(x)\,, \tag{3.2.7}$$

with

$$S_n(x) = \sum_{k=0}^{n} \frac{(2n-k)![2(n-k)]!}{k![(n-k)!]^2}(2x)^{2k-2n}\,. \tag{3.2.8}$$

We first note that x^2 in the expression for $\operatorname{Im}\alpha$ cancels with that in (3.2.7). Further, rewriting the members of the sum above as

$$s_{nk} = \frac{(2n-k)!}{n!} \cdot \frac{[2(n-k)]!}{(n-k)!} \cdot \frac{n!}{k!(n-k)!}(2x)^{2k-2n}$$

we immediately see that

$$s_{nk} \ge \left(\begin{array}{c} n \\ k \end{array} \right) (2x)^{2k-2n}$$

for all n, k. Hence

$$S_n \geq \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) 1^k \left(\frac{1}{(2x)^2} \right)^{n-k} = \left(1 + \frac{1}{(2x)^2} \right)^n$$

and therefore the imaginary part vanishes asymptotically.

Now turn to the investigation of the real part

$$\operatorname{Re} \alpha_n = -k \frac{j_n'(x) j_n(x) + y_n'(x) y_n(x)}{|h_n(x)|^2} \,. \tag{3.2.9}$$

By direct computation,

$$\operatorname{Re} \alpha_0 = \frac{1}{a} \,.$$

For $n \geq 1$, we replace j_n' and y_n', using the relation [1, 10.1.21],

$$f_n' = f_{n-1} - \frac{n+1}{x} f_n \,,$$

which holds for $f = j$ or $f = y$, to get

$$k^{-1} \operatorname{Re} \alpha_n = \frac{n+1}{x} - r_n(x) \tag{3.2.10}$$

with

$$r_n(x) := \frac{j_{n-1}(x) j_n(x) + y_{n-1}(x) y_n(x)}{|h_n(x)|^2} \,.$$

We claim that $|r_n| < 1$. Indeed, using the expressions for the modulus and phase of the Bessel functions [1, 10.1.26], we can write

$$r_n(x) = \frac{\frac{\pi}{2x} M_{n-\frac{1}{2}}(x) M_{n+\frac{1}{2}}(x) (\cos\theta_{n-\frac{1}{2}} \cos\theta_{n+\frac{1}{2}} + \sin\theta_{n-\frac{1}{2}} \sin\theta_{n+\frac{1}{2}})}{\frac{\pi}{2x} M_{n+\frac{1}{2}}^2(x)} \,,$$

where $M_{n+1/2}(x) = \sqrt{2x/\pi}\, |h_n(x)|$. Hence

$$r_n^2(x) \leq \frac{|h_{n-1}^2(x)|^2}{|h_n^2(x)|^2} = \frac{S_{n-1}(x)}{S_n(x)} \,,$$

where we used again (3.2.7). With the index transformations $l = n - 1 - k$ and $l = n - k$, respectively, we get

$$\begin{aligned} S_{n-1}(x) &= \sum_{l=0}^{n-1} \frac{(n-1+l)!(2l)!}{(n-1-l)!(l!)^2} (2x)^{-2l} \,, \\ S_n(x) &= \sum_{l=0}^{n} \frac{(n+l)!(2l)!}{(n-l)!(l!)^2} (2x)^{-2l} \,. \end{aligned}$$

We write the second sum in the form

$$S_n(x) = \sum_{l=0}^{n-1} a_l \frac{n+l}{n-l}(2x)^{-2l} + s_n(2x)^{-2n},$$

where a_l are the coefficients of $S_{n-1}(x)$. Now it is easy to see that

$$S_n(x) \geq S_{n-1}(x) + s_n(2x)^{-2n}.$$

It follows that $|r_n(x)| \leq 1$ as claimed. Then it follows from (3.2.10) that

$$\frac{n+1}{x} - 1 \leq k^{-1}\mathrm{Re}\,\alpha_n \leq \frac{n+1}{x} + 1. \tag{3.2.11}$$

To complete the proof of (3.2.6), observe that (3.2.9) is equivalently written as

$$\mathrm{Re}\,\alpha_n = -k\,\frac{(|h_n(x)|^2)'}{2|h_n(x)|^2} = \frac{|h_n(x)|'}{|h_n(x)|}.$$

Inserting $|h_n(x)| = x^{-1}\sqrt{S_n(x)}$ we get

$$\mathrm{Re}\,\alpha_n = k\,\frac{1 - x\,\frac{S_n'(x)}{2S_n(x)}}{x}.$$

By direct computation, it is easily checked that $-xS_n'(x)/S_n(x) \geq 0$. Replacing $x = ka$ above and in (3.2.11), the lower bound of (3.2.5) readily follows. The lemma is proved.

3.2.3 Truncation of the DtN Operator

Since the exact DtN operator involves an infinite series, the computation has to be truncated in practice, replacing G with

$$G_N = \sum_{n=0}^{N} \alpha_n u_n. \tag{3.2.12}$$

Now, the problem is well-posed only for the lower-order spherical harmonics. Indeed, $G_N u_n = 0$ if $n > N$. Hence $G_N f = 0$ for any function f that is expanded into higher-order harmonics only. Then the ABC (3.1.4) reduces to the Neumann condition $\partial_\nu f = 0$. The Neumann eigenvalue problem for the Laplace operator on a bounded domain has real eigenvalues (cf. [87, Sections 2.4, 4.1]), and hence there exist, in general, wave numbers k for which the reduced problem (3.1.4) is not uniquely solvable.

This problem is circumvented if one uses the modified truncated DtN operator (Grote and Keller [58])

$$G^* = (G_N - B_N) + B, \tag{3.2.13}$$

where B is any computationally efficient DtN operator with the property (3.1.6), and $(G_N - B_N)$ is the truncation of $G - B$ to the first $N+1$ modes. For example, one can simply take the so-called Sommerfeld operator $B = ik$. Grote and Keller [58] prove that the modified DtN condition renders well-posed problems for all $u \in H^1(\Omega_a)$. They also show in numerical experiments that the error $\|u - u^N\|$ (in maximum norm on Ω_a) is small also for $N \ll ka$. For the truncated DtN condition (3.2.12), Harari and Hughes [66] had proposed the rule

$$N \geq ka\,,$$

based on numerical experiments.

3.2.4 Localizations of the Truncated DtN Operator

We review the approach for the two-dimensional case from Feng [54]. The key is to replace the eigenvalues of the DtN operator (3.2.4) by the asymptotic expansion (for large a)

$$k\frac{H_n'(ka)}{H_n(ka)} \simeq ik \sum_{m=0}^{\infty} \left(\frac{i}{2ka}\right)^m c_m(n^2)\,. \tag{3.2.14}$$

The coefficients c_m are defined recursively. The first four coefficients are of low order in n^2:

$$c_0 = c_1 = 1, \qquad c_2 = 2(1 - \frac{n^2}{2}), \qquad c_3 = -4(1 - \frac{n^2}{4})\,.$$

Inserting into (3.2.4) and interchanging the order of summation, we get

$$\frac{\partial u}{\partial n}(a,\phi) \simeq ik \sum_{n=0}^{\infty} \left(\frac{i}{2ka}\right)^n \sum_{m=-\infty}^{\infty} c_n(m^2) u_m e^{im\phi}\,.$$

Now the sum over m can be interpreted as the action of the differential operator

$$c_n\left(-\frac{\partial^2}{\partial\phi^2}\right)$$

on the function $u = \sum_m u_m e^{im\phi}$. Then, by truncating the DtN operator to N terms, we obtain a local ABC,

$$G_N = ik \sum_{n=0}^{N} \left(\frac{i}{2ka}\right)^n c_n\left(-\frac{\partial^2}{\partial\phi^2}\right)\,. \tag{3.2.15}$$

The first four operators contain only lower derivatives (up to the second) with respect to ϕ. Since the coefficients c_n are written in even powers of

$\partial/\partial\phi$ only, the resulting ABC are symmetric in ϕ. By its construction, this DtN operator reproduces any solution that can be written as a linear combination of the first N Fourier modes on the circle S under the condition that a is sufficiently large (since the asymptotic expansion in r is used).

Remark 3.4. The expansion (3.2.14) is obtained from the expansion (cf. [1, 9.2.7-10])

$$H_n^{(1)}(x) = \sqrt{\frac{2}{\pi x}} e^{i(x-(\frac{1}{2}+\frac{1}{4})\pi)} \sum_{m=0}^{\infty} \left(\frac{i}{2x}\right)^m (n,m)$$

where (n,m) is an even polynomial of n.

Givoli and Keller [60] propose an idea that is similar to Feng's, except that the DtN condition is constructed in such a way that the first N Fourier modes are matched *exactly* by the numerical solution. This is achieved by *first* truncating the exact DtN condition and *then* expanding the finite sum into powers of m^2. To find the expansion, we require that the $N+1$ complex numbers

$$\lambda_n = -k\frac{H_n'(ka)}{H_n(ka)}, \qquad n = 0, 1, \ldots, N,$$

be written as a linear combination of the form

$$z_n = \sum_{m=0}^{N} c_m n^{2m}.$$

Now the unknown complex coefficients c_m can be computed from the linear system

$$\begin{bmatrix} 0^0 & 0^2 & 0^4 & \cdots & 0^N \\ 1^0 & 1^2 & 1^4 & \cdots & 1^N \\ & & \vdots & & \\ N^0 & N^2 & N^4 & \cdots & N^N \end{bmatrix} \begin{Bmatrix} c_0 \\ c_1 \\ \vdots \\ c_N \end{Bmatrix} = \begin{Bmatrix} \lambda_0 \\ \lambda_1 \\ \vdots \\ \lambda_N \end{Bmatrix}.$$

The solutions for $N = 0, 1$ are trivially $c_0 = \lambda_0,\ c_1 = \lambda_1 - \lambda_0$.

3.3 Absorbing Boundary Conditions

We are interested in absorbing boundary conditions (ABC) that are prescribed on Γ to replace the Sommerfeld condition. We distinguish between global (integral) and local (differential) ABC. Assuming that some discretization of the boundary is given and a set of nodal points $X_n := \{x_j\}_1^n$ defined on Γ_a, we call an operator global if at any point of interest $x \in X_n$

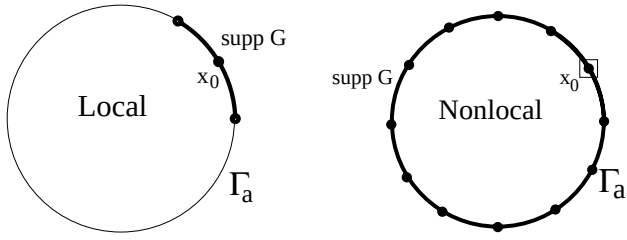


FIGURE 3.2. Local and nonlocal boundary operators.

it acts over all nodal points from X_n; see Fig. 3.2. On the other hand, for some $x_0 \in X_n$ a local operator on $u(x)$ acts only on a subset of M nodal points adjacent to x_0 where m is a fixed number that does not depend on n. We already discussed integral ABC with regular kernels obtained by truncation of the exact DtN map, as well as their localizations. In this section, we review some other popular local ABC. All local operators are differential operators, which are in practice implemented as difference operators.

3.3.1 Recursion in the Atkinson–Wilcox Expansion

We start from the time-dependent case. Assume that $P(\mathbf{x}, t)$, $\mathbf{x} \in \Omega^+$, is a solution of the scaled wave equation

$$\Delta P - \frac{\partial^2 P}{\partial t^2} = 0 \tag{3.3.1}$$

in the exterior domain Ω^+. We define an approximate solution P_N by imposing the condition

$$B_N P_N = 0 \tag{3.3.2}$$

on the coupling surface Γ_a, where the linear operators B_N are defined as follows.

$$\begin{aligned} B_1 &= L + \frac{1}{r}, \quad B_2 = \left(L + \frac{3}{r}\right)\left(L + \frac{1}{r}\right), \\ &\vdots \\ B_N &= \left(L + \frac{2N-1}{r}\right) B_{N-1}, \end{aligned} \tag{3.3.3}$$

with

$$L = \left(\frac{\partial}{\partial t} + \frac{\partial}{\partial r}\right).$$

By the Atkinson–Wilcox expansion (2.1.26),

$$P(r, \theta, \phi, t) = \frac{e^{ik(r-t)}}{r} \sum_{n=0}^{\infty} \frac{p_n(\theta, \phi)}{r^n} \tag{3.3.4}$$

holds in the exterior of any sphere enclosing the obstacle. It is easy to show by inductive argument that

$$B_N P = \sum_{n=1}^{\infty} a_n^N \frac{P_n}{r^{n+N}}, \qquad N = 1, 2, \dots, \tag{3.3.5}$$

where $P_n := e^{ik(r-t)} p_{n-1}$ and

$$a_n^N = (-1)^N (n-1)(n-2)\cdots(n-N). \tag{3.3.6}$$

Thus operator B_N annihilates ("absorbs") the first N terms of the Atkinson–Wilcox expansion of any propagating wave given in the form (3.3.4), and we have the residual condition

$$B_N P = O\left(r^{-(2N+1)}\right).$$

We can obtain time-harmonic variants of the operators B_N by replacing each time differentiation with the operator $-ik$ in the frequency domain. The corresponding time-harmonic solution is denoted by p_N. From (3.3.5), (3.3.6), and (3.3.2) we see that any radiating solution p satisfies the residual relation

$$B_N(p_N - p) = B_N p = \sum_{n=N+1}^{\infty} a_n^N \frac{p_n}{r^{n+N}}.$$

This residual relation shows, in particular, that the ABC will be more exact the larger the radius of the artificial boundary.

In cylindrical coordinates, similar considerations lead to the operator

$$B_N p = \sum_{n=1}^{\infty} a_n^N \frac{p_n}{r^{n+N-\frac{1}{2}}}, \qquad N = 1, 2, \dots \tag{3.3.7}$$

with a_n^N as in (3.3.6). This formula is developed from the expansion

$$p = \sqrt{\frac{2}{\pi k r}} e^{i(kr - \frac{\pi}{2})} \sum_{j=0}^{\infty} \frac{f_j(\phi)}{r^j},$$

which in turn follows from (2.1.27), using [1, 9.2.3].

A DtN operator G_N is defined from B_N by setting

$$G_N := -B_N + \frac{\partial}{\partial n}. \tag{3.3.8}$$

In practice, one usually employs only the first two approximate conditions. We consider here the spherical conditions for the case of axial symmetry (no dependence on ϕ) [20]. From

$$\begin{aligned} B_1 &= \frac{\partial}{\partial r} - ik + \frac{1}{r}, \\ B_2 &= \frac{\partial^2}{\partial r^2} + \left(\frac{4}{r} - 2ik\right) + \left(\frac{2}{r} - 4ik\right)\frac{1}{r} - k^2 \end{aligned}$$

we obtain, using the Bessel equation (2.1.8) to eliminate $\partial^2/\partial r^2$,

$$G_1 = ik - \frac{1}{r}, \tag{3.3.9}$$

$$G_2 = ik - \frac{1}{r} + \frac{\dfrac{\partial}{\partial\theta}\left(\sin\theta\dfrac{\partial}{\partial\theta}\right)}{2r^2\left(\dfrac{1}{r} - ik\right)\sin\theta}. \tag{3.3.10}$$

3.3.2 Localization of a Pseudodifferential Operator

We first consider cartesian coordinates in $\mathbf{R}^2$. Let the solution be given in the form of a wave packet (cf. (2.1.3)),

$$p(x,y) = \int_{-\infty}^{\infty} p(\alpha)e^{i(\alpha x+\beta y)}d\alpha\,,$$

with the dispersion relation

$$\alpha = k\sqrt{1-\sigma^2}, \qquad \sigma = \frac{\beta}{k}\,, \tag{3.3.11}$$

characterizing an outgoing solution. We seek an ABC at the boundary $x =$ constant in two dimensions. If we could transform the dispersion relation exactly into an ABC then all wave packets would pass the boundary without reflection. However, this is not possible if we look for a DtN condition in the form of a local differential operator. Since (3.3.11) is a nonrational expression, it represents a pseudodifferential rather than a differential operator.[4] To obtain local operators, we approximate the nonrational relation by rational functions. Then we deduce the approximate DtN conditions from those differential operators that are the preimages of the approximate Fourier symbols. Enquist and Majda [52] show that the Padé approximations[5] lead to stable and well-posed formulations. In Table 3.1, we list the Padé approximations with residual orders (in σ), the corresponding differential operators for $N = 0, 1, 2$, and the resulting DtN operators for the scaled wave equation (3.3.1). As an example, let us outline the operator for $N = 1$:

$$p_{,x} = i\alpha p \simeq ik\left(1 - \frac{\sigma^2}{2}\right)p = \left(ik + \frac{i(i\beta)^2}{2k}\right)p = \left(ik + \frac{i}{2k}\frac{\partial^2}{\partial y^2}\right)p\,.$$

---

[4]Here it suffices to say that the Fourier symbol of a differential operator is always a polynomial. Conversely, only polynomials can be symbols of differential operators. The theory of pseudodifferential operators allows for more general Fourier symbols and is thus applicable also to integral operators.

[5]We review the basic ideas of Padé approximation in the appendix to this chapter.

TABLE 3.1. Padé approximations and differential operators at $x = 0$.

Order N	Approximation	Residual	Boundary operator
0	1	$O(\sigma^2)$	ik
1	$1 - \frac{1}{2}\sigma^2$	$O(\sigma^4)$	$ik + \frac{i}{2k}\frac{\partial^2}{\partial y^2}$
2	$\frac{1 - \frac{3}{4}\sigma^2}{1 - \frac{1}{4}\sigma^2}$	$O(\sigma^6)$	$ik + \frac{3i}{4k}\frac{\partial^2}{\partial y^2} - \frac{1}{4k^3}\frac{\partial^2}{\partial x \partial y^2}$

We see that the first two Padé expansions are just the Taylor expansions. The approximation for $N = 2$ is the ratio of quadratic polynomials. For $N \geq 3$, the expansions involve higher-order polynomials and consequently lead to higher-order differential operators.

Enquist and Majda also develop boundary conditions in polar coordinates; cf. Section 3.3.3.

Unlike the previous approaches, Enquist and Majda do not attempt modal annihilation. The quality of the ABC is assessed from the amplitude reduction in the reflected modes. A plane wave P_{inc} hitting the boundary $x = 0$ at angle θ produces a spurious reflection P_{ref}; see Fig. 3.3. Considering again the scaled wave equation (3.3.1), we write the incoming wave as

$$P_{\text{inc}} = e^{ik(x\cos\theta + y\sin\theta - t)}.$$

Then

$$P_{\text{ref}} = R_0 e^{ik(-x\cos\theta + y\sin\theta - t)}.$$

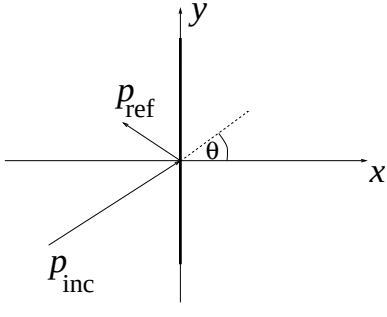


FIGURE 3.3. Reflection of a plane wave at an artificial boundary $x = 0$.

The reflection coefficient R_0 is computed from the zero-order boundary condition $(p_{\text{inc}} + p_{\text{ref}}),_x - ik(p_{\text{inc}} + p_{\text{ref}}) = 0$ at $x = 0$ as

$$R_0 = \frac{\cos\theta - 1}{\cos\theta + 1}\,. \tag{3.3.12}$$

Hence any plane wave incident in the normal direction is completely absorbed, whereas spurious reflections occur for $\theta \neq 0$. One can show [52] that

$$|R_N| = \left|\frac{\cos\theta - 1}{\cos\theta + 1}\right|^{(N+1)}$$

holds for the higher-order approximations. The amplitude of the spurious reflections decreases with the order N.

3.3.3 Comparison of ABC

In Table 3.2 we list the different two-dimensional DtN operators in polar coordinates, supposing that the artificial surface is a sphere of radius a.

TABLE 3.2. DtN conditions in polar coordinates. The abbreviations in the first column stand, respectively, for "Bayliss–Gunzburger–Turkel", "Enquist–Majda", and "Feng".

Authors	DtN conditions G_N		
	$N = 0$	$N = 1$	$N = 2$
BGT	–	$ik - \dfrac{1}{2a}$	$\dfrac{-2k^2 - \dfrac{3ik}{a} + \dfrac{3}{4a^2} + \dfrac{1}{a^2}\dfrac{\partial^2}{\partial\theta^2}}{2\left(ik - \dfrac{1}{a}\right)}$
EM	–	$ik - \dfrac{1}{2a}$	$ik - \dfrac{1}{2a} + \dfrac{1}{2k^2a^2}\left(ik + \dfrac{1}{a}\right)\dfrac{\partial^2}{\partial\theta^2}$
F	ik	$ik - \dfrac{1}{2a}$	$ik - \dfrac{1}{2a} - \dfrac{i}{8ka^2} - \dfrac{i}{2ka^2}\dfrac{\partial^2}{\partial\theta^2}$

Remark 3.5. The Sommerfeld operator ik is formally obtained for the approximation order $N = 0$ only by the approach of Feng.

The corresponding expressions from the Givoli–Keller approach are

$$k\frac{H_o'(ka)}{H_o(ka)}, \qquad k\left[\frac{H_o'(ka)}{H_o(ka)} - \left(\frac{H_1'(ka)}{H_1(ka)} - \frac{H_o'(ka)}{H_o(ka)}\right)\frac{\partial^2}{\partial\theta^2}\right],$$

for $N = 1$ or $N = 2$, respectively (the expression for $N = 3$ involves the fourth derivative in θ). We observe:

1. All conditions are essentially equal for $N = 1$.[6] The condition (3.1.6) for well-posedness is satisfied.

2. For $N = 2$, the expressions differ. However, all conditions, except those of Enquist and Majda, are based on the annihilation of the first three modes in the exterior expansion. Hence one expects similar approximation errors if the radius of the artificial sphere is sufficiently large (recall that Feng uses an expansion that is exactly annihilating only asymptotically). The same applies to the satisfaction of condition (3.1.6). For sufficiently large radius a, all conditions are dominated by the positive definite Sommerfeld operator ik.

3. Feng's as well as Enquist's and Majda's conditions always lead to symmetric (in θ) DtN operators whereas the Bayliss–Gunzburger–Turkel approach does not. In our evaluation, this is visible for $N = 2$.

4. The wave number enters the higher-order operators with inverse power. Thus the improvement in accuracy by higher N is expected to decrease with growing k.

The various ABC have been compared in numerical experiments by Shirron [110]. The accuracy of the conditions is assessed by defining and solving "canonical problems" for rigid scattering from the unit circle; see Fig. 3.4.

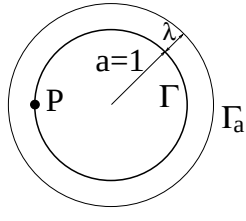


FIGURE 3.4. Cylindrical scatterer and artificial boundary.

Shirron isolates modal information by assuming that the cylinder is subject to a signal consisting of the nth mode only (this is the nth canonical problem). Then the scattered signal also consists of that mode only. The first 30 canonical problems are solved by FEM in the annular domain, imposing the various ABC with orders $N = 0, 1, 2$. The approximation error

[6] A simple calculation shows that

$$k\frac{H_0'(ka)}{H_0(ka)} \simeq ik - \frac{1}{2a}$$

for large ka.

is measured at a point P on the wet surface, and the artificial surface Γ_a is a circle located at the distance λ from Γ. The computations are done for wave numbers $k = 1, 5, 10, 20$. From the numerical data in [110], we draw here two conclusions for $N = 2$.

First, the Bayliss–Gunzburger–Turkel conditions are the most accurate ones, especially for the lower modes. In Shirron's computations, the artificial boundary is very close to the scatterer, i.e., not in an asymptotic range of the radius. Recall that Feng's condition is based on the asymptotic expansion with respect to the radius.

TABLE 3.3. Comparison of relative errors in the first- (BGT1) and second-order (BGT2) ABC after Bayliss–Gunzburger–Turkel for different k. The first column shows the ratio of the errors.

k	BGT1/BGT2	BGT2
1	177	3e-05
5	22	4.3e-03
10	8	2.1e-02
20	4	7.3e-02

Second, the maximal error for each k occurs for $n \approx k$ [110, p. 27]. While the error grows with k, the improvement in $N = 2$ compared to $N = 1$ decreases; see Table 3.3 (data from [110, pp. 28–30]). Consequently, since the maximal error for larger k is contributed by higher modes, the advantage of the Bayliss–Gunzburger–Turkel condition with respect to Feng's and Enquist's and Majda's conditions becomes insignificant in that case.

Resuming our evaluation, we note that the ABC discussed here are qualitatively equivalent. The differences are negligible except for small radius and low wave number. Thus the choice of the "best" ABC can be based on practical considerations, such as symmetry of the resulting DtN operator or the order of the derivatives involved.

3.3.4 The PML Method

A new method for the construction of ABC has been recently proposed by Bérenger [22, 23]. The idea is to introduce an exterior layer at the artificial boundary in such a way that all plane waves are totally absorbed. This means that no reflection occurs for an arbitrary angle of incidence, and the transmitted wave vanishes at infinity, whence the name perfectly matched layer (PML) method. In practice, the computation is truncated at some finite distance within the layer. But the resulting artificial reflections are small, due to the exponential decay.

The idea originated from electromagnetic compuations. Here we outline the method for the acoustic equations in $\mathbf{R}^2$. We consider the artificial boundary $x = 0$ in cartesian coordinates, assuming that the standard acou-

stic equations hold in the domain $\Omega_- = \{x \leq 0\}$ and the infinite layer is $\Omega^+ = \{x \geq 0\}$.

We start from the linearized continuity equation (1.1.2)

$$\frac{\partial \rho}{\partial t} + \rho_0 \mathrm{div}\, \mathbf{V} = 0\,,$$

which we rewrite as a system

$$\frac{\partial \rho_x}{\partial t} = -\rho_0 V_{x,x}\ , \qquad \frac{\partial \rho_y}{\partial t} = -\rho_0 V_{y,y}$$

$$\rho_x + \rho_y = \rho\,.$$

Here ρ_x, ρ_y are just formal variables with no physical meaning. We further recall the linearized Euler equation (1.1.3), which we write componentwise in the form

$$\frac{\partial V_x}{\partial t} = -\frac{1}{\rho_0}\frac{\partial P}{\partial x}\,, \qquad \frac{\partial V_y}{\partial t} = -\frac{1}{\rho_0}\frac{\partial P}{\partial y}\,.$$

Finally, the pressure and density are coupled by the material law

$$P = c^2 \rho\,.$$

In the PML Ω^+, we reformulate those equations that contain a derivative orthogonal to the boundary (here, the x-direction). In these equations we add an absorption term. For our example, the modified equations read

$$\frac{\partial \rho_x}{\partial t} + \sigma(x)\rho_x = -\rho_0 V_{x,x}\ , \qquad \frac{\partial V_x}{\partial t} + \sigma(x) V_x = -\frac{1}{\rho_0}\frac{\partial P}{\partial x}\,.$$

The homogeneous solutions of these equations are of the form $A\exp(-\sigma(x)t)$, and hence we require $\sigma(x) \geq 0$ for $x \geq 0$ to assure decay. Further, we require that $\sigma \equiv 0$ for $x \leq 0$ and that $\sigma(x) \in C^1(\mathbf{R})$.

We are ready to deduce the modified time-harmonic equation for p. Replacing the time-derivatives with $-i\omega$, we obtain

$$\begin{aligned} \rho_x &= -\frac{\rho_0}{\sigma - i\omega} v_{x,x} = \frac{1}{\sigma - i\omega}\frac{\partial}{\partial x}\left(\frac{1}{\sigma - i\omega}\frac{\partial p}{\partial x}\right) \\ \rho_y &= -\frac{1}{\omega^2}\frac{\partial^2 p}{\partial y^2}\,. \end{aligned}$$

Writing the material law in the form $c^{-2}p = \rho_x + \rho_y$ and inserting the relations above, we finally arrive at

$$\frac{i\omega}{\sigma - i\omega}\frac{\partial}{\partial x}\left(\frac{i\omega}{\sigma - i\omega}\frac{\partial p}{\partial x}\right) + \frac{\partial^2 p}{\partial y^2} + k^2 p = 0\,. \tag{3.3.13}$$

For $x \leq 0$, we have $\sigma = 0$ and (3.3.13) reduces to the Helmholtz equation.

Let now $x \geq 0$. Looking at the first term in (3.3.13), it is natural to introduce a new coordinate x' such that

$$\frac{\partial x'}{\partial x} = \frac{i\omega}{\sigma - i\omega}\,.$$

With the transformation

$$x' = x + \frac{i}{\omega}\int_0^x \sigma(\xi)\,d\xi\,, \tag{3.3.14}$$

we thus formally recover the Helmholtz equation in the (x', y) coordinate system. We can write normalized solutions of this equation in the form of a wave packet

$$p(x', y) = \int_{-\infty}^{\infty} e^{i(\alpha x' + \beta y)} d\alpha\,, \qquad \alpha^2 + \beta^2 = k^2\,.$$

Then, for an arbitrarily fixed $\alpha = k_x$, we have

$$e^{ik_x x'} = e^{ik_x x} \exp\left(-k_x \int_0^x \frac{\sigma(\xi)}{\omega}\,d\xi\right)\,,$$

and hence the wave is decaying in the positive x-direction (recall that $\sigma > 0$). No spurious reflection occurs at $x = 0$, since there is no jump in the material properties between the acoustic medium and the PML.

Consider now the practical case that the layer is truncated by a Dirichlet condition at $x = \delta > 0$. Then reflection occurs. Writing, respectively, $p_{\text{inc}} = e^{i(k_x x' + k_y y)}\,, \quad p_{\text{ref}} = Re^{i(-k_x x' + k_y y)}$ for the incoming and reflected waves and setting $p_{\text{inc}} + p_{\text{ref}} = 0$ at $x = \delta$, we see that the the amplitude of the reflected wave is

$$R = -\exp\left(2ik_x \int_0^\delta 1 + \frac{i\sigma(\xi)}{\omega}\,d\xi\right)\,. \tag{3.3.15}$$

Remark 3.6. The introduction of the damping factor σ has no physical significance. It is a mathematical transformation which in effect is an analytical continuation of the elementary solutions into the complex plane; cf. Collino and Monk [38].

3.4 The Finite Element Method in the Near Field

We consider the reduced boundary value problem (3.1.4) and review the main steps of its solution with finite element methods. We consider here only the case of a p-uniform regular mesh. Adjacent elements either share a vertex or a common edge (there are no "hanging nodes"), and the polynomial degree is constant throughout the mesh. The finite element method is applied to the weak formulation (3.1.5). We assume that g and G are such that the problem is well posed.

3.4.1 Finite Element Technology

The numerical solution of (3.1.5) by the FEM involves the following steps (we consider a two-dimensional domain for simplicity).

(1) Triangulation of the Domain:

Domain $\Omega_{\rm a}$ is divided into quadrilaterals and/or triangles τ_i (with straight or curvilinear boundaries) — the finite elements. The set of all elements is called the triangulation

$$T = \{\tau_i,\ i = 1, \ldots, N\}\,.$$

We assume the ideal case that

$$\Omega_{\rm a} = \bigcup_{i=1}^{N} \tau_i\,;$$

that is, the finite elements are an exact partition of $\Omega_{\rm a}$. Then

$$\int_{\Omega_{\rm a}} (\nabla u \nabla \bar{v} - k^2 u \bar{v})\, dV \ = \sum_{i=1}^{N} b_i(u, v)\,,$$

where

$$b_i(u, v) = \int_{\tau_i} (\nabla u \nabla \bar{v} - k^2 u \bar{v})\, dV$$

is the restriction of the form $b(u, v)$ to the element τ_i.

Prescribing some measure function that maps each element to its size h_i (for example, one can take the radius of the inscribed circle for straight-lined triangles), one defines the mesh size

$$h = \max_i h_i$$

of the triangulation T.

(2) Mapping from Master Elements:

Each of the elements τ_i is mapped from master elements by a transform

$$(x, y) = \mathbf{Q}(\xi, \eta)\,, \tag{3.4.1}$$

where ξ, η are defined either on a master square (e.g., $-1 \le \xi, \eta \le 1$) or on a master triangle (e.g., $0 \le \xi \le 1$, $0 \le \eta \le \xi$); see Fig. 3.5. For example, a straight-lined quadrilateral is mapped from the square master element by the linear map

$$\begin{aligned} x &= \frac{1}{4}(X_1(1-\xi)(1-\eta) + X_2(1+\xi)(1-\eta) \\ &+ X_3(1+\xi)(1+\eta) + X_4(1-\xi)(1+\eta))\,, \\ y &= \frac{1}{4}(Y_1(1-\xi)(1-\eta) + Y_2(1+\xi)(1-\eta) \\ &+ Y_3(1+\xi)(1+\eta) + Y_4(1-\xi)(1+\eta))\,, \end{aligned}$$

where (X_i, Y_i) are the coordinates of the element corners in the global x, y coordinate system. In general, the map $\mathbf{Q}$ is nonlinear.

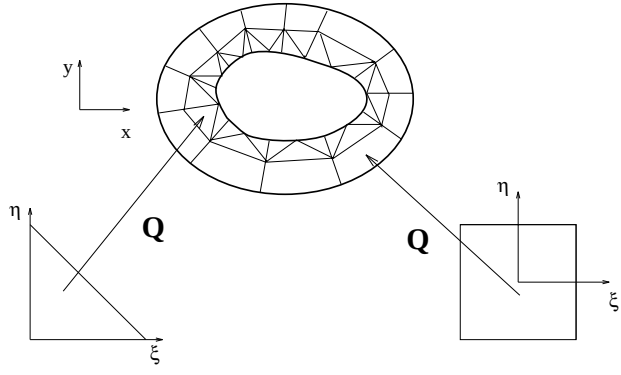


FIGURE 3.5. Triangulation of domain Ω and mapping from master elements.

The local forms b_i are computed on the master elements A as

$$\begin{aligned} b_i &= \int_A \left(\mathbf{J}^{-1}\nabla_{\xi\eta} u(\mathbf{Q}(\xi,\eta))\right)^T \left(\mathbf{J}^{-1}\nabla_{\xi\eta}\bar{v}(\mathbf{Q}(\xi,\eta))\right) \det \mathbf{J}\, d\xi d\eta \\ &\quad -k^2 \int_A u(\mathbf{Q}(\xi,\eta))\bar{v}(\mathbf{Q}(\xi,\eta)) \det \mathbf{J}\, d\xi d\eta\,, \end{aligned}$$

where the gradients are computed in local coordinates (ξ, η), and $\mathbf{J}$ is the Jacobian matrix

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

The design of the mesh has to be such that $\mathbf{J}^{-1}$ exists for each element.

(3) Approximation:

The trial functions u and test functions v are approximated in local coordinates ξ, η by the linear combination

$$u_h(\xi,\eta) = \sum_{i=1}^{n(p)} a_i N_i(\xi,\eta)\,, \tag{3.4.2}$$

where the a_i are unknown complex coefficients and N_i are polynomials of maximal degree p. Each of the coefficients a_i corresponds to a degree of freedom (DOF) of the element τ_i. The functions N_i are called shape functions. In a quadrilateral element, the first four DOF are identified with the nodal values of the unknown function u_h at the corners of the element. The first four shape functions $N_1, \ldots, N_4$ are the bilinear (nodal) shape functions $\frac{1}{4}(1\pm\xi)(1\pm\eta)$. In the hierarchical concept of *hp*-approximation, we further

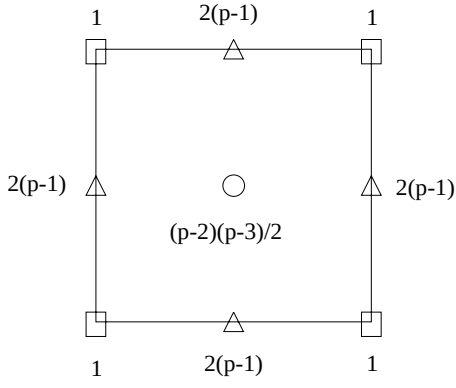


FIGURE 3.6. Degrees of freedom in square element with hierarchical basis: □, nodal DOF; △, edge DOF; ◯ – internal DOF.

have $4(p-1)$ edge modes and $(p-2)(p-3)/2$ internal modes ("bubble functions"); see Fig. 3.6. Here, p is the maximal degree of the polynomial shape functions. Edge modes arise for $p \geq 2$, and internal modes arise only for $p \geq 4$. The mapping $\mathbf{Q}$ transforms the elemental shape functions into element-level basis functions for approximation in global coordinates. The quality of the FE approximation depends on the mesh design (i.e., the size and shape of the elements), the degree of approximation inside the elements, and also on the choice of the transform $\mathbf{Q}$.

The local shape functions form the linear polynomial spaces $S^p(\square)$ or $S^p(\triangle)$, respectively. For the square master elements, the space $S^p(\square)$ consists of all polynomials that can be written as a linear combination of monomials $\xi^i \eta^j,\ 0 \leq i, j \leq p,\ i + j \leq p$, plus the monomial $\xi\eta$ for $p = 1$ or the monomials $\xi^p \eta, \xi\eta^p$ for $p > 1$, respectively. As an example, see below

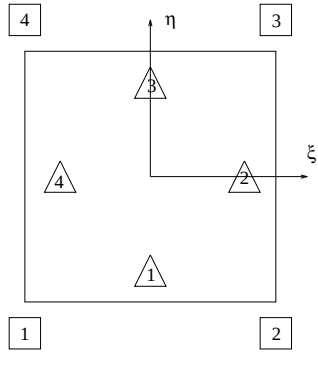


FIGURE 3.7. Square master element.

the hierarchical shape functions (cf. Szabó and Babuška [112, Chapter 6])

for a square master element with node and edge numbering as depicted in Fig. 3.7.

1. *Nodal modes:*

$$\begin{aligned} N_1(\xi,\eta) &= \frac{1}{4}(1-\xi)(1-\eta)\,, \\ N_2(\xi,\eta) &= \frac{1}{4}(1+\xi)(1-\eta)\,, \\ N_3(\xi,\eta) &= \frac{1}{4}(1+\xi)(1+\eta)\,, \\ N_4(\xi,\eta) &= \frac{1}{4}(1-\xi)(1+\eta)\,. \end{aligned}$$

2. *Edge modes:* The edge modes associated with edge 1 are

$$N_i^{(1)} = \frac{1}{2}(1-\eta)\phi_i(\xi), \qquad i = 2,\ldots,p\,,$$

with

$$\phi_i(\xi) = \sqrt{\frac{2i-1}{2}} \int_{-1}^{\xi} P_{i-1}(t)dt\,, \tag{3.4.3}$$

where $P_{i-1}(t)$ are the Legendre polynomials given in (2.1.12). Similarly, the edge modes on edge 2 are

$$N_i^{(2)} = \frac{1}{2}(1+\xi)\phi_i(\eta), \qquad i = 2,\ldots,p\,,$$

etc.

3. *Internal modes:* If $p \geq 4$, then there are $(p-2)(p-3)$ internal modes

$$\begin{aligned} N_1^{(0)}(\xi,\eta) &= \phi_2(\xi)\phi_2(\eta)\,, \\ N_2^{(0)}(\xi,\eta) &= \phi_3(\xi)\phi_2(\eta)\,, \\ N_3^{(0)}(\xi,\eta) &= \phi_2(\xi)\phi_3(\eta), \\ &\vdots \end{aligned}$$

Remark 3.7. It is possible to choose other polynomials than those defined in (3.4.3). For fixed degree of approximation p, the choice of the polynomial shape functions does not influence the quality of approximation. Different shape functions just form different bases of the same space S^p, which determines the degree of local approximation. The elemental shape functions can thus be chosen by practical aspects of the FEM technology such as geometrical approximation of the domain, method of mesh refinement and enrichment, efficient coding and data storage. For instance, the new basis in a p-enrichment step is an extension of the previous basis if hierarchical

shape functions are used; otherwise, a whole new basis has to be introduced if p is changed locally. Considerations of the optimal choice of basis functions for the FEM can be found in [10]. A performance analysis of different hierarchical square elements of order p in the example of the Laplace equation with h-uniform mesh is given in [9]. The choice of basis functions influences the conditioning of the global stiffness matrix. This aspect is of particular importance for very large problems where iterative solvers with preconditioning are used; cf. Babuška et al. [8]. A performance analysis of different hierarchical square elements of order p in the example of the Laplace equation with h-uniform mesh is given in [9].

(4) Computation of Local Stiffness and Mass Matrices:

On each element, the local form $b_i(u, v)$ is computed, where u, v are written on the master element as $\sum_j a_j N_j$, $\sum_i b_i N_i$, respectively. Carrying out the integrations (in general, by Gaussian quadrature), one obtains the local matrix

$$\mathbf{A}_{\mathrm{loc}} = \mathbf{K}_{\mathrm{loc}} - k^2 \mathbf{M}_{\mathrm{loc}} \,,$$

with

$$\begin{aligned} [\mathbf{K}_{\mathrm{loc}}]_{ij} &= \int_A \left(\mathbf{J}^{-1} \nabla N_j(\xi, \eta)\right)^T \left(\mathbf{J}^{-1} \nabla N_i(\xi, \eta)\right) \det \mathbf{J} \, d\xi d\eta \,, \\ [\mathbf{M}_{\mathrm{loc}}]_{ij} &= \int_A N_j(\xi, \eta) N_i(\xi, \eta) \det \mathbf{J} \, d\xi d\eta \,. \end{aligned}$$

(5) Assembly of Global Matrices:

On the global level, we look for an approximate solution $u_h(x, y)$ that is continuous across element junctions. Thus the DOF associated with the nodal and edge modes on neighboring elements have to be identified; see Fig. 3.8. The global stiffness matrix is obtained by summing up the contri-

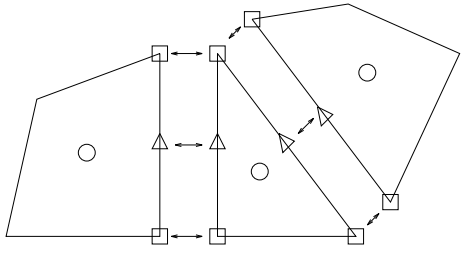


FIGURE 3.8. Identification of modes in assembly.

butions of local stiffness matrices for all elements that contain the vertex or edge under consideration. The internal modes are not connected across

element edges and can be eliminated prior to the assembly by static condensation. After assembly, we obtain the linear system

$$\mathbf{A}\mathbf{u} = \mathbf{g}\,, \tag{3.4.4}$$

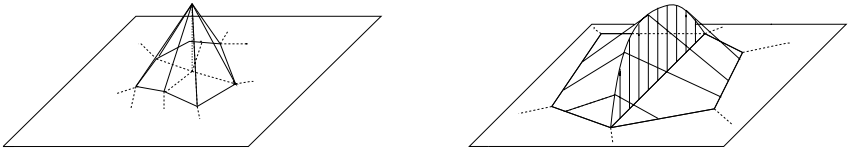
where the system matrix

$$\mathbf{A} = \mathbf{K} - k^2\mathbf{M} \tag{3.4.5}$$

is a linear combination of the stiffness matrix $\mathbf{K}$ and the mass matrix $\mathbf{M}$, and the vector $\mathbf{g}$ has been assembled from the discretization of the right-hand side. Provided that the bubble modes have been condensed, the vector $\mathbf{u}$ contains all DOF associated with the nodal and edge modes of the global mesh.

3.4.2 Identification of the FEM as a Galerkin Method

In the assembly procedure, one effectively constructs continuous basis functions $\Phi_l(x,y)$ over patches $P_l = \bigcup \tau_j$ of adjacent elements. In Fig. 3.9a) we show the nodal basis function on a patch consisting of four quadrilaterals. Similarly, one obtains $p-1$ edge basis functions on patches consisting of two adjacent elements; see Fig. 3.9b). Finally, the internal modes lead to global basis functions that are supported on one element only.



(a) Nodal basis function (b) Edge basis function

FIGURE 3.9. Global basis functions.

Since all basis functions vanish at the patch boundaries, they can be continuously extended by zero onto the whole domain. The trial function u_h is then written as the linear combination

$$u_h(x,y) = \sum_{l=1}^{N_{\text{glob}}} u_l \Phi_l(x,y)\,,$$

with unknown coefficients u_l. The global basis functions $\Phi_l(x,y)$ reduce on every element τ_i to polynomials $\Phi_l^{(i)}(\xi(x,y),\eta(x,y))$ in local coordinates. The linear span of these basis functions forms the finite-dimensional space $S_h^p(\Omega_{\text{a}})$. By definition, $S_h^p(\Omega_{\text{a}})$ is a subspace of the space $H^1(\Omega_{\text{a}})$ in which we seek the exact solution; hence the FEM is conforming.

The finite element solution is found from the variational problem:

$$\left\{ \begin{array}{l} \text{Find } u_h \in S_h^p(\Omega_\mathrm{a}) \ : \\ \qquad b(u_h, v) = \langle g, v \rangle_\Gamma \,, \qquad \forall v \in S_h^p(\Omega_\mathrm{a}) \,. \end{array} \right. \tag{3.4.6}$$

Equivalently, the testing is performed succesively with all the basis functions Φ_l, $l = 1 \ldots N_\mathrm{glob}$.

3.4.3 The h-Version and the hp-Version of the FEM

The FEM is a numerical method for the solution of boundary value problems. We suppose that a unique solution exists and demand that the numerical solution converges to the exact solution. In practice, convergence is either achieved by refining the mesh, letting the mesh size h approach zero asymptotically, or by increasing the degree of polynomial approximation p on a fixed mesh. The first procedure is called h-refinement, whereas the second procedure is called p-enrichment. Correspondingly, one speaks of the h-version and the p-version of the FEM. If convergence is achieved by a combination of h-refinement and p-enrichment, one speaks of the hp-version; see Szabó and Babuška [112].

We will investigate the convergence of the FEM for Helmholtz problems. In this context, we will use the name h-version to refer to the standard procedure of h-refinement with piecewise linear basis functions. If also higher-order polynomials are used in the approximation, we will speak of the hp-version. In our applied computations, we will concurrently use h-refinement and p-enrichment to achieve convergence.

3.5 Infinite Elements and Coupled Finite–Infinite Element Discretization

In Sections 3.2 and 3.3, we reviewed some numerical approaches in which the exterior domain Ω_a^+ outside the artificial boundary Γ_a was truncated. Alternatively, one can partition Ω_a^+ into so-called infinite elements. In this section, we review the infinite elements based on the radial expansion of Wilcox (cf. Section 2.1.4) after Burnett [32].

We first define infinite elements for the exterior of the unit sphere and then consider different variational formulations that lead to Petrov–Galerkin or Bubnov–Galerkin methods.

3.5.1 Infinite Elements from Radial Expansion

Consider the domain decomposition of the exterior as described in Section 3.1. Assume that the annular subdomain Ω_a is partitioned into finite

elements and a subspace $S_h^p(\Omega_a) \subset H^1(\Omega_a)$, based on conforming finite elements, is defined. This naturally induces a finite element partition and a finite element space $S_h^p(\Gamma_a)$ on the artificial boundary Γ_a. Assume further that the finite elements are coupled to infinite elements that form a partition of the exterior domain Ω_a^+, as sketched in Fig. 3.10.

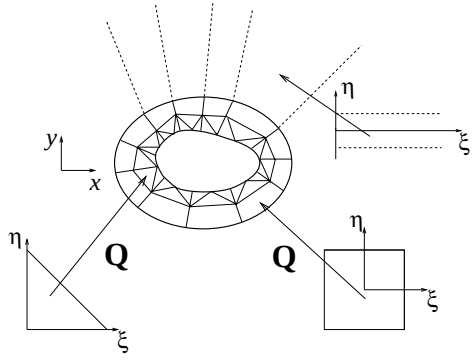


FIGURE 3.10. Partition of the exterior into finite and infinite elements.

In the general case, one assumes that Γ_a is described by a system of coordinates in such a way that the normal at any point of Γ_a is tangential to a coordinate line which goes out from Γ_a to infinity. One can imagine each infinite element bounded by those coordinate lines. Infinite elements based on prolate or oblate spheroids have been described by Burnett [32]. These infinite elements are of practical value if one computes the scattering of waves from elongated or flat obstacles, since they allow to fit the obstacle closely, keeping the finite element region small. If the artificial surface is a sphere, we use spherical coordinates and the coordinate lines are the radial lines.

In the following, we outline the basic ideas of the infinite element discretization for the most simple case that Γ_a is the surface of the unit sphere. The infinite element is a semianalytical construction that reduces to a finite element on the coupling surface. In addition, a finite number of analytical "shape functions" is defined in the radial direction. Thus the IE shape functions are obtained as a tensor product of the FE shape functions and the radial functions. That means, it is assumed that separation of variables can be applied in the exterior of the artificial boundary.

We seek a numerical solution u_h^N that in $\Omega_a^+ = \{r > 1\}$ can be written in the form

$$u_h^N(r,\theta,\phi) = U_N(r)u_h(\theta,\phi)\,, \qquad u_h \in S_h^p(\Gamma_a)\,, \tag{3.5.1}$$

with

$$U_N(r) = \sum_{j=1}^{N} a_j \varphi_j(r)\,,$$

where the radial "shape functions" are the first N members of the radial expansion (2.1.26),

$$\varphi_j(r) = \frac{e^{ikr}}{r^j}, \qquad j = 1, \ldots, N\,, \tag{3.5.2}$$

and the a_j are unknown complex coefficients; see also Fig. 3.11.

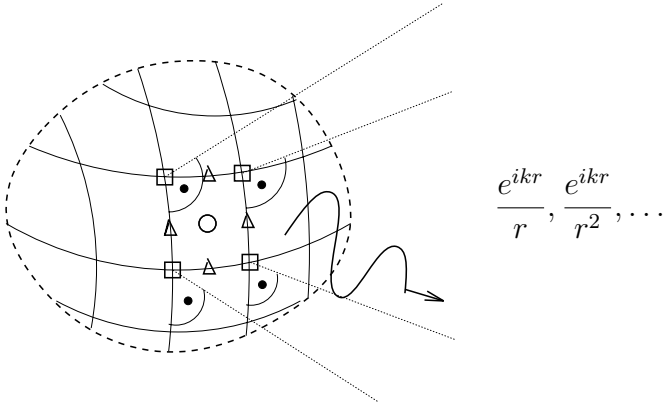


FIGURE 3.11. Infinite element.

The only nodes of the infinite elements are the FE corner nodes on the coupling surface. Similar to the edge and bubble modes in the hierarchical finite element approximation (see, e.g., Szabó and Babuška [112]), the radial degrees of freedom a_j need not be associated with physical locations.

3.5.2 Variational Formulations

A Petrov–Galerkin Formulation:

Let us first consider the variational formulation (2.3.10) with the weighted Sobolev spaces after Leis [87], as introduced in Section 2.3.2. The trial functions u_h^N defined in (3.5.1) lie in the space $H_w^{1+}(\Omega_{\rm a}^+)$. Indeed, the functions always satisfy the Sommerfeld condition by the choice of the radial basis functions, and it can be easily checked that $\|u_h^N\|_{1,w} < \infty$. Hence the trial space

$$V_1 := \left\{ u\,;\, u|_{\Omega_{\rm a}} \in S_h^p(\Omega_{\rm a}) \,\text{and}\, u|_{\Omega_{\rm a}^+} \in S_{h,w}^{pN}(\Omega_{\rm a}^+) \right\}, \tag{3.5.3}$$

where $S_{h,w}^{pN}(\Omega_{\rm a}^+)$ denotes the linear span of the trial functions (3.5.1), is a finite-dimensional subspace of the space $H_w^{1+}(\Omega^+)$. Each function in the trial space reduces on $\Gamma_{\rm a}$ to a finite element function $u_h \in S_h^p(\Gamma_{\rm a})$, whereas in radial direction it lies in the linear span of the shape functions (3.5.2).

The test functions v have to be defined in such a way that the test space V_2 is a subspace of the weighted Sobolev space $H^1_{w^*}(\Omega^+)$. This condition is satisfied if $v = O(r^{-3})$, $r \to \infty$. Defining the radial basis functions

$$\psi_j(r) = \frac{e^{ikr}}{r^j}, \qquad j = 3, \ldots, N+2\,, \tag{3.5.4}$$

we can write the test functions in the form

$$v = v_h^N(r,\theta,\phi) = V_N(r) v_h(\theta,\phi)\,, \tag{3.5.5}$$

with $v_h \in S_h^p(\Gamma_{\rm a})$ and

$$V_N(r) = \sum_{j=1}^{N} b_j \psi_j(r)\,.$$

The test space for the coupled finite–infinite element method is thus

$$V_2 := \left\{ v \in H^1_{w*}(\Omega^+_{\rm a})\,;\, v|_{\Omega_{\rm a}} \in S_h^p(\Omega_{\rm a})\,,\, v|_{\Omega^+_{\rm a}} \in S^{pN}_{h,w*}(\Omega^+_{\rm a}) \right\}\,.$$

where $S^{pN}_{h,w^*}(\Omega^+_{\rm a}) \subset H^{1+}_{w^*}(\Omega^+_{\rm a})$ is the linear span of (3.5.5). Hence the trial and test functions of the coupled finite–infinite element discretization are continuous functions on $\Omega^+ = \Omega_{\rm a} \cup \Omega^+_{\rm a}$ that reduce to piecewise polynomial functions on $\Omega_{\rm a}$ and to semianalytical infinite element functions on $\Omega^+_{\rm a}$.

With these notations, the variational problem is posed as

$$\left\{ \begin{array}{l} \text{Find } u_h^N \in V_1 \,: \\ \qquad B(u_h^N, v) = \langle g, v\rangle_\Gamma, \qquad \forall v \in V_2\,, \end{array} \right. \tag{3.5.6}$$

where

$$B(u,v) = \int_{\Omega^+} (\nabla u \nabla \bar{v} - k^2 u \bar{v})\, dV\,. \tag{3.5.7}$$

Separation of variables under the integral:

Let us write the test and trial functions in the form

$$u = U(r)\Phi(\theta,\phi)\,, \qquad v = V(r)\Psi(\theta,\phi)\,.$$

Then

$$B(u,v) = \int_1^\infty r^2 dr \int_{S_0} [\nabla(\Phi U)\nabla(\overline{\Psi V}) - k^2 U\overline{V}\Phi\overline{\Psi}]\, dS\,,$$

where S_0 denotes the surface of the unit sphere, and $dS = \sin\theta\, d\phi\, d\theta$. Writing out the gradient in spherical coordinates as

$$\nabla = \frac{\partial}{\partial r}\mathbf{e}_r + \frac{1}{r}\nabla_S\,, \quad \text{with} \quad \nabla_S = \frac{\partial}{\partial\theta}\mathbf{e}_\theta + \frac{1}{\sin\theta}\frac{\partial}{\partial\phi}\mathbf{e}_\phi\,,$$

we arrive at the separated form

$$\begin{aligned} B(u,v) &= \int_1^\infty r^2 \left(U_{,r}\,\overline{V}_{,r} - k^2 U\overline{V}\right) dr \int_{S_0} \Phi\overline{\Psi}\, dS \\ &+ \int_1^\infty U\overline{V} dr \int_{S_0} \nabla_S\Phi\,\nabla_S\overline{\Psi}\, dS\,. \end{aligned} \tag{3.5.8}$$

Bubnov–Galerkin Formulations:

Let us now consider variational formulations where the trial and test spaces are identical. This is possible if the spaces are finite-dimensional and the integration in the exterior is understood in the sense of the Cauchy principal value. We disregard the FE region $\Omega_{\rm a}$, considering only the infinite element discretization of $\Omega_{\rm a}^+$ here. The infinite elements are defined as in Section 3.5.1, but now we seek the discrete solution u_h^N from the variational problem

$$\left\{ \begin{array}{l} \text{Find } u_h^N \in V_1 \,: \\ \qquad b_{\rm c}(u_h^N, v) = \langle g, v\rangle_\Gamma, \qquad \forall v \in V_1\,, \end{array} \right. \tag{3.5.9}$$

where V_1 is the space defined in (3.5.3), and the variational form is

$$b_{\rm c}(u,v) := \lim_{R\to\infty} \left(\int_{\Omega_R} (\nabla u \nabla \bar{v} - k^2 u\bar{v})\, dV - ik \int_{S_R} u\bar{v}\, dS \right). \tag{3.5.10}$$

Here, as before, S_R denotes the surface of a large sphere with radius R that is concentric with $\Gamma_{\rm a}$, whereas Ω_R is the annual domain between the $\Gamma_{\rm a}$ and S_R. Unlike the form $B(u,v)$ in (3.5.7), the sesquilinear form $b_{\rm c}(u,v)$ contains a surface integral in the far field. As previously mentioned, this integral always vanishes in the limit $R\to\infty$ for test functions that are $O(r^{-3})$ and, therefore, it is not included in (3.5.7).

The following verification of the integration in (3.5.10) is given after Gerdes [55]. We rewrite the trial and test functions in the form

$$u = \frac{\exp(ikr)}{r^m} f_m\,, \quad v = \frac{\exp(ikr)}{r^n} f_n\,, \tag{3.5.11}$$

where the coefficients f_m and f_n are functions of the angular coordinates, and the summation convention applies over the range $1,\ldots,N$ of the indices. Substituting (3.5.11) into the sesquilinear form (3.5.10) and taking into account (3.5.8), we obtain

$$\begin{aligned} b_{\rm c}(u,v) &= \lim_{R\to\infty} \left\{ \int_1^R r^2 \left[\frac{\partial}{\partial r}\left(\frac{\exp(ikr)}{r^m}\right) \frac{\partial}{\partial r} \overline{\left(\frac{\exp(ikr)}{r^n}\right)} \right.\right. \\ &\qquad \left. - \frac{k^2}{r^{m+n}} \right] dr \int_{S_0} f_m \bar{f}_n\, dS \\ &+ \int_1^R \frac{dr}{r^{m+n}} \int_{S_0} \nabla_S f_m \nabla_S \bar{f}_n\, dS - ik \frac{R^2}{R^{m+n}} \left. \int_{S_0} f_m \bar{f}_n\, dS \right\}, \end{aligned}$$

where we have used the trivial relation

$$\left(\frac{\exp(ikr)}{r^m}\right)\overline{\left(\frac{\exp(ikr)}{r^n}\right)} = \frac{1}{r^{m+n}}\,.$$

Carrrying out the differentiation in r and cancelling powers of r leads to

$$\begin{aligned} b_{\rm c}(u,v) &= \lim_{R\to\infty}\left\{\int_1^R\left[\frac{k^2}{r^{m+n-2}}+\frac{ik(m-n)}{r^{m+n-1}}+\frac{nm}{r^{n+m}}\right.\right.\\ &\quad\left.-\frac{k^2}{r^{m+n-2}}\right]dr\int_{S_0} f_m\bar f_n\,dS\\ &\quad+\left.\int_1^R\frac{1}{r^{m+n}}\,dr\int_{S_0}\nabla_S f_m\nabla_S\bar f_n\,dS-ik\int_{S_0}\frac{R^2}{R^{m+n}}f_m\bar f_n\,dS\right\}. \end{aligned}$$

The integration above is well-defined since the terms k^2/r^{m+n-2} (which are singular for $m+n\le 3$) cancel, and the term $ik(m-n)/r^{m+n-1}$ vanishes for $m=n=1$. For the remaining terms, the integrals in radial direction exist and can be easily computed. The surface integral is evaluated to be

$$\lim_{R\to\infty}\frac{1}{R^{m+n-2}}\int_{S_0} f_m\bar f_n\,dS = \begin{cases} 0, & n+m>2\\ \displaystyle\int_{S_0}|f_1|^2\,dS & n=m=1\,. \end{cases}$$

It is also possible to formulate a variational problem, using a bilinear form instead of the sesquilinear form $b_{\rm c}$. We define the variational form $b_{\rm uc}$ ("unconjugated") in the same way as $b_{\rm c}$ ("conjugated"), but without the complex conjugation of the test functions:

$$b_{\rm uc}(u,v) := \lim_{R\to\infty}\left(\int_{\Omega_R}(\nabla u\nabla v-k^2uv)\,dV-ik\int_{S_R}uv\,dS\right). \qquad (3.5.12)$$

The corresponding variational problem is obtained by replacing $b_{\rm c}$ with $b_{\rm uc}$ in (3.5.9).

Let us verify the integration in this case. Writing out the trial and test functions, we now obtain

$$\begin{aligned} b_{\rm uc}(u,v) &= \lim_{R\to\infty}\left\{I_{mn}\int_1^R r^2\left[\frac{\partial}{\partial r}\left(\frac{\exp(ikr)}{r^n}\right)\frac{\partial}{\partial r}\left(\frac{\exp(ikr)}{r^m}\right)\right.\right.\\ &\quad\left.-k^2\frac{\exp(2ikr)}{r^{m+n}}\right]dr\\ &\quad+\left.J_{mn}\int_1^R\frac{\exp(2ikr)}{r^{m+n}}dr-ik\frac{\exp(2ikR)}{R^{m+n-2}}I_{mn}\right\}, \end{aligned}$$

where we have denoted the surface integrals as

$$I_{mn} := \int_{S_0} f_m f_n \, dS \,, \qquad J_{mn} := \int_{S_0} \nabla_S f_m \nabla_S f_n \, dS \,.$$

Carrying out the differentiation in r and cancelling powers of r, we now obtain

$$\begin{aligned} b_{\mathrm{uc}}(u,v) &= \lim_{R\to\infty} \left\{ I_{mn} \int_1^R \frac{\exp(2ikr)}{r^{m+n}} \left[-2(kr)^2 - i(n+m)kr + nm \right] dr \right. \\ &+ \left. J_{mn} \int_1^R \frac{\exp(2ikr)}{r^{m+n}} \, dr - ik \frac{\exp(2ikR)}{R^{m+n-2}} I_{mn} \right\} . \end{aligned}$$

The integrals of the form

$$\int_1^R \frac{\exp(2ikr)}{r^j} dr, \qquad j \geq 1$$

can be computed, using the sine and cosine integrals [1, Chapter 5]. For details, we refer to Burnett [32].

It remains to consider the limit (for $m = n = 1$)

$$L := \lim_{R\to\infty} \left(\int_1^R -2k^2 \exp(2ikr) dr - ik \exp(2ikR) \right) \int_{S_0} f_1^2 \, dS \,.$$

Performing the integration in radial direction, we observe that the R-dependent terms within the parantheses cancel and the limit is simply

$$L = -ik \exp(2ik) \int_{S_0} f_1^2 \, dS \,.$$

3.5.3 Remarks on the Analysis of the Finite–Infinite Element Method

The numerical analysis of the infinite element discretizations is a matter of ongoing research. Computational tests on various model problems have been reported by Shirron [110] and Gerdes [55]. Let us collect some observations here. For brevity, we will refer to the different formulations as PG, BGC, and BGU (Petrov–Galerkin, Bubnov–Galerkin conjugated, and Bubnov–Galerkin unconjugated, respectively).

In practice, one is not necessarily interested in computing the far-field results directly from the discrete model. Rather, one may use the coupled finite–infinite element discretization to obtain an approximate solution of the near field problem in Ω_{a} only. In the second step, one then computes the far-field pattern from the Helmholtz integral equation (1.1.18), using

the numerical solution on a "collection surface" in the near field. If such an approach is taken, the discretization with infinite elements is effectively used for "mapping" numerically the far-field behavior onto the near field. This interpretation reminds us of the numerical DtN operators discussed in Section 3.2. Let us identify the DtN operator G_N that corresponds to the coupled finite–infinite element discretization. Consider the coupled equations (3.1.2). Assume, for the sake of argument, that we know the Dirichlet datum u_+. Then we can solve the exterior Dirichlet problem $(3.1.2)_{3,5,6}$ with an "infinite element method" based on the Petrov–Galerkin formulation (3.5.6). Taking the normal derivative of the solution on the artificial surface Γ_a, we have constructed an approximate DtN operator G_N.

We now give a more precise definition of the DtN operators G and G_N. Let us write the exterior Dirichlet problem $(3.1.2)_{3,5,6}$ equivalently as $Bu_+ = \hat{u}$, where $\hat{u} \in H^{1/2}(\Gamma_a)$ is the trace of u on Γ_a, and $B : H^{1+}_w(\Omega_a) \to H^{1/2}(\Gamma_a)$ is a linear operator. This problem is uniquely solvable (cf., e.g., Leis [87, Section 4.4]); i.e., the operator B has a bounded inverse operator $B^{-1} : H^{1/2}(\Gamma_a) \to H^{1+}_w(\Omega_a)$ such that $u = B^{-1}\hat{u}$. Identifying u with its trace $\gamma u \in H^{1/2}(\Gamma_a)$ and interpreting the normal derivative as a linear operator $\partial_\nu : H^{1/2}(\Gamma_a) \to H^{-1/2}(\Gamma_a)$, we finally see that $G = \partial_\nu \circ \gamma \circ B^{-1}$.

Similarly, we write the Galerkin formulation of $(3.1.2)_{3,5,6}$ as $B_N u^N = \hat{u}$ where $B_N : V_1^N \subset H^{1+}_w(\Omega_a^+) \to H^{1/2}(\Gamma_a)$ is the linear operator associated with the restriction to the finite-dimensional subspace. Assuming that this problem is also well-posed, we follow the existence of a bounded inverse $B_N^{-1} : H^{1/2}(\Gamma_a) \to H^{1+}_w(\Omega_a^+)$. Concluding similarly as above, we find that $G_N = \partial_\nu \circ \gamma \circ B_N^{-1}$.

Let us now deduce a sufficient condition for convergence of $u_N \to u$, where u_N, u are, respectively, the solutions of (3.1.4) or (3.1.7). With the definitions of G, G_N as above, we seek the exact solution u from

$$\left\{ \begin{array}{l} \text{Find } u \in H^1(\Omega_a) : \\ \qquad b(u,v) = B(u,v) + (Gu,v) = (g,v)\,, \quad \forall v \in H^1(\Omega_a)\,, \end{array} \right. \tag{3.5.13}$$

whereas the approximate solution u^N is obtained from

$$\left\{ \begin{array}{l} \text{Find } u^N \in H^1(\Omega_a) : \\ \qquad b_N(u^N,v) = B(u^N,v) + (G_N u^N,v) = (g,v)\,, \quad \forall v \in H^1(\Omega_a)\,. \end{array} \right. \tag{3.5.14}$$

We see that $b_N(u^N, v) = b(u, v)$ holds for all $v \in H^1(\Omega_a)$. The following theorem is shown in [45].

Theorem 3.8. *Suppose that the variational problem (3.1.4) or, equivalently, (3.5.13), has a unique solution $u \in H^1(\Omega_a)$ with the trace $\hat{u} \in H^\alpha(\Gamma_a)$, $\alpha \geq 1/2$. Assume further that the discrete problem(3.1.7) satisfies the inf–sup condition with $\beta_N > 0$, and let u^N be the solution of (3.5.14). Then the error $u - u^N$ is bounded as*

$$\|u - u^N\|_1 \leq C(\alpha, \Gamma_a, u)\, \beta_N^{-1}\, \|G - G_N\|_{\mathcal{L}(H^\alpha, H^{-1/2})}\,.$$

Hence u^N converges to u if

$$\beta_N^{-1} \, \|G - G_N\|_{\mathcal{L}(H^\alpha, H^{-1/2})} \to 0 \,, \quad N \to \infty \,.$$

For the proof, observe that

$$\begin{aligned} |b_N(u - u^N, v)| &= |b_N(u, v) - b(u, v) + b(u, v) - b_N(u^N, v)| \\ &= |b(u, v) - b_N(u, v)| \\ &= |\langle (G - G_N)u, v \rangle| \,, \qquad \forall v \in H^1(\Omega_\mathrm{a}) \,. \end{aligned}$$

Then, by the discrete inf–sup condition,

$$\begin{aligned} \beta_N \|u - u^N\|_1 &\leq \sup_{0 \neq v \in H^1(\Omega_\mathrm{a})} \frac{|b_N(u - u^N, v)|}{\|v\|_1} \\ &= \sup_{0 \neq v \in H^1(\Omega_\mathrm{a})} \frac{|\langle (G - G_N)u, v \rangle|}{\|v\|_1} \\ &\leq C_\gamma \|G - G_N\|_{\mathcal{L}(H^\alpha, H^{-1/2})} \|\hat{u}\|_{H^\alpha(\Gamma_\mathrm{a})} \,, \end{aligned}$$

where C_γ is the trace constant. Hence

$$\|u - u^N\|_1 \leq C_\gamma \beta_N^{-1} \|G - G_N\|_{\mathcal{L}(H^\alpha, H^{-1/2})} \, \|\hat{u}\|_{H^\alpha(\Gamma_\mathrm{a})} \,,$$

yielding the statement with $C = C_\gamma \, \|\hat{u}\|_{H^\alpha(\Gamma_\mathrm{a})}$.

Remark 3.9. In essence, Theorem 3.8 states a specialization of the well-known fact that convergence follows from stability and approximability. Here, stability is shown if $\beta_N^{-1} < \infty$; whereas approximability is measured in the distance $\|G - G_N\|$. Note that we use the space H^α instead of $H^{1/2}$. As usual, the speed of convergence depends on the regularity of the solution. At this point, we require just the minimal regularity $\alpha = 1/2$, which is naturally satisfied, since $u \in H^1$.

The theorem holds for any shape of the artificial boundary. If Γ_a is a sphere, we expand the trial and test functions into spherical harmonics. As has been shown for the exact DtN operator in Section 3.2, we then find complex numbers α_n^N, $n = 1, 2, \ldots$, such that

$$\langle G_N u, u \rangle = \sum_{n=0}^{\infty} \alpha_n^N \sum_{m=-n}^{n} |u_{mn}|^2, \qquad u \in H^{1/2}(\Gamma_\mathrm{a}) \,. \tag{3.5.15}$$

The α_n^N can be determined numerically by solving modal problems. The following lemma is related to the approximability of the exact solution on the coupling surface.

Lemma 3.10. *Let α_n^N, α_n, respectively, be the coefficients in the spherical expansions (3.5.15) and (3.2.2), (3.2.3). Then for integer $N \geq 1$*

$$\alpha_n^N = \alpha_n, \qquad n = 0, 1, \ldots, N \tag{3.5.16}$$

holds.

To show (3.5.16), we recall (cf. Section 2.1.2) that the exact solution u of the exterior Dirichlet problem with boundary data $\hat{u}$ is

$$u(r, \theta, \phi) = \sum_{n=0}^{\infty} \frac{h_n(kr)}{h_n(ka)} u_n(\theta, \phi)\,,$$

with

$$u_n = \sum_{m=-n}^{n} \hat{u}_{mn} Y_{mn}\,.$$

It is well-known (cf. [1, 10.1.16]) that the Hankel functions $h_n(kr)$ can be expressed as a finite sum of n functions $\varphi_j(r)$ from the infinite element approximation (3.5.2). Hence selecting the Dirichlet data in such a way that $\hat{u}_{mn} \equiv 0$ for $n > N$, the exact solution lies in the trial space of the Galerkin method for the computation of the discrete eigenvalues. Therefore, the exact solution, and hence the exact eigenvalues of the DtN operator, are reproduced for all spherical modes with $n \leq N$.

Using the lemma, one can show that the approximation error $\|G - G_N\|$ decays exponentially as N is increased [45]. Recall that convergence depends on approximability and stability, as expressed in the error estimate (2.5.10). Approximability is quantified by the $\inf_{\chi \in V_1^N} \|u - \chi\|_{V_1}$; i.e., it depends only on the choice of the trial space V_1. Now, all three methods use the same finite-dimensional trial space $V_1 \subset H_w^{1+}(\Omega_{\mathrm{a}}^+)$. Thus all formulations have the same approximability in the exterior Sobolev norm. The approximation error decays exponentially (the typical behavior for spectral approximation). Differences in the convergence behavior must then result from different stability properties. A measure for the stability is given by the discrete inf–sup constant β_N, as discussed in Chapter 2 (see Theorem 2.26). The three forms used for PG, BGC, and BGU, respectively, obviously satisfy the same continuity condition. Thus the specific convergence behavior of each formulation is determined by its inf–sup constant only.

Shirron [110] computes the constants β_N^{BGC} and β_N^{BGU} for a model problem. His results suggest that the BGC constants approach zero with algebraic rate,

$$\beta_N^{\mathrm{BGC}} = O(N^{-2})\,,$$

whereas the BGU constants decay exponentially. We therefore expect that the BGC converges everywhere in the exterior, since the exponential rate of approximability is asymptotically stronger than the algebraic rate of the

stability loss. On the other hand, the BGU formulation does not, in general, converge in the exterior.

Things look different if we restrict ourselves to convergence on the artificial boundary only. Shirron [110] and Gerdes [55] measure the error in maximum norm

$$\|e\|_\infty = \max_{x\in\Gamma_a} |u(x) - u_G(x)|\,,$$

where u_G is the Galerkin solution. Shirron considers a sequence of modal equations that is deduced from the cylindrical scattering problem. Gerdes computes the solutions for rigid scattering from the unit sphere. Both authors observe that all three formulations converge in the maximum norm on Γ_a. Moreover, the BGU formulation converges much more quickly than BGC and PG. For illustration, see results from Gerdes [55] in Table 3.4.

TABLE 3.4. Relative error $\|u-u_{hN}^p\|_\infty/\|u\|_\infty$ on Γ_a, in percent, for the PG, BGC, and BGU formulation. Comparison of results with $N = 1, 3, 6$ radial functions and angular approximation $p = 4$ or $p = 5$.

	p=4			p=5		
	N=1	N=3	N=6	N=1	N=3	N=6
PG	100	87.64	27.43	100	84.0	9.2
BGC	100	100	31.96	100	100	14.7
BGU	80.26	33.23	29.08	58.63	14.04	9.07

3.6 Summary

We have described different approaches to the finite-element analysis of exterior problems. The decomposition of the exterior domain into a near field (inside an artificial surface enclosing the obstacle) and a far field is common to all the approaches. The bounded near-field domain is partitioned into finite elements, whereas the far field is either partitioned into infinite elements, or it is truncated by applying absorbing boundary conditions to the near-field problem. We have reviewed a number of those conditions, writing them in the form of approximate Dirichlet-to-Neumann conditions. We then show that the infinite-element discretization can also be interpreted as an approximate Dirichlet-to-Neumann condition. All the far-field approximations treated in this chapter are related to the series representation of the exact solution, as discussed in Chapter 2.

3.7 Bibliographical Remarks

Absorbing boundary conditions, in the context of the solution of the wave equation by FEM, were first proposed by Enquist and Majda [52]. The work on the other ABC and the DtN conditions originates in the early 1980s. The article by Feng [54] is somewhat hard to locate. It also seems to contain some misprints in the final formulae, which we have corrected here. The pioneering work on infinite elements is due to Bettess and Zienkiewicz [24], see also the monograph [25] of Bettess. A detailed overview of numerical methods for exterior problems is given in Givoli's monograph [60].

Our present review of approximation methods for the far field of the exterior solution does not attempt to be complete. For example, we discuss here only approaches that are based on the series representation of the exact solution. This concerns, in particular, the infinite elements, where we elaborate only on the spectral approximation based on radial expansion. The original paper by Burnett [32] contains a detailed discussion of other formulations. Burnett uses a bilinear form in his coupled finite–infinite-element formulation. The mathematical formulation in weighted Sobolev spaces was first introduced by Demkowicz and Gerdes [46]. It is closely related to the "wave-envelope" formulation by Astley et al. [4].

The finite-element method is treated in a large number of textbooks. Our outline mostly follows Szabó and Babuška [112].

Appendix: Padé Approximation

We seek an approximation of a smooth function $f(x)$ by rational polynomials. First the function $f(x)$ is replaced by a power series expansion $L(x)$, e.g., its Taylor series. As an example, we consider $f(x) = (1+x)^{1/2}$. Then we have

$$f(x) = 1 + \frac{x}{2} - \frac{x^2}{8} + \frac{3x^3}{48} + O(x^4) := L(x)\,.$$

Now we look for a rational function

$$r_{mn}(x) = \frac{p_m(x)}{q_n(x)}\,,$$

where p_m, q_n are polynomials of order m, n, respectively, such that the function p_m approximates the product Lq_n with order $O(x^{m+n+1})$. Let $p_m(x) = a_o + a_1 x + \cdots + a_m x^m$, $q_n(x) = b_0 + b_1 x + \cdots + b_n x^n$, and $L(x) = c_j x^j$. Then the function r_{mn} has the required property if the equations

$$\begin{aligned} c_0 b_0 &= a_0\,, \\ c_1 b_0 + c_0 b_1 &= a_1\,, \\ &\vdots \\ c_m b_0 + c_{m-1} b_1 + \cdots + c_0 b_m &= a_m\,, \end{aligned}$$

and

$$c_{m+1} b_0 + c_m b_1 + \cdots c_{m-n+1} b_n = 0\,,$$

$$\begin{aligned} &\vdots \\ c_{m+n}b_0 + c_{m+n-1}b_1 + \cdots + c_m b_n &= 0 \end{aligned}$$

are satisfied. Functions r_{mn} are called the *Padé* approximations of $L(x)$ (or $f(x)$). The table

$$\begin{array}{cccc} r_{oo} & r_{o1} & r_{o2} & \dots \\ r_{1o} & r_{11} & r_{12} & \dots \\ r_{2o} & r_{21} & r_{22} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{array}$$

is called the Padé table of L. As an example, we compute r_{11} for $f(x)$ as given above. The systems of equations are

$$\begin{aligned} c_2b_0 + c_1b_1 &= 0\,, \\ c_0b_0 &= a_0\,, \\ c_1b_0 + c_0b_1 &= a_1\,, \end{aligned}$$

with $c_0 = 1,\ c_1 = 1/2,\ c_2 = -1/8$. Setting $b_0 = 1$, we get $b_1 = 1/4,\ a_0 = 1$, and $a_1 = 3/4$, which gives

$$r_{11} = \frac{1 + \frac{3x}{4}}{1 + \frac{x}{4}}\,.$$

This rational function approximates $f(x)$ with order $O(x^3)$. For further information, see Jones and Thron [80].

This page intentionally left blank

4

Finite Element Error Analysis and Control for Helmholtz Problems

This chapter is devoted to the finite element analysis of Helmholtz problems on bounded domains. Speaking of "Finite Element Analysis," we elaborate on the computational analysis with the FEM as well as on the numerical analysis of the FEM. We will use the abbreviation "FE" for "finite element," as in "FE mesh," "FE solution" etc.

We assume throughout that we are solving a well-posed problem, one, for example, that results from domain decomposition of an exterior domain $\Omega^+ = \Omega_{\rm a} \cup \Omega_{\rm a}^+$ as described in Section 3.1. In particular, we are interested in the FE solution of the variational problem (3.1.5) in the bounded near-field domain $\Omega_{\rm a}$. The FE discretization of $\Omega_{\rm a}$ has been described in Section 3.4. Solving (3.1.5) on the finite-dimensional subspace $V_h \in H^1(\Omega_{\rm a})$, we get the FE solution $u_h \in V_h$. In general, $u_h \neq u$, and we wish to estimate the error function $e = u - u_h$. We distinguish two cases:

- First, *a priori* error estimation. The error function is estimated in a suitable functional norm *without quantitative input from the computed solution*. The estimates are based on the approximation properties of the subspace where the numerical solution is sought and on the stability properties of the differential operator or variational form. The estimates are generally global; i.e., the error function is estimated in an integral norm computed over the whole solution domain.

- Second, *a posteriori* error estimation. The error function is estimated *employing the computed solution of the discrete model* as data for the estimates. In practice, the estimates are usually part of an adaptive mesh refinement methodology. For mesh refinement, one needs local

information on the error. A posteriori error estimation should therefore be given in a norm that is defined on a single element or on a patch of adjacent elements.

Since the FEM is a Galerkin method, the first natural step of error analysis is to specify the convergence results from Section 2.5 for the present case. This will be our starting point. We give the basic definitions and review the well-known FE convergence theory for positive definite forms. We then turn to the convergence of the FEM for indefinite forms. In particular, we will discuss the specializations of Theorems 2.26 (inf–sup condition) and 2.27 (asymptotic convergence for coercive forms) for Helmholtz problems. Both theorems yield error estimates for the FE solution. However, the upper bound from the inf–sup condition is too crude if the wave number k is large, whereas the asymptotic estimate from the Gårding inequality holds on very fine meshes only (again, if k is large).

The main purpose of this chapter is therefore the proof of so-called preasymptotic estimates. We will show convergence theorems that hold on meshes with the restriction $kh < 1$, which is a standard assumption in engineering practice. The new error bounds contain a pollution term that is related to the loss of stability with large wave numbers. We will give a precise description of this effect. The question arises of whether it is possible to reduce the pollution effect. We first show that FEM with higher-order polynomial approximation (the *hp* version of the FEM) work well towards this goal. Then we review the so-called stabilized FEM, which attempt to correct the loss of stability in the Helmholtz operator. Finally, we analyze the problem of *a posteriori* error estimation and draw conclusions for practical computations. Most of the theorems in this chapter are given with their proofs. Only in cases where proofs seemed too lengthy or technical, we refer the reader to the literature for details.

4.1 Convergence of Galerkin FEM

The FEM is a Galerkin method with piecewise polynomial trial and test functions. Hence the general theory of the Galerkin method applies, while the specific ingredients are the approximation properties of the FE subspaces. The size of the Galerkin subspaces V_N is inversely proportional to the mesh size h of the FE mesh and proportional to the degree of polynomial approximation p. FE error estimates are therefore generally given as a function of these two numerical parameters.

4.1.1 Error Function and Residual

Let V be a Hilbert space, and let a sesquilinear form $b : V \times V \to \mathbf{C}$ be given. Consider the abstract variational problem

$$\begin{cases} \text{Find } u \in V : \\ \qquad b(u, v) = f(v)\,, \qquad \forall v \in V\,, \end{cases} \tag{4.1.1}$$

where $f \in V^*$ is an antilinear functional. With the FEM, we solve instead

$$\begin{cases} \text{Find } u_h \in V_h : \\ \qquad b(u_h, v) = f(v)\,, \qquad \forall v \in V_h\,, \end{cases} \tag{4.1.2}$$

where $V_h \subset V$ is a proper subspace (conforming FEM). Assume that the solutions u, u_h exist. We call

$$e = u - u_h \tag{4.1.3}$$

the error function. Since $V_h \subset V$, we have the standard orthogonality condition

$$b(e, v) = 0, \qquad \forall v \in V_h\,. \tag{4.1.4}$$

The FE solution u_h does not, in general, satisfy the original problem (4.1.1), and we define the residual $r \in V^*$ by

$$r(v) := f(v) - b(u_h, v), \qquad v \in V\,. \tag{4.1.5}$$

Replacing above $f(v) = b(u, v)$, we see that the error function satisfies the residual equation

$$b(e, v) = r(v), \qquad \forall v \in V\,. \tag{4.1.6}$$

If the form b satisfies the assumptions of the Lax–Milgram theorem (or the inf–sup condition), we immediately get the error estimate

$$\|e\|_V \le C\|r\|_{V^*},$$

where C is a stability constant. This type of residual estimate is usually applied *a posteriori*, i.e., after u_h has been computed.

4.1.2 Positive Definite Problems

We first consider a sesquilinear form $a(u, v) : V \times V \to \mathbf{C}$ that satisfies the assumptions of the Lax–Milgram theorem (Theorem 2.10). Then solutions u, u_h of (4.1.1) exist, and the error $u - u_h$ satisfies the error estimate (Céa's Lemma, cf. Theorem 2.25)

$$\|u - u_h\|_V \le \frac{M}{\alpha} \inf_{v \in V_h} \|u - v\|_V\,, \tag{4.1.7}$$

where α, M are, respectively, the ellipticity and the continuity constant. This error estimate allows us to identify the two factors influencing the convergence behavior of any FEM:

- The infimum characterizes the *approximability* of the exact solution in the subspace that is spanned by the FE shape functions. If the infimum is reached on an element $u_{\rm ba} \in V_h$, then this element is the *best approximation* of u in the subspace V_h (in the norm of the "full space" V). Geometrically speaking, $\|u - u_{\rm ba}\|_V$ is the minimal distance between u and V_h. Therefore, $u_{\rm ba}$ is also frequently called the V_h-projection of u; see Fig. 4.1. The error of the best approximation,

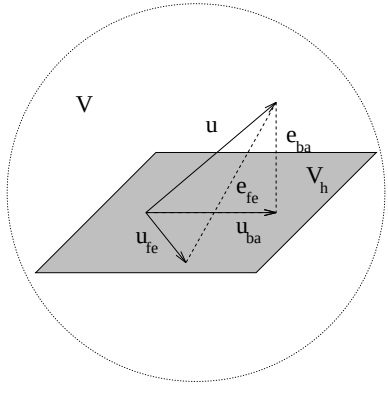


FIGURE 4.1. Best approximation of a function.

$u - u_{\rm ba}$, is therefore V-orthogonal to all functions from the subspace,

$$(u - u_{\rm ba}, v)_V = 0, \qquad \forall v \in V_h\,, \tag{4.1.8}$$

where $(\cdot,\cdot)_V$ is the scalar product of the space V. Thus $u_{\rm ba}$ can be simply computed from the variational problem

$$\left\{ \begin{array}{l} \text{Find } u_{\rm ba} \in V_h : \\ \qquad (u_{\rm ba}, v)_V = (u, v)_V\,, \qquad \forall v \in V\,, \end{array} \right. \tag{4.1.9}$$

provided that *the exact solution is known.*

- The factor $M\alpha^{-1}$ characterizes the *stability* of the problem. In positive definite problems, the stability constant is generally not large, and thus approximability is the major factor influencing convergence.

Resuming, we can verbally express Céa's lemma by the "equality"

$$\boxed{\text{Convergence} = \text{Approximability} + \text{Stability}.}$$

In general, even if the approximation error is small, stability problems may lead to poor convergence.

The approximability by the FE shape functions can be related to the well-known error of interpolation of a given function (here the exact solution)

by piecewise polynomials of degree p. Let $m = p + 1$ and let $\mathcal{I}^p u$ be a pth-order interpolant of u (for example, the continuous, piecewise polynomial, of degree p, Lagrangian interpolant within each element), and let h be the FE mesh size. Then the estimate (cf. Brenner–Scott [31, p. 104])

$$\|u - \mathcal{I}^p u\|_{H^s(\Omega)} \leq C h^{m-s} \|u\|_{H^m(\Omega)}, \qquad 0 \leq s \leq m \tag{4.1.10}$$

holds. It is supposed that the domain Ω and function u are such that the norm on the right-hand side is well-defined. By definition, the error of best approximation is smaller than or equal to the interpolation error; hence the estimate (4.1.10) applies also for the best approximation. The constant C depends, in general, on m, n as well as on the shape of the elements (as characterized by its "chunkiness parameter" [31, p. 97]) and on the norm of the local–global mapping.

4.1.3 Indefinite Problems

Discrete inf–sup Condition:

Unlike the coercivity condition, the inf–sup condition, if proven for the "full space" V, does not automatically hold on the FE subspace V_h. Let a sesquilinear form $b : V_1 \times V_2 \to \mathbf{C}$ on Hilbert spaces V_1, V_2 be given. Assume that the conditions of the Babuška theorem (Theorem 2.15) are satisfied on $V_1 \times V_2$, and let $W_1 \subset V_1, W_2 \subset V_2$ be proper subspaces. If form b satisfies, in addition,

(1) the discrete inf–sup condition

$$\exists \beta_h > 0 : \qquad \beta_h \leq \sup_{0 \neq v \in W_2} \frac{|b(u,v)|}{\|u\| \|v\|}, \qquad \forall 0 \neq u \in W_1, \tag{4.1.11}$$

(2) the transposed condition

$$\sup_{0 \neq u \in W_1} |b(u,v)| > 0, \qquad \forall 0 \neq v \in W_2, \tag{4.1.12}$$

then there exists a unique element $u_h \in W_1$ such that

$$b(u_h, v) = f(v), \qquad \forall v \in W_2 .$$

As a corollary (cf. Theorem 2.26), we find that the error $u - u_h$ can be estimated as

$$\|u - u_h\|_{V_1} \leq \left(1 + \frac{M}{\beta_h}\right) \inf_{v \in W_1} \|u - v\|_{V_1} . \tag{4.1.13}$$

Coercive Problems:

We consider a sesquilinear form $b(\cdot,\cdot)$ on $H^1(\Omega)$, where Ω is a bounded domain and assume that b satisfies the Gårding inequality (2.4.15) as well as the continuity condition $|b(u,v)| \leq M\|u\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)}$. Assume further that the continuous problem (4.1.1) is uniquely solvable, and let u be the exact solution. As before, we denote by V_h the finite-dimensional FE subspaces.

Let $\{u_h \in V_h\}$ denote the FE solutions on a sequence of meshes with decreasing mesh size h (as obtained from successive mesh refinements). On such a sequence, the assumptions of Theorem 2.27 apply. It follows[1] that there exists a "threshold" value h_0 such that the discrete variational problem (4.1.2) has a unique solution u_h for all $h < h_0$, and the error $u - u_h$ satisfies the estimate

$$\|u - u_h\|_V \leq C \inf_{v \in V_h} \|u - v\|_V, \qquad \forall h < h_0 \ , \tag{4.1.14}$$

where C is a constant not depending on h. Estimates of the form (4.1.14) are called quasioptimal.

Remark 4.1. Observe that the error estimates for the indefinite forms are formally similar to the positive definite case. However, the stability constants in (4.1.7) and (4.1.13) are different (cf. the corresponding discussion in Section 2.5; see Remark 2.17 in particular). Estimate (4.1.14) states that the error of the coercive problem behaves asymptotically (practically speaking, on sufficiently fine meshes) as in the positive definite case. However, we will see that the "asymptotic" meshes for Helmholtz forms do not, in general, lie in the range of engineering applications.

4.2 Model Problems for the Helmholtz Equation

We will use three model problems for time-harmonic wave propagation, each featuring essential properties of the general three-dimensional problem of elastic scattering. First, we consider the propagation of a plane wave in a homogeneous exterior domain. The problem is one-dimensional, and hence the unbounded domain can be truncated by imposing the Sommerfeld condition at finite distance. This problem will be used to exemplify the problem of FEM approximation at high wave numbers. In particular, we will show preasymptotic *a priori* error estimates and discuss the pollution effect.

[1] The corresponding theorem on convergence of the Galerkin FEM for coercive problems was proven by Schatz [108]. We reproduce this proof in its specialization to the Helmholtz equation in Section 4.4.2.

The second problem is a two-dimensional Helmholtz equation on a square. Here we will address the dependence of error estimation on the direction of the waves.

The third problem is again one-dimensional, describing the propagation of a plane wave in an inhomogeneous medium. This will allow us to highlight typical numerical effects in the solution of fluid–structure interaction problems.

4.2.1 Model Problem I: Uniaxial Propagation of a Plane Wave

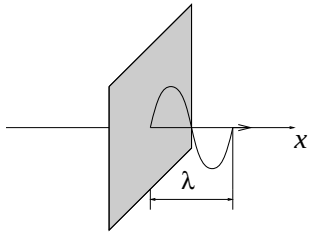


FIGURE 4.2. Uniaxial propagation of plane wave.

The propagation of a time-harmonic plane wave along the x-axis (Fig. 4.2) leads to the boundary value problem that we already introduced in (2.4.16),

$$\begin{aligned} -u'' - k^2 u &= f \qquad \text{on } \Omega = (0,1)\,, \\ u(0) &= 0\,, \\ u'(1) - iku(1) &= 0\,, \end{aligned} \tag{4.2.1}$$

with the corresponding variational formulation

$$\begin{cases} \text{Find } u \in H^1_{(0}(\Omega) : \\ \qquad b(u,v) = (f,v), \qquad \forall v \in H^1_{(0}(\Omega)\,, \end{cases} \tag{4.2.2}$$

where

$$b(u,v) = \int_0^1 u'\bar{v}' - k^2 \int_0^1 u\bar{v} - iku(1)\bar{v}(1)\,, \tag{4.2.3}$$

$$(f,v) = \int_0^1 f(x)\bar{v}(x)dx\,,$$

and $H^1_{(0}(\Omega)$ has been defined in (2.4.18) as the subspace of all H^1-functions that satisfy the Dirichlet condition $u(0) = 0$. For datum $f \in L^2(0,1)$, the solution of the boundary value problem (4.2.1) can be written in the form

$$u(x) = \int_0^1 G(x,s) f(s)\, ds\,, \tag{4.2.4}$$

using the Green's function

$$G(x,s) = \frac{1}{k} \begin{cases} \sin(kx)\, e^{iks} & 0 \le x \le s\,, \\ \sin(ks)\, e^{ikx} & s \le x \le 1\,. \end{cases} \tag{4.2.5}$$

For the numerical solution with finite element methods, let a set of nodes

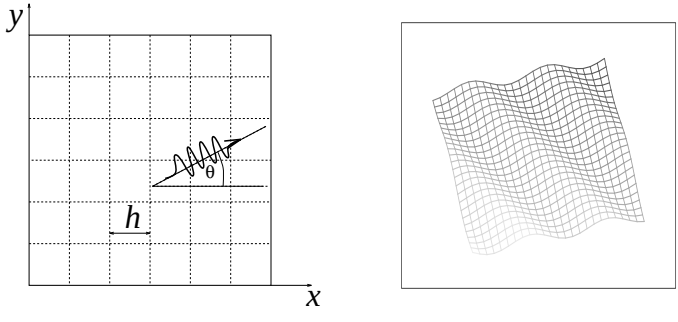
$$X_h := \{x_i;\ 0 = x_0 < x_1 < x_2 < \ldots < x_N = 1\} \tag{4.2.6}$$

be given on $\Omega = (0,1)$. We call X_h the FE mesh and

$$h = \max_{1 \le i \le N} (x_i - x_{i-1}) \tag{4.2.7}$$

the mesh size. The intervals $\tau_i = (x_{i-1}, x_i)$ are called finite elements. The mesh is called uniform if all elements have the same size $h = N^{-1}$.

4.2.2 Model Problem II: Propagation of Plane Waves with Variable Direction



(a) FE mesh

(b) Plane wave in direction θ

FIGURE 4.3. Model Problem II: Domain, FE mesh, and exact solution

We solve the Helmholtz equation in two dimensions on the square $\Omega = (0,1) \times (0,1)$. Figure 4.3a shows the domain and a partition into finite elements with uniform mesh size h. The problem is formulated as

$$-\Delta u - k^2 u = 0 \quad \text{in } \Omega\,,$$

with boundary conditions

$$iku + \frac{\partial u}{\partial n} = g \quad \text{on } \Gamma = \partial\Omega\,,$$

where the function g is chosen such that the exact solution is a plane wave

$$u_{\text{ex}} = e^{i\mathbf{k}\cdot\mathbf{x}}$$

propagating in direction θ; i.e., $\mathbf{k} = k\{\cos\theta, \sin\theta\}$. The corresponding variational formulation is

$$\begin{cases} \text{Find } u \in H^1(\Omega): \\ \qquad b(u,v) = (g,v), \qquad \forall v \in H^1(\Omega)\,, \end{cases} \tag{4.2.8}$$

where

$$b(u,v) = \int_\Omega \left(\nabla u \nabla \bar{v} - k^2 u\bar{v}\right)\, dxdy + ik\int_\Gamma u\bar{v}\, ds\,, \tag{4.2.9}$$

and $(g,v) = \int_\Gamma g\bar{v}\, ds$.

4.2.3 Model Problem III: Uniaxial Fluid–Solid Interaction

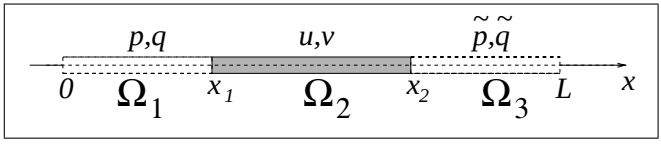


FIGURE 4.4. Fluid-solid interaction model.

The plot of Fig. 4.4 shows a simple case of one-dimensional acoustoelastic fluid–solid interaction. The "fluid" regions Ω_1, Ω_3 have material properties ρ_{f} (density) and c_{f} (speed of sound in fluid). The "solid" region Ω_2 has properties E (Young's modulus) and $\rho_{\text{s}}, c_{\text{s}}$ (density, speed of sound in solid). The domain of computation is $\Omega = \Omega_1 \cup \Omega_2 \cup \Omega_3$, where $\Omega_1 = (0, x_1)$, $\Omega_2 = (x_1, x_2)$, and $\Omega_3 = (x_2, L)$, $0 < x_1 < x_2 < L$. The general physical relations given in Chapter 1 are here specialized to the system of Helmholtz equations

$$p_{,xx} + k^2 p = -g_1 \qquad \text{in } \Omega_1\,, \tag{4.2.10}$$

$$au_{,xx} + nk^2 u = -f \qquad \text{in } \Omega_2\,, \tag{4.2.11}$$

$$\tilde{p}_{,xx} + k^2 \tilde{p} = -g_2 \qquad \text{in } \Omega_3\,, \tag{4.2.12}$$

with the boundary conditions

$$p_{,x}(0) + ikp(0) = 0\,, \tag{4.2.13}$$

$$\tilde{p}_{,x}(L) - ik\tilde{p}(L) = 0\,, \tag{4.2.14}$$

and the transmission conditions

$$p_{,x}(x_1) - k^2 u(x_1) = 0\,, \tag{4.2.15}$$

$$p(x_1) + au_{,x}(x_1) = 0\,, \tag{4.2.16}$$

$$\tilde{p}_{,x}(x_2) - k^2 u(x_2) = 0\,, \tag{4.2.17}$$

$$\tilde{p}(x_2) + au_{,x}(x_2) = 0\,, \tag{4.2.18}$$

with the notations $n = \rho_{\rm f}/\rho_{\rm s}$ and $a = E/B$, where B is the bulk modulus of the fluid; cf. (1.2.9). The real parameter $k = \omega/c_{\rm f}$, where ω is a given frequency, is the wave number in the fluid. The pressures are nondimensional (scaled by the bulk modulus).

The variational formulation is given on the (test and trial) space $V = H^1(\Omega_1) \times H^1(\Omega_2) \times H^1(\Omega_3)$, where $H^1(\Omega_i)$ are the usual Sobolev spaces on the subdomains. Let the trial functions be denoted by $\mathcal{U} = (p, u, \tilde{p})$ and consider the variational form

$$\begin{aligned} b(\mathcal{U}, \mathcal{V}) \;=\; & -ikp(0)\bar{q}(0) + \int_{\Omega_1} p_{,x}\, \bar{q}_{,x}\, dx - k^2 \int_{\Omega_1} p\bar{q} dx \\ & -k^2 u(x_1)\bar{q}(x_1) - k^2 p(x_1)\bar{v}(x_1) \\ & +k^2 \left(\int_{\Omega_2} a u_{,x}\, \bar{v}_{,x}\, dx - nk^2 \int_{\Omega_2} u\bar{v} dx \right) \\ & +k^2 \tilde{p}(x_2)\bar{v}(x_2) + k^2 u(x_2)\bar{\tilde{q}}(x_2) \\ & + \int_{\Omega_3} \tilde{p}_{,x}\, \bar{\tilde{q}}_{,x}\, dx - k^2 \int_{\Omega_3} \tilde{p}\bar{\tilde{q}} dx - ik\tilde{p}(L)\bar{\tilde{q}}(L), \end{aligned} \tag{4.2.19}$$

with trial functions $\mathcal{V} = (q, v, \tilde{q})$. Defining the L^2-type inner product on space V by the weighted sum

$$(\mathcal{U}, \mathcal{V})_0 = (p, q) + k^2 (u, v) + (\tilde{p}, \tilde{q})\,, \tag{4.2.20}$$

we seek the solution $\mathcal{U}$ that satisfies

$$\left\{ \begin{array}{l} \text{Find } \mathcal{U} \in V : \\ \qquad b(\mathcal{U}, \mathcal{V}) = (\mathcal{F}, \mathcal{V})_0, \qquad \forall \mathcal{V} \in V\,, \end{array} \right. \tag{4.2.21}$$

with $\mathcal{F} = (g_1, f, g_2)$.

4.3 Stability Estimates for Helmholtz Problems

Throughout this section, we consider Model Problem I.

4.3.1 The inf–sup Condition

Let us show that the inf–sup constant for the one-dimensional model problem is of order $O(k^{-1})$.

Theorem 4.2. *Let $V = H^1_{(0}(0,1)$, and let $b : V \times V \to \mathbf{C}$ be the sesquilinear form defined in (4.2.3). Then the inf-sup constant*

$$\beta = \inf_{0 \neq u \in V} \sup_{0 \neq v \in V} \frac{|b(u,v)|}{|u|_1 |v|_1}$$

satisfies

$$\frac{C_1}{k} \leq \beta \leq \frac{C_2}{k} \tag{4.3.1}$$

for constants C_1, C_2 not depending on k.

Let us first prove the left inequality of (4.3.1). We will show that for any given $u \in V$ there exists an element $v_u \in V$ such that

$$|b(u, v_u)| \geq \frac{C}{k} |u|_1 \, |v_u|_1 \, . \tag{4.3.2}$$

Let $u \in V$ be given. Define $v_u := u + z$, where z is a solution of the adjoint variational problem

$$b(w, z) = k^2 (w, u), \qquad \forall w \in V \, . \tag{4.3.3}$$

Furthermore, since $u \in H^1(0,1) \subset L^2(0,1)$, the function z is also a solution of the corresponding adjoint boundary value problem[2] with datum $k^2 u$; cf. the discussion of regularity in Section 2.5. Hence

$$z = k^2 \int_0^1 \overline{G(x,s)} u(s) \, ds \, ,$$

with the Green's function $G(x,s)$ from (4.2.5). Then

$$\begin{aligned} |b(u, v_u)| &\geq \text{Re } b(u, v_u) \\ &= \text{Re } (b(u,u) + b(u,z)) \\ &= \text{Re } \big(b(u,u) + b(u,z) + k^2(u,u) - k^2(u,u)\big) \\ &= \text{Re } b(u,u) + k^2 \|u\|^2 = |u|_1^2 . \end{aligned}$$

Now, if we show that

$$|u|_1 \geq \frac{C}{k} \, |v_u|_1, \tag{4.3.4}$$

we have proved (4.3.2) and the lower bound in (4.3.1) follows. To obtain (4.3.4), we integrate by parts the Green's function representation of z,

$$z(x) = k^2 \left(H(x,1) u(1) - \int_0^1 H(x,s) u'(s) ds \right) ,$$

where

$$H(x,s) := \int_0^s \overline{G(x,t)} dt.$$

[2]The adjoint problem to (4.2.1) is obtained by changing the sign in the Sommerfeld condition.

Differentiating this equation and taking absolute values, we get by the triangle inequality

$$\begin{aligned} |z'(x)| &\leq k^2 \left(|H_{,x}(x,1)||u(1)| + \int_0^1 |H_{,x}(x,s)u'(s)|ds \right) \\ &\leq k^2 \left(|H_{,x}(x,1)| + \|H_{,x}\| \right) |u|_1 . \end{aligned}$$

By direct computation $|H_{,x}(x,1)| \leq k^{-1}, \|H_{,x}\| \leq k^{-1}$, and hence

$$|z|_1 \leq 2k|u|_1 .$$

Consequently,

$$|v_u|_1 \leq |u|_1 + |z|_1 \leq (1+2k)|u|_1 ,$$

which proves (4.3.4) for sufficiently large k.

To prove the upper bound of (4.3.1), it is sufficient to find an element $z_0(x) \in V$ that satisfies

$$\frac{|b(z_0,v)|}{|z_0|_1} \leq \frac{C}{k}|v|_1, \qquad \forall v \in V .$$

Consider the function

$$z_0(x) = \varphi(x)\,\frac{\sin kx}{k} ,$$

where $\varphi \in C^\infty(0,1)$ is chosen in such a way that it does not depend on k and

$$z_0(0) = z_0(1) = z_0'(0) = z_0'(1) = 0 \tag{4.3.5}$$

holds. We further require that $|z_0|_1 \geq \alpha$ for some $\alpha > 0$ not depending on k (take, for example, $\varphi(x) = x(x-1)^2$). Then it is easy to see that

$$b(z_0,v) = -\int_0^1 (z_0''(x) + k^2 z_0(x))\bar{v}(x)dx$$

holds for all $v \in V$. Defining

$$u(x) := \int_0^x \left(z_0''(s) + k^2 z_0(s) \right) ds , \tag{4.3.6}$$

one easily derives the bound

$$|b(z_0,v)| = \left| u(1)\bar{v}(1) - \int_0^1 u(x)\bar{v}'(x)\,dx \right| \leq (|u(1)| + \|u\|)|v|_1 .$$

On the other hand, integrating by parts in (4.3.6) and using the definition of z_0, we see that the terms in the right-hand side above can be estimated as follows:

$$|u(1)| \leq \frac{1}{k}\|\varphi''\|_\infty$$

and

$$\|u\| \leq \frac{1}{k} \left(\|\varphi''\|_\infty + 2\|\varphi'\|_\infty \right) .$$

Hence there exists a constant C_1 such that

$$(|u(1)| + \|u\|) \leq \frac{C_1}{k}$$

holds. Collecting results, we can write

$$\frac{|b(z_0, v)|}{|z_0|_1} \leq \frac{1}{\alpha} |b(z_0, v)| \leq \frac{C}{k} |v|_1, \qquad \forall v \in V$$

with $C = \alpha^{-1} C_1$, where α is by definition the lower bound of $|z_0|_1$. This completes the proof of (4.3.1).

Now (2.4.12) immediately gives the following.

Corollary 4.3. *Let $V = H^1_{(0}(0,1)$ and $u \in V$ be the solution of (4.2.2) for data $f \in V^* = H^{-1}(0,1)$. Then*

$$\|u\|_{H^1} \leq Ck \|f\|_{H^{-1}} . \tag{4.3.7}$$

We remark that a similar result on the inf–sup constant holds on the FE subspaces on uniform meshes with mesh size h, i.e., we can show that the discrete inf–sup condition (4.1.11) holds with $\beta_h = O(k^{-1})$; see (4.5.7) below.

4.3.2 Stability Estimates for Data of Higher Regularity

From the general regularity theory it follows that the variational problem (4.2.2) yields solutions $u \in H^{s+1}(0,1)$ for data $f \in H^{s-1}(0,1)$. In the case that $s = 0$, the corresponding stability estimate (4.3.7) has been obtained as a corollary from the Babuška theorem. For our analysis of the h-version and the hp-version of the FEM, we will also need stability estimates for data $f \in H^s(0,1)$, $s \geq 0$. We first treat the case $s = 0$.

Theorem 4.4. *Let $u \in H^1_{(0}(0,1)$ be the variational solution of (2.4.16). If the data $f \in L^2(0,1)$, then the solution u lies in the Sobolev space $H^2(\Omega)$, and the stability estimates*

$$\|u\| \leq \frac{1}{k} \|f\| , \tag{4.3.8}$$

$$|u|_1 \leq \|f\| , \tag{4.3.9}$$

$$|u|_2 \leq (1+k) \|f\| \tag{4.3.10}$$

hold.

Indeed, applying the Cauchy–Schwarz inequality to the integral representation

$$u(x) = \int_0^1 G(x,s) f(s)\, ds\,,$$

with $G(x,s)$ from (4.2.5), we get

$$|u(x)| \le \int_0^1 |G(x,s)|\, |f(s)|\, ds \le \frac{1}{k} \int_0^1 |f(s)|\, ds\,,$$

for all $x \in (0,1)$. The expression on the right can be written as the L^2-inner product $(1, |f|)$, hence

$$|u(x)| \le \|f\|\,,$$

again by the Cauchy–Schwarz inequality. Then (4.3.8) follows if one squares both sides of the inequality and integrates over $(0,1)$.

The second bound (4.3.9) is obtained similarly if one first takes the derivative in x on both sides of the integral representation. This operation is well-defined since $G_{,x}\,(x,s) \in L^2(0,1)$.

Regarding (4.3.10), we first remark that $u \in H^2(0,1)$ by Proposition 2.24. Hence $u'' = f - k^2 u$ holds, and (4.3.10) follows from the triangle inequality and (4.3.8).

Remark 4.5. Similar stability estimates have been shown for Helmholtz problems on convex domains $\Omega \in \mathbf{R}^2$ by Melenk [91].

Remark 4.6. It is easy to show that the seminorm $|\cdot|_1 = \|u'\|$ is a norm on $H^1_{(0}(0,1)$. Indeed, with $u(0) = 0$ we have $u(x) = \int_0^x u'(t)dt$, and then, by the Cauchy–Schwarz inequality,

$$|u(x)| \le \left(\int_0^x 1^2 dt \right)^{1/2} \left(\int_0^1 |u'(t)|^2 dt \right)^{1/2},$$

whence

$$\|u\|^2 = \int_0^1 |u(x)|^2 dx \le \int_0^1 x\, dx\, |u|_1^2 = \frac{1}{2} |u|_1^2\,.$$

We have proven a Poincaré inequality for $u \in H^1_{(0}(0,1)$. Hence on $H^1_{(0}(0,1)$, the seminorm $|\cdot|_1$ is equivalent to the norm $\|u\|_1$; cf. Example 2.13.

The stability estimate of Theorem 4.4 can be generalized for $s \ge 0$. It follows from the regularity theory (cf. Hackbusch [62, Theorem 9.1.16.]) that data $f \in H^{s-1}(0,1)$ is mapped to a solution $u \in H^{s+1}(0,1)$. The matter of interest here is the power to which the wave number k enters the

stability constant.

Theorem 4.7. *Let f be the data and u the solution of (4.2.2). Assume, for $l > 1$, that $f(x) \in H^{l-1}(0,1)$. Then $u \in H^{l+1}(0,1)$, and*

$$|u|_{l+1} \leq C_{\mathrm{s}}(l) k^{l-1} \|f\|_{l-1} \tag{4.3.11}$$

holds for a constant $C_{\mathrm{s}}(l) \leq Dl$, where D does not depend on k and l.

We give the proof for $l = 2$, showing $|u|_3 \leq Ck\|f\|_1$. The solution u is written as

$$u(x) = \int_0^1 G(x,s) f(s) ds\,, \tag{4.3.12}$$

with the Green's function $G(x,s)$ from (4.2.5). By partial integration,

$$u(x) = [H(x,s)f(s)]_{s=0}^{s=1} - \int_0^1 H(x,s) f'(s) ds \tag{4.3.13}$$

where

$$H(x,s) := \int G(x,s) ds = -\frac{1}{k^2} \begin{cases} i\sin(kx)e^{iks} + 1, & 0 \leq x \leq s, \\ \\ \cos(ks)e^{ikx}, & s \leq x \leq 1\,. \end{cases} \tag{4.3.14}$$

For any fixed s (or x, respectively), $H(x,s)$ is an H^2-function of x (or s, respectively). In the boundary points, $H(x,0)$ and $H(x,1)$ are smooth (C^∞) functions. We now estimate

$$|u(x)| \leq |H(x,0)||f(0)| + |H(x,1)||f(1)| + \sup_{x,s} |H(x,s)| \|f'\|\,.$$

From (4.3.14) we have directly

$$\forall x, s: \quad |H(x,s)| \leq \frac{2}{k^2}\,.$$

It is easy to show that

$$\forall s: \quad |f(s)| \leq \sqrt{2}\|f\|_1\,,$$

whence

$$\|u\| \leq \frac{2}{k^2}\left(1 + 2\sqrt{2}\right)\|f\|_1\,. \tag{4.3.15}$$

Next we want to estimate the derivatives of u. Differentiating in (4.3.13), we obtain

$$u'(x) = [H_{,x}(x,s)f(s)]_{s=0}^{s=1} - \int_0^1 H_{,x}(x,s) f'(s) ds\,, \tag{4.3.16}$$

which leads directly to

$$|u|_1 \leq \frac{1}{k}\left(1+2\sqrt{2}\right)\|f\|_1. \tag{4.3.17}$$

Similarly, since $H(x,\cdot) \in H^2(\Omega)$, we obtain, differentiating in (4.3.16),

$$|u|_2 \leq \left(1+2\sqrt{2}\right)\|f\|_1. \tag{4.3.18}$$

Finally, since $u \in H^3(\Omega)$, the differential equation $u''' + k^2 u' = f'$ holds (in the weak sense). Hence

$$|u|_3^2 \leq k^4|u|_1^2 + 2k^2|u|_1|f|_1 + |f|_1^2\,,$$

and with (4.3.17) we obtain

$$|u|_3^2 \leq D^2k^2\|f\|_1^2 + 2Dk\|f\|_1^2 + \|f\|_1^2\,,$$

or, equivalently,

$$|u|_3 \leq C\,k\,\|f\|_1\,, \tag{4.3.19}$$

which proves the statement for $l = 2$. The argument for the higher derivatives is similar; see [73] for details.

4.4 Quasioptimal Convergence of FE Solutions to the Helmholtz Equation

Throughout this section, we consider Model Problem I of Section 4.2.1. We denote by $S_h(0,1) \subset H^1_{(0}(0,1)$ the space of continuous piecewise linear functions (with the nodal values in the points of the mesh X_h), satisfying the Dirichlet condition at $x = 0$.

4.4.1 Approximation Rule and Interpolation Error

The functions $\sin(kx)$ and $\cos(kx)$ are elementary solutions of the Helmholtz equation in one dimension. These solutions are periodic with wavelength $\lambda = 2\pi/k$. It is intuitively clear that a "rule of thumb"

$$n_{\text{res}} = \frac{\lambda}{h} \approx \text{constant} \tag{4.4.1}$$

should be applied in the design of the mesh for given k. The number n_{res} is called the resolution of the mesh – see Fig. 4.5. The choice $n_{\text{res}} = 10$ is usually recommended in practice. This rule leads to reliable results if we *interpolate* an oscillatory function u.

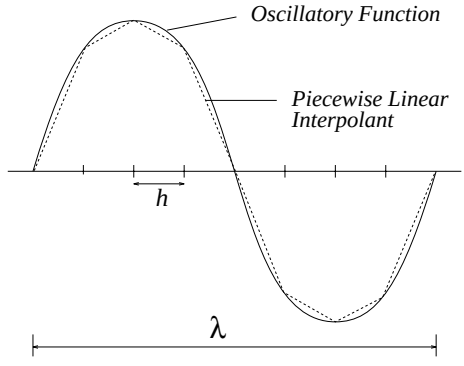


FIGURE 4.5. Resolution of a wave with $n_{\text{res}} = 8$.

Indeed, the interpolation error satisfies the following well-known estimates (see, e.g., Strang and Fix [111, p. 45]):

Lemma 4.8. *Let $u \in H^2(0,1)$, and let $u_I \in S_h(0,1)$ be the piecewise linear interpolant of u on a mesh with mesh size h. Then*

$$\begin{aligned} \|u - u_I\| &\leq \left(\frac{h}{\pi}\right)^2 |u|_2 \,, \\ |u - u_I|_1 &\leq \left(\frac{h}{\pi}\right) |u|_2 \,, \\ \|u - u_I\| &\leq \left(\frac{h}{\pi}\right) |u - u_I|_1 \,. \end{aligned} \tag{4.4.2}$$

Now, assuming that u and u' do not vanish identically, we can divide on both sides of the first two estimates of (4.4.2) by the norms $\|u\|$ and $|u|_1$, respectively, to obtain estimates for the relative error. In our one-dimensional model, the solution u is a linear combination of the elementary solutions $\sin(kx)$, $\cos(kx)$, and we therefore can find constants such that

$$\frac{|u|_2}{\|u\|} \leq C_1 k^2, \quad \frac{|u|_2}{|u|_1} \leq C_2 k \,. \tag{4.4.3}$$

Hence the relative errors of interpolation of an oscillatory solution u satisfy

$$\frac{\|u - u_I\|}{\|u\|} \leq C_1 h^2 k^2 \,, \tag{4.4.4}$$

$$\frac{|u - u_I|_1}{|u|_1} \leq Chk \,. \tag{4.4.5}$$

Since $kh = n_{\text{res}}/(2\pi)$, we conclude that the resolution rule (4.4.1) controls the interpolation error.

It is easy to see that in one dimension (we show it below for $u \in H^1(0,1)$) the nodal interpolant u_I is the best approximation of a given function u in the H^1-seminorm; i.e.

$$|u - u_I|_1 = \inf_{v \in S_h(0,1)} |u - v|_1 \tag{4.4.6}$$

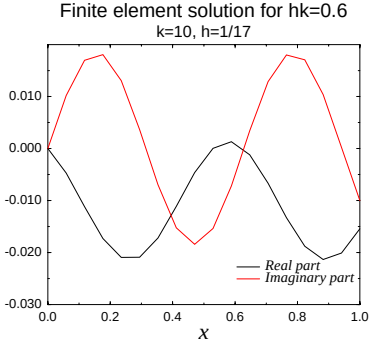
holds. Indeed, the best approximation satisfies

$$0 = ((u - u_{\rm ba})', v') = \int_0^1 (u - u_{\rm ba})' \bar{v}', \qquad \forall v \in S_h(0,1)\,.$$

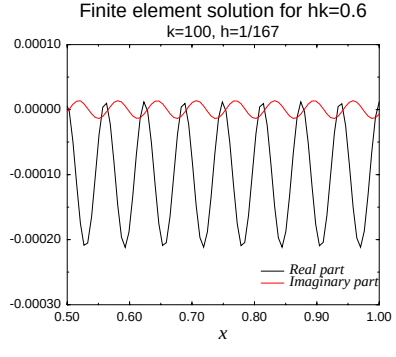
Then, splitting the integral and integrating by parts, we have

$$\begin{aligned} 0 &= \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} (u - u_{\rm ba})' \bar{v}' \\ &= \sum_{i=1}^{N} \left[(u - u_{\rm ba})\, \bar{v}' \right]_{x_{i-1}}^{x_i} - \int_{x_{i-1}}^{x_i} (u - u_{\rm ba}) \bar{v}'', \qquad \forall v \in V_h\,. \end{aligned}$$

The second member on the right vanishes since $v'' \equiv 0$ in the element interiors. Hence the first member must also be zero for all $v \in V_h$, which can be true only if $u = u_{\rm ba}$ in all nodal points. Thus (4.4.6) holds.



(a) $k = 10$



(b) $k = 100$

FIGURE 4.6. Finite element solution for different wave numbers.

We thus have shown that the error of the best approximation is controlled by the "rule of thumb" (4.4.1). On the other hand, it is known from computations that the FE error grows with the wave number also on meshes where the rule of thumb is satisfied. In Fig. 4.6, we show the piecewise linear finite element solution of Model Problem I for two different wave numbers, $k = 10$ and $k = 100$. Both solutions are computed with

$n_{\text{res}} = 10$. The graphs of both FE solutions look fine in the "eyeball norm": both are oscillatory as expected. However, the error (in H^1-seminorm) of the first solution is 21%, while the error of the solution for $k = 100$ is larger than 100%. This example illustrates the problem with the application of the quasioptimal estimate (4.1.14) from the general convergence theory for Galerkin methods. The error of best approximation is the same for both wave numbers (18%), but the FE solution for $k = 100$ cannot be considered a reliable approximation of the exact solution.

Resuming the discussion of this section, we can say that the estimates (4.4.2) determine the size of the optimal error in the solution space of the h-version of the FEM. This optimal error is bounded independently of the wave number k if one applies the rule (4.4.1) for the mesh size h. However, the optimal order of the error is not, in general, achieved in FE computations.

We will now show that the condition $k^2 h \ll 1$ is sufficient for quasioptimality of the FE error.

4.4.2 An Asymptotic Error Estimate

The Helmholtz variational form (4.2.3) is $H^1_{(0}$-coercive. Hence we can specify the quasioptimal estimate (4.1.14) for Model Problem I. We consider here the case of piecewise linear approximation. The corresponding FE subspace is denoted by $S_h(0,1)$.

Theorem 4.9. *Let $f \in L^2(0,1)$, and let $u \in V =: H^2(0,1) \cap H^1_{(0}(0,1), u_h \in V_h =: S_h(0,1)$ be, respectively, the exact and the finite element solutions of (4.2.2) . Assume that h and k are such that the denominators of the constants in the following estimates are positive. Then*

$$|u - u_h|_1 \le C_{\text{s}} \inf_{v \in V_h} |u - v|_1 \tag{4.4.7}$$

holds with the stability constant

$$C_{\text{s}} := \frac{2\left(1 + \left(\frac{hk}{2\pi}\right)^2\right)^{\frac{1}{2}}}{\left(\frac{1}{2} - 6C_1^2 k^2 h^2 (1+k)^2\right)^{\frac{1}{2}}}, \tag{4.4.8}$$

where

$$C_1 := \frac{2}{(1 - 2(1+k)\frac{k^2h^2}{\pi^2})\,\pi}.$$

To prove the statement, let $z \in H^1_{(0}(0,1)$ be such that

$$b(v,z) = (v,e), \qquad \forall v \in H^1_{(0}(0,1)\,,$$

where we denote, as usual, $e := u - u_h$. In particular, $b(e,z) = (e,e)$. Then, for all $w \in V_h$,

$$\begin{aligned}\|e\|^2 &= (e,e) = b(e, z-w) \\ &= \int e' \, (\overline{z-w})' - k^2 \int e \, (\overline{z-w}) - ike(1) \, (\overline{z(1) - w(1)}) \\ &\leq \| \, (z-w)' \| \, \|e'\| + k^2 \|z-w\| \, \|e\| + k|z(1) - w(1)||e(1)| \, ,\end{aligned}$$

where we use the orthogonality $b(e,w) = 0$, and the third line follows from the second by the Cauchy–Schwarz inequality. To estimate the last term above, we use the inequality $|v(1)| \leq \sqrt{2} \|v\|^{\frac{1}{2}} \|v'\|^{\frac{1}{2}}$, which leads to

$$\begin{aligned}k|z(1) - w(1)||e(1)| &\leq 2k \| \, (z-w)' \|^{\frac{1}{2}} \|e'\|^{\frac{1}{2}} \|z-w\|^{\frac{1}{2}} \|e\|^{\frac{1}{2}} \\ &\leq k^2 \|z-w\| \, \|e\| + \| \, (z-w)' \| \, \|e'\| \, , \quad (4.4.9)\end{aligned}$$

where we have also applied the trivial inequality $2ab \leq a^2 + b^2$. Hence

$$\|e\|^2 \leq 2 \left(\| \, (z-w)' \| \, \|e'\| + k^2 \|z-w\| \, \|e\| \right)$$

holds for all $w \in V_h$. Choosing $w = z_I \in V_h$ (the piecewise linear interpolant of z) and recalling that z by definition is the solution to the variational problem with datum e ("Nitsche trick"), we can apply the approximation properties of Lemma 4.2 and the stability conditions from Theorem 4.4, to obtain

$$\begin{aligned}\|e\|^2 &\leq \left(\| \, (z - z_I)' \| \, \|e'\| + k^2 \|z - z_I\| \, \|e\| \right) \\ &\leq 2 \left((1+k) \frac{h}{\pi} \|e'\| \, \|e\| + k^2 \frac{h^2}{\pi^2} (1+k) \, \|e\|^2 \right) .\end{aligned}$$

Dividing both sides of the inequality by the common factor $\|e\|$, we see that

$$\|e\| \leq C_1 \, (1+k) \, h \, \|e'\| \qquad (4.4.10)$$

holds with

$$C_1 := \frac{2}{(1 - 2(1+k)\frac{k^2 h^2}{\pi^2}) \, \pi}$$

under the assumption that k, h are such that the denominator of C_1 is positive.

Let us now derive an estimate for $|e|_1$. By the definition of the error function, the trivial identity $b(e,e) = b(e, u - u_h)$ holds. Using the orthogonality property of the error, we can replace u_h above with an arbitrary function from the subspace: $b(e,e) = b(e, u-v) \; \forall v \in V_h$. Writing out this equation, we have

$$\begin{aligned}&\int_0^1 e' \overline{e}' - k^2 \int_0^1 e\overline{e} - ik|e(1)|^2 = \\ &\qquad \int_0^1 e' (\overline{u-v})' - k^2 \int_0^1 e(\overline{u-v}) - ike(1)(\bar{u}(1) - \bar{v}(1)) \, ,\end{aligned}$$

and therefore

$$\begin{aligned}\|e'\|^2 &\leq k^2\|e\|^2 + k|e(1)|^2 + \|e'\|\,\|(u-v)'\| \\ &\quad + k^2\|e\|\,\|u-v\| + k|e(1)|\,|u(1)-v(1)| \\ &\leq k^2\|e\|^2 + 2k\|e'\|\,\|e\| + 2\|e'\|\,\|(u-v)'\| + 2k^2\|e\|\,\|u-v\|\,,\end{aligned}$$

where the terms at $x = 1$ have been estimated as in (4.4.9). We now use the so-called ε-inequality

$$2ab \leq \varepsilon^2 a^2 + \frac{b^2}{\varepsilon^2}\,, \qquad \varepsilon \neq 0\,,$$

to get the estimates

$$\begin{aligned}2k\,\|e'\|\,\|e\| &\leq \frac{1}{4}\,\|e'\|^2 + 4k^2\|e\|^2\,, \\ 2\,\|e'\|\,\|(u-v)'\| &\leq \frac{1}{4}\,\|e'\|^2 + 4\|(u-v)'\|^2\,, \\ 2k^2\,\|e\|\,\|u-v\| &\leq k^2\,\|e\|^2 + k^2\|u-v\|^2.\end{aligned}$$

Introducing these estimates into the previous inequality leads to

$$\|e'\|^2 \leq 6k^2\|e\|^2 + \frac{1}{2}\|e'\|^2 + 4\|(u-v)'\|^2 + k^2\|u-v\|^2\,,$$

which holds for all $v \in V_h$. Now we choose $v = u_I$. Then, using (4.4.10) and the approximation properties (4.4.2) we finally obtain

$$\frac{1}{2}\|e'\|^2 \leq 6k^2C_1^2(1+k)^2h^2\|e'\|^2 + 4|u-u_I|_1^2 + k^2\left(\frac{h}{\pi}\right)^2|u-u_I|_1^2\,,$$

and the statement of the theorem readily follows, completing the proof.

4.4.3 Conclusions

The constant $C_{\rm s}$ in (4.4.8) does not depend on h, k if h is small enough that $k^2h \ll 1$. Thus (4.4.7) holds only under the condition that

$$h \ll \frac{1}{k^2}\,,$$

by which we have specified the size of the threshold value h_0 (cf. Section 4.1.3) for the Helmholtz equation. Moreover, one can show (cf. Demkowicz [42], see also Ihlenburg and Babuška [74]) that $C_{\rm s} = 1$ asymptotically:

$$|u-u_h|_1 \to \inf_{v\in V_h}|u-v|_1, \quad h \to 0\,.$$

Practically speaking, the estimate (4.4.7) predicts that the convergence of the FE error on very fine meshes is similar to that of the interpolation error (and thus is controlled by the "rule of thumb"). However, for large k, meshes of size $h = O(k^{-2})$ are not used in practice. It is impossible to determine from the asymptotic order of convergence the size of the actual error on a mesh with $h = O(k^{-1})$.

Further, in practice one is interested in a rule that keeps the error below some tolerance level. However, a mesh design by $hk^2 \leq \alpha$ would lead to a *decreasing* error for increasing k. Indeed,

$$\tilde{e}_1 = \frac{|u - u_h|_1}{|u|_1} \leq \frac{C_{\mathrm{s}}}{|u|_1} \inf_{v \in V_h} |u - v|_1 \leq C_{\mathrm{s}} C h \frac{|u|_2}{|u|_1} \leq Chk \leq C\,\frac{\alpha}{k}\,.$$

Hence, asymptotically (which means here: on meshes with $k^2 h < \alpha$) the FE errors *tend towards 0* as k is increased. However, it is only required that the error *be bounded* for all k.

Finally, we remark that Theorem 4.9 can be proven also for the *hp*-version of the FEM; i.e., for $V_h = S_h^p(0,1)$, see Ihlenburg and Babuška [73].

4.5 Preasymptotic Error Estimates for the *h*-Version of the FEM

We have seen that the asymptotic error estimate does not, in general, characterize the error behavior (of FE solutions to the Helmholtz equation) in the range of engineering computations. In this section, we show estimates that hold under the assumption $kh < 1$. We will find that the error of Galerkin–FE solutions to Helmholtz problems always contains a so-called pollution term that depends on the wave number. While this term is negligible for low wave numbers or on extremely refined meshes, it does have a major influence for large wave numbers with standard mesh resolution.

4.5.1 Dispersion Analysis of the FE Solution

We set out in search of new error estimates that hold under the assumption $kh < 1$. We start from a spectral analysis of the stiffness matrix for Model Problem I. On a uniform mesh, this matrix can be written as

$$\mathbf{K} = \begin{bmatrix} 2S(kh) & R(kh) & & & \\ R(kh) & 2S(kh) & R(kh) & & \\ & & \ddots & & \\ & R(kh) & 2S(kh) & R(kh) & \\ & & R(kh) & S(kh) - ikh \end{bmatrix}, \tag{4.5.1}$$

with

$$R(kh) = -1 - \frac{(kh)^2}{6}\,, \qquad S(kh) = 1 - \frac{(kh)^2}{3}\,.$$

Except for the first and last lines, each line of this tridiagonal matrix corresponds to the homogeneous difference equation

$$R(kh)(u(x_i - h) + u(x_i + h)) + 2S(kh)u(x_i) = 0\,. \tag{4.5.2}$$

We seek a solution of this equation in the form

$$y(x_h) = e^{i\tilde{k}x_h}, \quad x_h \in X_h,$$

with unknown "discrete wave number" $\tilde{k}$. Then (4.5.2) transforms into the algebraic equation

$$R(kh)\lambda^{j-1} + 2S(kh)\lambda^j + R(kh)\lambda^{j+1} = 0\,,$$

where

$$\lambda = e^{i\tilde{k}h}\,.$$

The solutions of this equation,

$$\lambda_{1,2} = -\frac{S(kh)}{R(kh)} \pm \sqrt{\frac{S^2(kh)}{R^2(kh)} - 1}\,,$$

are

$$\text{(i)} \quad \text{complex} \quad \text{if} \quad \left|\frac{S(kh)}{R(kh)}\right| < 1\,,$$

$$\text{(ii)} \quad \text{real} \quad \text{if} \quad \left|\frac{S(kh)}{R(kh)}\right| \geq 1\,.$$

From the definition of λ we see that $\tilde{k}$ is either real (in case (i)) or pure complex (case (ii)). Thus case (i) describes a propagating wave, whereas case (ii) describes an evanescent wave. Solving the inequality $|S(kh)/R(kh)| < 1$, we see that the propagating case corresponds to the bound

$$h \leq \frac{\sqrt{12}}{k}\,. \tag{4.5.3}$$

Assuming that this inequality is satisfied, we can compute the discrete wave number in terms of kh, setting

$$\cos(\tilde{k}h) = -\frac{S(kh)}{R(kh)} \tag{4.5.4}$$

to obtain

$$\tilde{k} = \frac{1}{h}\arccos\left(-\frac{S(kh)}{R(kh)}\right). \tag{4.5.5}$$

Writing the Taylor expansion of the right-hand side, we find that

$$\tilde{k} = k - \frac{k^3 h^2}{24} + O(k^5 h^4). \tag{4.5.6}$$

Three conclusions can be made at this point:

(1) The discrete solutions are *dispersive*; i.e., their phase velocity depends on the frequency ω (recall from the discussion in Section 1.1.2. that the exact solution is nondispersive).

(2) The (piecewise linear) FE solution is a propagating wave, provided that the *minimal resolution* condition (4.5.3) is satisfied. Since $\sqrt{12} \approx \pi$, this condition means that a wavelength should be resolved by more than two elements.

(3) If the discrete solution is a propagating wave, then the phase velocity of this "numerical wave" differs from that of the exact wave. In our case, the phase difference is characterized by (4.5.6).

4.5.2 The Discrete inf–sup Condition

We consider again Model Problem I. The discrete inf–sup constant is of the same order $O(k^{-1})$ as the continuous constant (cf. (4.3.1)).

Theorem 4.10. *Let $V_h = S_h(0,1) \subset H^1_{(0}(0,1)$, and let $b : V_h \times V_h \to \mathbf{C}$ be the sesquilinear form defined in (4.2.3). Then the inf-sup constant*

$$\beta_h = \inf_{0 \neq u \in V_h} \sup_{0 \neq v \in V_h} \frac{|b(u,v)|}{|u|_1 |v|_1}$$

satisfies

$$\frac{C_1}{k} \leq \beta_h \leq \frac{C_2}{k} \tag{4.5.7}$$

for constants C_1, C_2 not depending on k or h.

The proof is similar to that of the continuous case, using techniques from the theory of finite difference methods. We refer to [72] for details.

Inserting $\beta_h = O(k^{-1})$ into the error estimate (4.1.13) and recalling that the continuity constant $M = O(k^2)$, we obtain the bound

$$|u - u_h|_1 \leq C(1 + k^3) h \, |u|_2 \,, \tag{4.5.8}$$

with some constant C independent of h, k. Dividing by $|u|_1$, we conclude that the relative error is bounded as

$$\tilde{e} := \frac{|u - u_h|_1}{|u|_1} \leq C(1 + k^3) kh \,. \tag{4.5.9}$$

Here, we have assumed again that the solution is oscillatory in the sense of (4.4.3); i.e., an extra power of k is introduced in each differentiation. The

bound in estimate (4.5.9) consists of two terms. The first term is of order hk, representing the interpolation error (cf. (4.4.5)). However, for $k \gg 1$, the error bound is dominated by the second term of order $O(k^4 h)$. This term estimates the numerical pollution in the error. We will discuss this issue in more detail in Section 4.6.

Computational experiments show that though the estimate (4.5.9) is too pessimistic (with respect to the power to which the wave number k enters the pollution term), it does reflect the nature of the error for Helmholtz problems. Let us explain at this point why this estimate may well be too pessimistic: In (4.5.9) we have used the stability constant $M\beta_h^{-1}$, which is obtained as the ratio of the continuity constant and the inf–sup constant. Both these constants are independently valid for any function in the spaces $H^1_{(0}(\Omega)$ or $S_h(0,1)$, respectively.

Consider, for example, the linear function $u = x \in H^1_{(0}(\Omega)$. The norm of this function is

$$|u|_1 = \left(\int_0^1 |u'|^2 \, dx \right)^{1/2} = 1 \, .$$

On the other hand, we have

$$|b(x,x)| = \left| \int_0^1 dx - k^2 \int_0^1 x^2 dx - ik \right| = \frac{\sqrt{9 + 3\,k^2 + k^4}}{3} = O(k^2) \, .$$

This shows that the continuity constant is, in general, no better than $O(k^2)$. However, for error estimation we are not interested in a bound that holds for all members of the Sobolev space $H^1_{(0}(\Omega)$ at the same time. For instance, the linear function of the example above is not a solution of the boundary value problem (4.2.1). Quite to the contrary, the terms $(u', v') - k^2(u, v)$ cancel for the solutions of this problem. Thus we can expect finer error estimates from employing solution-specific stability constants such as those derived in Theorem 4.4.

4.5.3 A Sharp Preasymptotic Error Estimate

Using the discrete wave number, we can construct a discrete Green's function (cf., e.g., Samarski [106])

$$G_h(x,s) = \frac{1}{h \sin \tilde{k} h} \begin{cases} \sin \tilde{k} x \left(A \sin \tilde{k} s + \cos \tilde{k} s \right) \, , & x \le s \, , \\ \sin \tilde{k} s (A \sin \tilde{k} x + \cos \tilde{k} x) \, , & s \le x \le 1 \, , \end{cases} \tag{4.5.10}$$

where

$$A = \frac{(kh)^2 \sin \tilde{k} \cos \tilde{k} + i\sqrt{12}\sqrt{12 - (kh)^2}}{12 - (kh)^2 \cos^2 \tilde{k}} .$$

Obviously, $|A|$ is bounded independently of $\tilde{k}$ if $kh \leq \alpha < \sqrt{12}$. Employing the discrete Green's function, the solution is now written as

$$u_h(x_i) = \sum_{j=1}^{n} G_h(x_i, s_j) r_h(s_j)\,, \tag{4.5.11}$$

where the discrete right-hand side is computed in the standard way as the scalar product of the data f and the nodal shape functions:

$$r_h(s_j) = h \int_0^1 f(s)\phi_j(s)\,ds\,.$$

Remark 4.11. Relating (4.5.11) to the FE solution procedure, we see that

$$[\,[G_h(x_i, s_j), \quad j = 1, \ldots, n], \, i = 1, \ldots, n]$$

is precisely the inverse of the stiffness matrix $\mathbf{K}$.

Using the discrete Green's function, we can show the following estimate ("discrete" stability).

Lemma 4.12. *Let $V_h := S_h(0,1)$ be the space of piecewise linear functions on a uniform partition of $\Omega = (0,1)$ with mesh size h, and let $u_h \in V_h$ be the FE solution to the variational problem (4.2.2) for given data $f \in L^2(0,1)$. Then, if $hk \leq 1$, there exists a constant C not depending on h and k such that*

$$|u_h|_1 \leq C\|f\| \tag{4.5.12}$$

holds.

The proof of this lemma employs some techniques from the finite difference theory. Details can be found in [72].

Observe that (4.5.12) holds under the condition $hk < 1$. We will call bounds with this constraint "preasymptotic", distinguishing the results from the asymptotic case $k^2 h \ll 1$. Recall that the meshes designed by the "rule of thumb" are in the preasymptotic range.

We are ready to derive an error estimate for the FE error $u - u_h$. We start from the trivial inequality

$$|u - u_h|_1 \leq |u - u_I|_1 + |u_I - u_h|_1\,,$$

where u_I denotes again the piecewise linear nodal interpolant of u. Let us show that $z := u_h - u_I$ is the solution of the variational problem

$$b(z, v) = k^2(u - u_I, v), \qquad \forall v \in V_h\,. \tag{4.5.13}$$

Indeed, adding and subtracting the exact solution, we get

$$b(z, v) = -b(u - u_h, v) + b(u - u_I, v) = b(u - u_I, v)$$

by the b-orthogonality of $u - u_h$. The right-hand side is

$$b(u - u_I, v) = \int_0^1 (u - u_I)'\bar{v}' - k^2 \int_0^1 (u - u_I)\bar{v} - ik(u(1) - u_I(1))\bar{v}(1)$$

The boundary term on the right is zero since $u = u_I$ at all nodes (in particular, at the node $x = 1$). Further, the first integral on the right also vanishes. To see this, one simply integrates by parts, integrating the first and differentiating the second member of the inner product. The term v'' is zero in all element interiors, since v is piecewise linear. The jump terms at the nodes vanish, since u_I is the interpolant.

Applying now the stability bound (4.5.12) to (4.5.13) and inserting the result into the first inequality, we have

$$\begin{aligned} |u - u_h|_1 &\leq |u - u_I|_1 + C\,k^2 \|u - u_I\| \\ &\leq (1 + Ck^2 h)\,|u - u_I|_1\,, \end{aligned}$$

where the second line follows from the first by the approximation properties (4.4.2). Taking into account that the interpolant u_I is the best approximation of u in the H^1-seminorm, we have shown the following theorem.

Theorem 4.13. *Let $f \in L^2(0,1)$ and let $u \in V$, $u_h \in V_h$ be the exact and FE solutions, respectively, of (4.2.2), where the spaces are defined as in Theorem 4.9. Then, if $hk < 1$, the error satisfies*

$$|u - u_h|_1 \leq \left(1 + C\,k^2 h\right) \inf_{v \in V_h} |u - v|_1\,, \tag{4.5.14}$$

where C is a constant not depending on k, h.

Remark 4.14. Note that the asymptotic result (4.4.7) directly follows from (4.5.14) if $k^2 h \ll 1$. Thus Theorem 4.13 is a generalization (for the special case of Model Problem I) of the well-known asymptotic theory.

Let us conclude by deriving the preasymptotic bound for the relative error. Introducing the approximation properties of the interpolant and assuming oscillatory behavior of the solution (i.e., $|u|_2/|u|_1 = O(k)$), we find that $\tilde{e}_1$ satisfies

$$\boxed{\tilde{e}_1 \leq C_1 hk + C_2 k^3 h^2\,, \qquad hk < 1.} \tag{4.5.15}$$

The second term in (4.5.15) is significantly smaller (for large k) than the term $C_2 k^4 h$ in the corollary of the inf–sup condition; cf. (4.5.9).

We will now show with computational results that the estimate (4.5.14) is sharp. Considering the relative error in H^1-norm, it turns out that the upper bound in (4.5.15) cannot be further improved: a pollution term of the size $O(k^3 h^2)$ is indeed seen in numerical experiments.

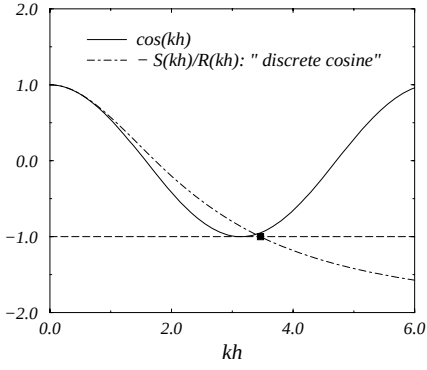


FIGURE 4.7. Discrete and exact cosine.

4.5.4 Results of Computational Experiments

We consider Model Problem I with datum $f = 1$. The FE meshes X_h are uniform with mesh size $h = N^{-1}$.

Discrete Wave Number and Phase Lag:

We first illustrate the dispersion relation (4.5.6). In Fig. 4.7, we show the "discrete cosine" $\cos \tilde{k}h$, computed from (4.5.4), compared to the exact $\cos kh$. The FE solution is evanescent if $kh > \sqrt{12}$. The cutoff value $kh = \sqrt{12}$ is marked in the plot with a fat dot. For normalized discrete wave numbers below the cutoff value, the finite element solution is a propagating wave with a dispersive discrete wave number. We consider only this case in the following.

The phase lag of the FE solution with respect to the exact solution, as quantified in (4.5.6), is illustrated in Fig. 4.8.

Error of the Best Approximation:

The best approximation of the exact solution u in the H^1-seminorm $|\cdot|_1$ is the nodal interpolant u_I; cf. (4.4.6). The convergence of the interpolation error is shown for different wave numbers in Fig. 4.9. The log-log plot shows the predicted convergence rate of N^{-1}. The error stays at 100% on coarse meshes and starts to decrease at a certain mesh size, which we call the *critical number of degrees of freedom (DOF)* for approximability. More precisely, for any fixed k and f the critical number of DOF is the minimal number $N(k, f)$ for which, first, $\tilde{e}_1(n, k) < 1$ and, second, $\tilde{e}_1(n, k)$ is monotone decreasing with respect to n for $n > N(k, f)$.

The critical number of DOF for approximability is determined by the rule that the mesh size for interpolation by piecewise linear functions should be smaller than one-half of the wavelength of the exact solution, i.e. $hk < \pi$.

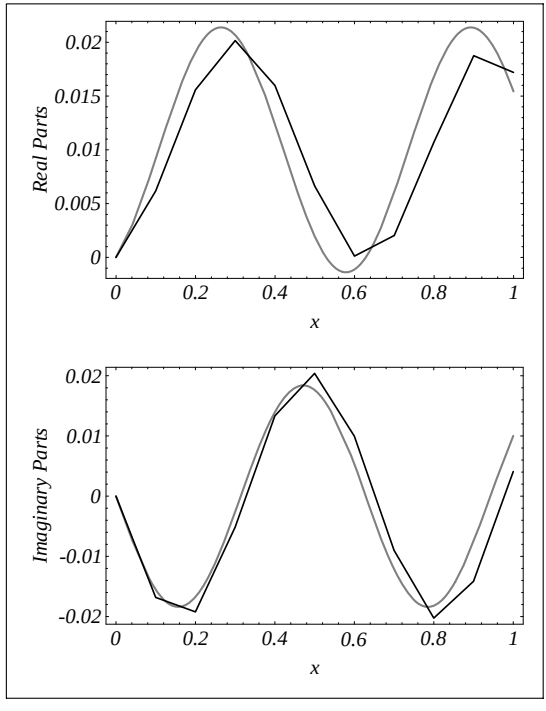


FIGURE 4.8. Phase lag of the finite element solution: exact and finite element solution for $k = N = 10$.

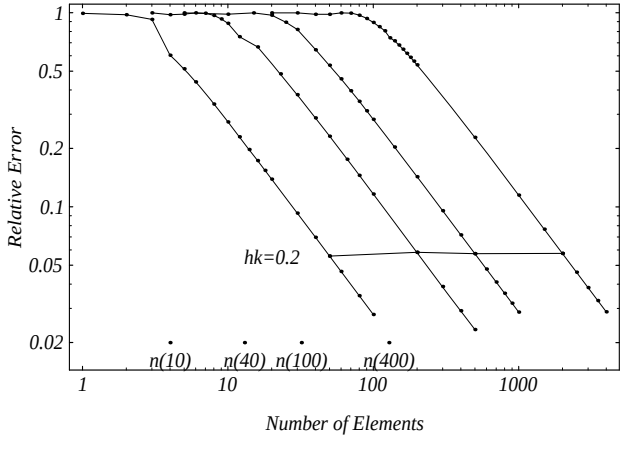


FIGURE 4.9. Relative error of best approximation in the H^1-seminorm, wave numbers (from left to right) $k = 10, 40, 100, 400$.

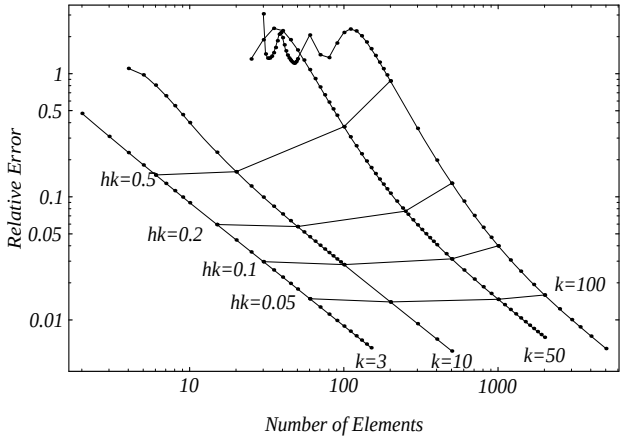


FIGURE 4.10. Relative error of the FE solution in the H^1-seminorm for $k = 3, 10, 50, 100$. The horizontal lines connect points with hk =constant.

The critical point N_0, computed from

$$N_0 = \left[\frac{k}{\pi}\right] \tag{4.5.16}$$

is plotted for different k. The positions of these points approximately coincide with the beginning of convergence on all curves. The figure also shows that the error of the best approximation is controlled by the magnitude hk. For illustration, the points $hk = 0.2$ are connected in the convergence curves for the different k. The connecting line neither increases nor decreases significantly with the change of k. Applying the "rule of thumb" $\lambda/h \approx 10$ thus guarantees a constant (with respect to k) error of about 17%.

TABLE 4.1. Resolution needed to maintain a relative error of 10% in the H^1-seminorm.

k	100	200	300	400	600	800	1000
n	38	57	63	82	94	107	120

Finite Element Error:

The relative error of the finite element solution for different k is plotted in Fig. 4.10. Unlike the error of the best approximation, the FE error is not controlled by the magnitude of hk. We see that the lines hk=constant in Fig. 4.10 increase with k. In Table 4.1 we display the resolution needed to maintain an error of 10% for increasing k. Clearly, the rule kh =constant is not sufficient to control the FE error. We observe a pollution effect at large wave numbers.

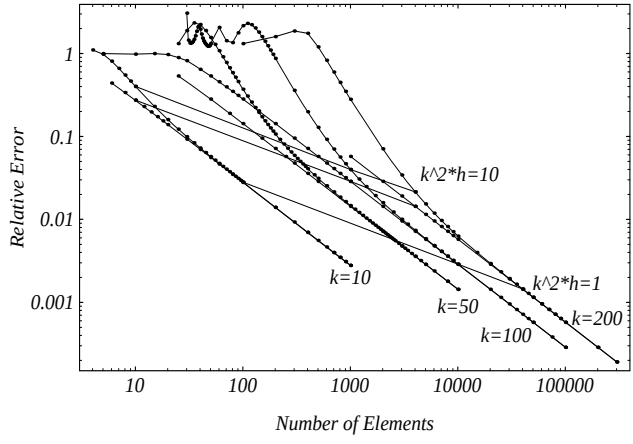


FIGURE 4.11. Relative error of the FE solution compared with the best approximation (BA) in the H^1-seminorm. For each k, the lower lines show the BA error. Points with k^2h=constant on the BA or FE lines, respectively, are connected.

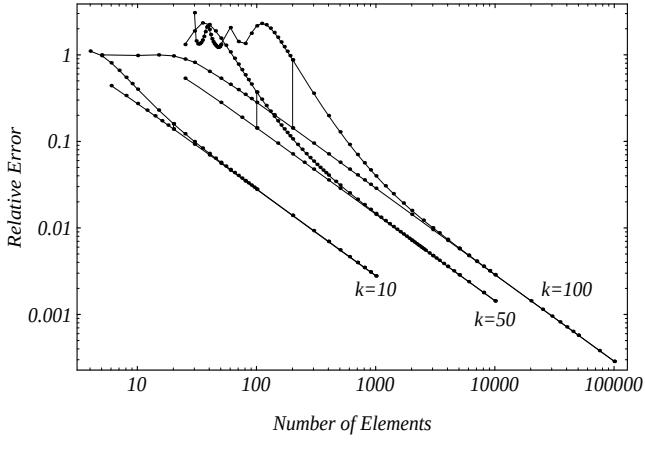


FIGURE 4.12. Relative error of the FE solution compared to the best approximation in the H^1-seminorm: stability constant $C(k)$ in the preasymptotic range (kh=constant).

In Figures 4.11 and 4.12 we illustrate the error behavior, respectively, in the asymptotic and in the preasymptotic range. The plots show the relative FE error $|u - u_h|_1/|u|_1$ and BA (best approximation) error $|u - u_{\mathrm{ba}}|_1/|u|_1$ for different wave numbers k. Points with k^2h=constant on the FE and BA error lines, respectively, are connected in Fig. 4.11. Along the line corresponding to $k^2h = 10$, we observe a constant (with respect to k) distance between the FE and BA error lines. Since we have plotted in the log-log-scale, this corresponds to a constant ratio $C_{\mathrm{s}} = |u - u_h|_1/|u - u_{\mathrm{ba}}|_1$ as established in the quasioptimal estimate (4.4.7) of Theorem 4.9. This shows that an assumption on k^2h seems to be necessary for asymptotic convergence. For $k^2h = 1$ the distance is indistinguishable within the resolution of the plot, corresponding to the fact that $C_{\mathrm{s}} \to 1$ as $h \to 0$; cf. Section 4.4.3. We also observe that the error *decreases* along the lines k^2h=constant.

On the other hand, the stability constant C_{s} *grows* with k in the preasymptotic range kh=constant, as shown in Fig. 4.12. The ratio $|u-u_h|_1/|u-u_{\mathrm{ba}}|_1$ now behaves as predicted by the estimate (4.5.14) of Theorem 4.13.

4.6 Pollution of FE Solutions with Large Wave Number

The error estimate (4.5.15) establishes a bound of order $O(k^3h^2)$ for the relative error of the FE solution to Model Problem I. Computations show that this upper bound is reached for large k; see Fig. 4.13.

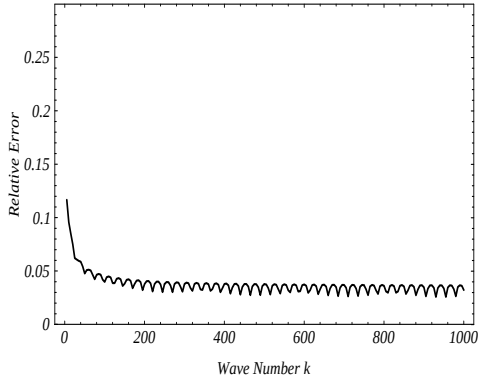


FIGURE 4.13. Relative error of the FE solution in the H^1-seminorm as function of wave number k with constraint $k^3h^2 = 1$.

In this context, we speak of numerical pollution in the FE error: viewing the error of interpolation as the "natural" error that is inherent to any approximation, we interpret the second term in (4.5.15) as a "pollution" of this natural error. We will give a precise definition in the following. We also

address the influence of boundary conditions on the convergence behavior and discuss error estimation in the L^2-norm.

4.6.1 Numerical Pollution

Definition: Consider a Helmholtz problem with wave number k on a normed space V. Let $u \neq 0 \in V$ and $u_h \in V_h \subset V$, respectively, be the exact and the finite element solution, and assume that an estimate of the form

$$\tilde{e} = \frac{\|u - u_h\|_V}{\|u\|_V} \leq C(k) \inf_{v \in V_h} \frac{\|u - v\|_V}{\|u\|_V} \tag{4.6.1}$$

holds. Then, if C can be written in the form

$$C(k) = C_1 + C_2 k^\beta (hk)^\gamma \,, \tag{4.6.2}$$

where $\beta > 0, \quad \gamma \geq 0$, and C_1, C_2 are independent of h and k, then the finite element solution is said to be *polluted*, and the term $C_2 k^\beta (hk)^\gamma$ is called *pollution term*.

Remark 4.15. Generally, one speaks of pollution if the error consists of two parts, one that is locally determined and another that is of nonlocal nature. For instance, in FE computations for elasticity problems, pollution occurs in regions away from a singularity (corner, crack, point load) if an error from insufficient resolution at the singularity spreads into the domain of computation, superposing the local interpolation error also in areas that are distant from the singularity.

Comparing (4.5.15) with (4.5.6), we see that the pollution term for the error in Helmholtz problems has the size of the phase lag. Indeed, in the proof of the error estimate, the pollution term results from the evaluation of $\|u_h - u_I\|$, where obviously, u_I is in phase with the exact solution. The phase lag is a *global* effect that builds up over the whole domain of computation. This shows the similarity between the notions of pollution in positive definite and indefinite problems. The difference is that the pollution in our case cannot be related to an insufficient resolution of a *local* singularity but is caused by the dispersive character of the discrete wave number, which is a global property. This topic will be further discussed in the section on *a posteriori* error estimation.

The pollution term also determines the critical number of DOF for the finite element error, as Fig. 4.14 shows. The position of the critical points, as marked over the abscissa of the plot, has been computed from the formula

$$N_0 = \sqrt{k^3/24} \,. \tag{4.6.3}$$

This formula can be obtained by heuristic argument from a simple estimate of the amplitude error at the end of the interval [72].

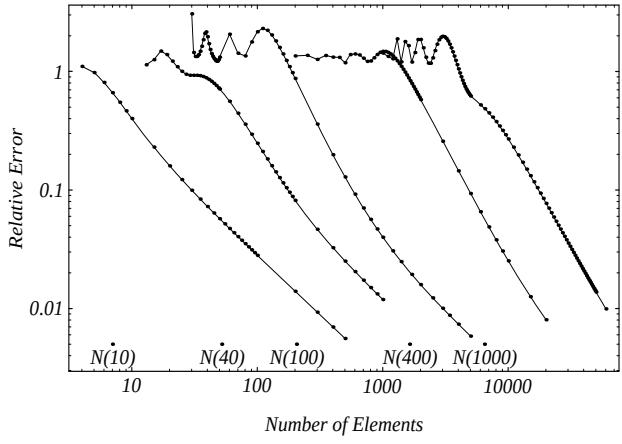


FIGURE 4.14. Relative error of the FE solution in H^1-seminorm and critical number of DOF.

Remark 4.16. The size of the critical number of DOF may be important, for instance, for the application of multilevel methods to the Helmholtz equation, if the coarse-grid solutions still have to be able to approximate the exact solution. Our experiments show that the coarsest grid should then be found from some constraint on k^3h^2.

4.6.2 The Typical Convergence Pattern of FE Solutions to the Helmholtz Equation

We have concluded from the preasymptotic estimate (4.5.14) that a constraint $k^2h < c$ is sufficient for optimal convergence

$$\|u - u_h\|_1 \leq C\|u - u_I\|_1 \,,$$

with a constant C not depending on k, h. The plot in Fig. 4.11 has shown (cf. the corresponding remarks in Section 4.5.4) that this constraint also seems to be necessary.

We are now in a position to describe precisely the FE error behavior throughout the region of convergence. In Fig. 4.15, the finite element error and the interpolation error are plotted for $k = 100$. We distinguish four intervals on the abscissa, where the number of elements N is shown:

(1) $N \leq n_o$: The point

$$n_0 = \frac{k}{\pi} \approx 32$$

is the "limit of resolution" (two elements per wavelength, $hk = \pi$). To the left of n_0, the error of the best approximation is 100%.

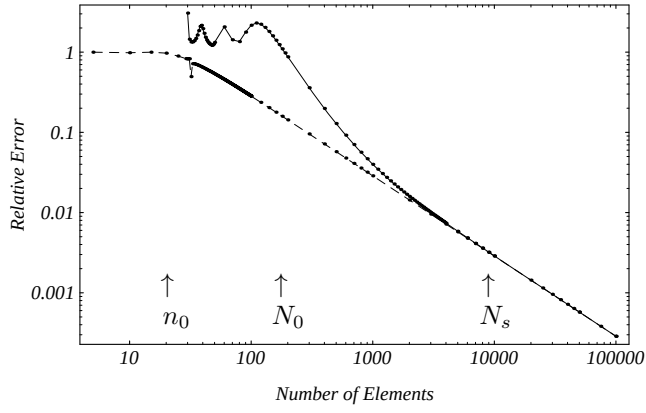


FIGURE 4.15. Convergence of relative errors of FE solution (continuous line) and best approximation (dashed line) for $k = 100$.

(2) $n_0 < N \leq N_0$: The point

$$N_0 = \sqrt{k^3/24} \approx 200$$

is the minimal meshsize for which the FE error is (and stays, if the mesh is refined) below 100%. In this interval, the error of the best approximation goes down at the rate -1 as h decreases, whereas the error of the finite element solution oscillates with amplitudes of more than 100%. The finite element solution is well-defined, but it does not approximate the exact solution. The pollution term in the error estimate (4.5.15) is dominant in this region. For the "rule of thumb" $\lambda/n = 10$ (read at $N = 5k/\pi \approx 160$ in Fig. 4.15), the minimal error is $\approx 15\%$, whereas the finite element error still exceeds 100%.

(3) $N_0 < N \leq N_s$: To the right of $N = N_s$, the influence of the pollution term in the FE error is negligible. The point $N = N_s$ marks the beginning of asymptotic convergence; hence

$$N_s = \frac{1}{c}k^2,$$

where, theoretically, c is a small constant. Numerical experience shows that c need not be small, but a constraint on k^2h is needed to ensure quasi-optimality in the H^1-norm. To the left of N_s, the convergence behavior of the finite element solution is still governed by the pollution term. Note that the prevalence of this term leads to a *superoptimal* rate of convergence: one observes a decay N^{-2} compared with N^{-1} for the best approximation.

(4) $n > N_s$: The condition for quasi-optimal convergence of the finite element solution is satisfied, independent of k.

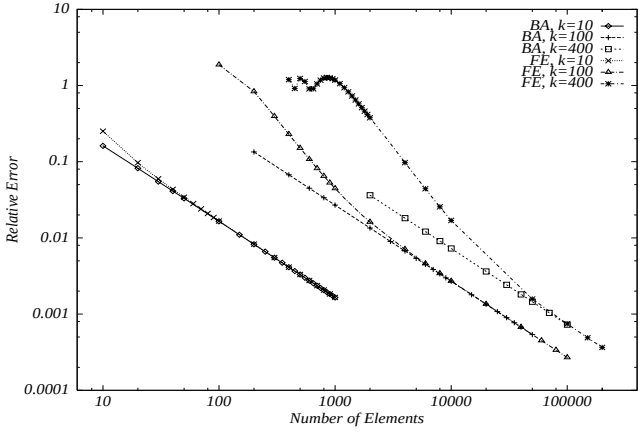


FIGURE 4.16. Errors of finite element solution (FE) and best approximation (BA) for Sommerfeld boundary conditions.

The third region is generally the region of practical interest. The finite element error is sufficiently small, while the mesh is still reasonably coarse. In particular, the FE error lies within this region if one constrains the magnitude of k^3h^2; cf. estimate (4.5.15). With a constraint of this form, reliability of the finite element solution is assured for all k.

We remark that similar behavior of the errors has been observed in computations for a one-dimensional example with highly irregular mesh; cf. Babuška and Sauter [17].

4.6.3 Influence of the Boundary Conditions

In order to test the influence of different boundary conditions on the behavior of the FE solution, we impose a Sommerfeld condition instead of the Dirichlet condition at $x = 0$. We then get the modified variational equality

$$b^*(u, v) = (u', v') - k^2(u, v) - ik\langle u, v\rangle = (f, v), \tag{4.6.4}$$

with

$$\langle u, v\rangle = u(0)\bar{v}(0) + u(1)\bar{v}(1)\,.$$

No Dirichlet condition is imposed, and hence the solution is sought in $V = H^1(\Omega)$. In this space, the H^1-seminorm is not a norm. Instead we use

$$\|u\|_* = \left(\|u'\|^2 + k^2|u(0)|^2\right)^{1/2}\,,$$

which is equivalent to the H^1-norm. The convergence behavior of the finite element solution and the best approximation is shown in Fig. 4.16 for different wave numbers. The convergence behavior is similar to that in the case of the Dirichlet condition. Further details can be found in Ihlenburg and Babuška [74].

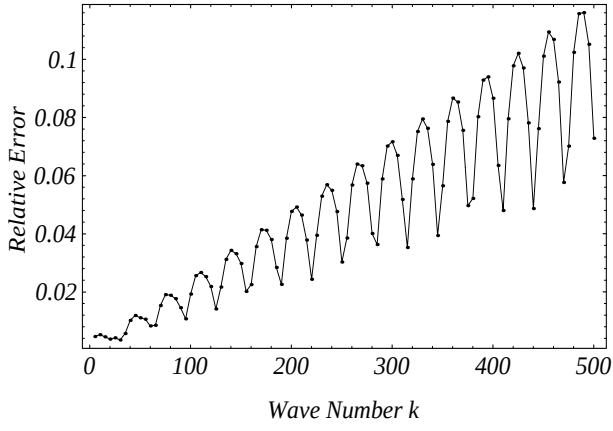


FIGURE 4.17. Relative error of the FE solution in the l^2-norm for constant resolution with $kh = 0.2$.

4.6.4 Error estimation in the L^2-norm

So far, estimates have been given in the H^1-norm. This norm measures both the function and its derivative, as opposed to the L^2-norm, which measures only the function itself.

An estimate for the error in the L^2-norm is obtained in the same way as for the H^1-norm. Assuming again that $hk < 1$, one can show[3] the discrete stability bound

$$\|u_h - u_I\| \leq Ck\|u - u_I\|$$

with C independent of h, k. Thus

$$\|u - u_h\| \leq (1 + k)\|u - u_I\| .$$

Inserting on the right-hand side the interpolation estimate from (4.4.2), we arrive at

$$\|u - u_h\| \leq (1 + Ck) \left(\frac{h}{\pi}\right)^2 |u|_2 . \tag{4.6.5}$$

For an oscillatory solution, $|u|_2/\|u\| = O(k^2)$, and hence we can estimate the relative error of an oscillatory solution as

$$\tilde{e}_0 := \frac{\|u - u_h\|}{\|u\|} \leq Ck(kh)^2 , \tag{4.6.6}$$

where C is a constant that does not depend on k, h.

The estimate (4.6.6) states that the L^2-errors are dispersive also in the asymptotic range. That is, on meshes designed by the "rule of thumb," the

[3]Recall that $z = u_h - u_I$ solves the Helmholtz problem with the right-hand side $k^2(u - u_I)$, and compare the continuous stability bound (4.3.9).

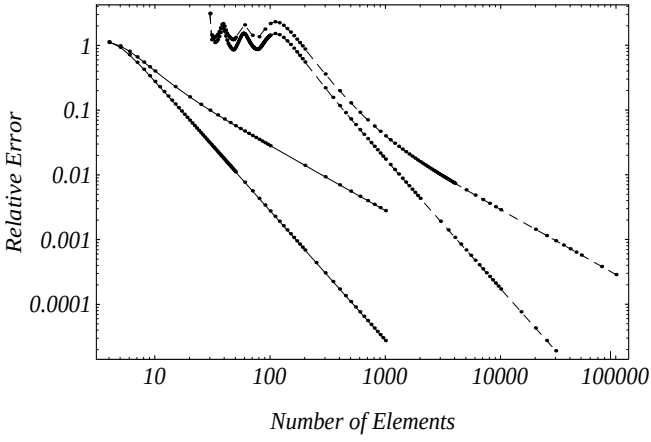


FIGURE 4.18. Relative errors of the FE solution in the H^1-norm and in the l^2-norm for $k = 10$ and $k = 100$.

error grows linearly with k also for small h. This is confirmed in numerical experiments. We consider again the FE solution of Model Problem I. In Fig. 4.17, we show the growth of the relative error in the l^2 vector norm on meshes with resolution $n = 10\pi$. In Fig. 4.18, the convergence of the relative error in the H^1-seminorm is compared to the relative error in the discrete l^2-norm. The errors for $k = 10$ (left two lines) and $k = 100$ are plotted. In each case, the lower lines in the plot correspond to the error in the l^2-norm, whereas the upper lines are the results in the H^1-norm.

The figure illustrates the typical convergence behavior for H^1-coercive forms. If the perturbation term is large, it governs the error behavior in the preasymptotic range (in both the H^1- and L^2-norms). However, while the positive definite term determines the asymptotic behavior in the H^1-norm, this effect is not observed if the error is measured in the norm of the perturbation (L^2-norm).

4.6.5 Results from 2-D Computations

Now consider Model Problem II. The square domain $\Omega = (0,1) \times (0,1)$ is partitioned into regular square elements. Bilinear shape functions are used for approximation. In the present example, boundary conditions are chosen such that the exact solution is a plane wave with the wave vector $\mathbf{k} = k\{\cos\theta, \sin\theta\}$ for $\theta = \pi/8$. The FE solutions were obtained with discretizations ranging from 16×16 to 1024×1024 elements.

In Fig. 4.19 the relative errors in the H^1-seminorm for $k = 10$, $k = 50$, and $k = 100$ are plotted. We observe that the error behavior is completely similar to the one-dimensional case; cf. Fig. 4.15. In Fig. 4.20, the errors of the finite element solution and best approximation in the H^1-norm and

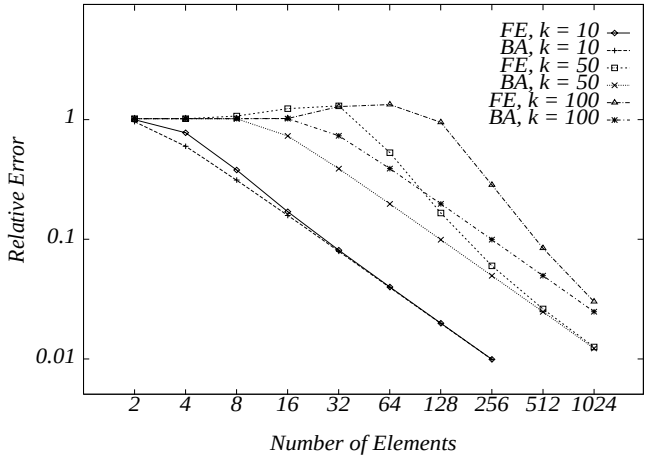


FIGURE 4.19. Relative errors in the H^1-norm for the 2-D Helmholtz problem: finite element solution (FE) vs. best approximation (BA) for $k = 10$, $k = 50$ and $k = 100$.

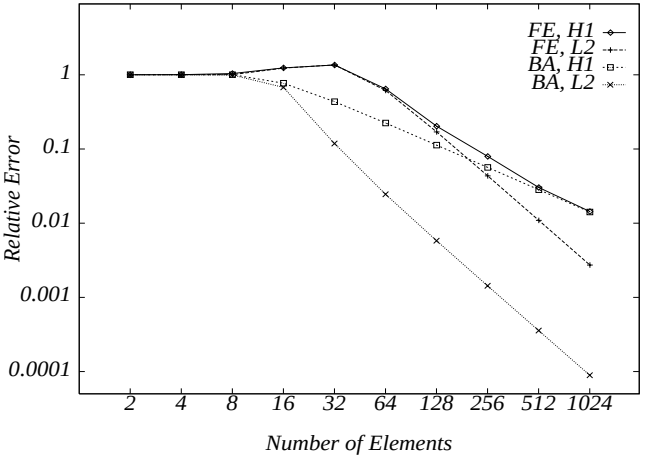


FIGURE 4.20. Relative errors for the 2-D Helmholtz problem: finite element solution (FE) vs. best approximation (BA) in the L^2-norm and the H^1-seminorm for $k = 50$.

the L^2-norm, respectively, are shown for $k = 50$. Again, the results are similar to those shown in Fig. 4.18. Similar results of computations in three dimensions are discussed in Gerdes and Ihlenburg [57].

4.7 Analysis of the *hp* FEM

We have seen that the relative error of the standard (h-version) Galerkin FEM can be written in the form (cf. (4.5.15))

$$\tilde{e}_1 = O(\theta + k\theta^2), \qquad \theta < 1\,, \tag{4.7.1}$$

with $\theta = kh$. The second term in this estimate characterizes numerical pollution, due to the phase difference between the exact and FE solutions. In this section we will investigate FE approximations with polynomial shape functions of degree $p \geq 2$. It will become evident that formula (4.7.1) can be generalized to higher-order polynomial approximations, except that now

$$\theta := \left(\frac{kh}{p}\right)^p\,. \tag{4.7.2}$$

Towards this goal, we outline the specifics of *hp*-approximation and describe in detail the FE solution procedure, including static condensation. This leads to the investigation of the phase lag by dispersion analysis. Finally, the preasymptotic error estimate (4.7.1), with θ from (4.7.2), is obtained. The material for this section is taken mostly from Ihlenburg and Babuška [73].

The analysis of the problem has some new ingredients compared to the h-version. First of all, we need information on the approximation of functions by higher-order polynomials. The error of approximation (measured in the H^1-norm) of the function u with polynomials of order p can be bounded, in general, by an expression of the form $h^p|u|_{p+1}$. To make use of this property, the derivatives of higher order should exist (in the weak sense). Hence we have to address the question of regularity. For Helmholtz problems, we need to know how the wave number k enters the stability estimates for functions of higher regularity. Also, it is possible to bound the error of approximation in negative norms (up to $1 - p$).

After studying these two issues — approximability and stability in higher-order norms — we turn to the actual error analysis. As for the h-version, the key lemma addresses discrete stability. We describe the process of static condensation that is used to separate local analysis from global analysis. In particular, the error analysis in negative norms is of local character.

4.7.1 hp-Approximation

We consider again Model Problem I with uniform meshes X_h of mesh size h. Hence by the *hp*-version we here mean h-convergence (i.e., error reduction

by global mesh refinement) with shape functions of degree p. The master element in the one-dimensional case is the interval $I = (-1, 1)$. The polynomials in $S^p(I)$ are written as linear combinations of the nodal shape functions

$$N_1(\xi) = \frac{1-\xi}{2} \ , \ N_2(\xi) = \frac{1+\xi}{2} \ ; \ -1 \leq \xi \leq 1 \, ,$$

and (if $p > 1$) the internal shape functions

$$N_l(\xi) = \phi_{l-1}(\xi) \ ; \ l = 3, 4, \ldots, p+1 \, ,$$

where the ϕ_l have been defined in (3.4.3). We denote by $S_h^p(\tau_i)$ the set of shape functions on the element $\tau_i = (x_{i-1}, x_i)$, where $x = \mathbf{Q}(\xi)$ is the standard linear map $I \to \tau_i$. The global space of *hp*-approximation is $S_h^p(\Omega) \subset H^1(\Omega)$. The approximation properties of the *h*-version (cf. (4.4.2)) can then be generalized. To write the corresponding theorem, we define antiderivatives of a function as follows. As usual, $f^{(m)}$, $m \geq 0$ denotes the mth derivative of a function f. Generalizing this for negative integers, we call any function F such that

$$\partial^m F = f$$

an mth antiderivative of f and write

$$F = f^{(-m)} \, .$$

The definition is specified for the interval $(0, 1)$ as follows. Let $f \in L^2(0, 1)$, then

$$f^{(-1)}(x) := -\int_x^1 f(t)dt \, ,$$

and similarly,

$$f^{(-m-1)}(x) := -\int_x^1 f^{(-m)}(t)dt \, .$$

Lemma 4.17. *Let l, p be integers with $1 \leq l \leq p$ and let $u \in H^{l+1}(0, 1)$. There exists an $s \in V_h = S_h^p(0, 1)$ such that*

$$\forall x_i \in X_h : \quad s^{(m)}(x_i) = u^{(m)}(x_i) \, , \quad m = -p+1, \ldots, 0 \, , \tag{4.7.3}$$

and

$$\|(u-s)^{(m)}\| \leq C_{\mathrm{a}}(l) C_{\mathrm{a}}(-m) \left(\frac{h}{2p}\right)^{l-m+1} |u|_{l+1} \, , \quad m = -p+1, \ldots, 1 \, , \tag{4.7.4}$$

hold, where C_{a} satisfies:

1. $C_{\mathrm{a}}(-1) = 1$ *(formal definition),*
2. $C_{\mathrm{a}}(0) = 1$,

3. C_{a} *decreases for* $0 \le l \le \sqrt{p}$,

4. C_{a} *increases for* $l > \sqrt{p}$ *and*

$$C_{\mathrm{a}}(p) = \max_{1 \le l \le p} C_{\mathrm{a}}(l) = \left(\frac{e}{2}\right)^p (\pi p)^{-1/4}. \tag{4.7.5}$$

The proof is based on Babuška et al. [14], where the cases $l = 0, 1$ are treated. Let τ_i be a finite element and let $I = (-1, 1)$. A function $u'(\xi) \in H^l(I) \subset L^2(I)$ can be expanded as

$$u'(\xi) = \sum_{i=0}^{\infty} a_i P_i(\xi)\,,$$

where $P_i(\xi)$ are the Legendre polynomials of order i and equality is understood in the L^2 sense. Setting

$$s'(\xi) := \sum_{i=0}^{p-1} a_i P_i(\xi)$$

and defining the integrals $(i = 0, 1, 2, \ldots)$

$$u^{(-i)}(\xi) = u^{(-i)}(1) - \int_{\xi}^{1} u^{(-i+1)}(\tau) d\tau\,, \tag{4.7.6}$$

$$s^{(-i)}(\xi) = u^{(-i)}(1) - \int_{\xi}^{1} s^{(-i+1)}(\tau) d\tau\,, \tag{4.7.7}$$

we now prove that (4.7.3) holds. First, let $i = 0$. Then from (4.7.6), (4.7.7) we have trivially $u(1) = s(1)$. Further, by definition,

$$\begin{aligned} u(-1) &= u(1) - \int_{-1}^{1} u'(\tau) d\tau = u(1) - \sum_{j=0}^{\infty} a_j \int_{-1}^{1} P_j(\tau) d\tau = u(1) - 2a_0 \\ &= u(1) - \int_{-1}^{1} s'(t) dt = s(-1) \end{aligned}$$

by (4.7.7). Now we integrate $u'(\xi)$. Using

$$P_i(\tau) = (P'_{i-1}(\tau) - P'_{i+1}(\tau))/(2i+1)\,,$$

we obtain

$$u(\xi) = u(1) + a_0(P_1(\xi) - P_0(\xi)) + \sum_{i=1}^{\infty} a_i \frac{P_{i+1}(\xi) - P_{i-1}(\xi)}{2i+1}.$$

Integrating once more (we write $U := u^{(-1)}$),

$$U(-1) = U(1) - \int_{-1}^{1} u(\tau)d\tau = U(1) - 2u(1) + 2a_0 + \frac{2a_1}{3}.$$

Obviously, the same result is obtained from the integration of the polynomial $s(\xi)$, since only the coefficient of P_0 influences the result of integration over the whole interval I. Using a similar argument we conclude that integration of the polynomial s on the one hand and the function u on the other hand leads to the same result exactly $p-1$ times. Indeed, by replacing repeatedly $P_i(\tau) = (P'_{i-1}(\tau) - P'_{i+1}(\tau))/(2i+1)$, we see that with the ith succesive integration of $u(\xi)$ or $s(\xi)$ the coefficient a_i enters the set of coefficients multiplying P_0. Since the norms of $u^{(i)}$ and $s^{(i)}$ depend only on the coefficient of P_0, both norms are equal until P_0 is multiplied by a_{p-1}; i.e., in general, $u^{(-p+1)}(\xi) = s^{(-p+1)}(\xi)$ and $u^{(-p)}(\xi) \neq s^{(-p)}(\xi)$. Thus nodal exactness, (4.7.3), is proved on an arbitrary element and hence it holds globally.

Let us now prepare the proof of estimate (4.7.4). With the above definitions, the error of approximation is

$$e'(\xi) := u'(\xi) - s'(\xi) = \sum_{i=p}^{\infty} a_i P_i(\xi)\,,$$

and from the orthogonality property of the Legendre polynomials we have

$$\|e'\|^2 = \sum_{i=p}^{\infty} \frac{2}{2i+1} a_i^2\,. \tag{4.7.8}$$

It can be proven (see Babuška et al. [14], Chapter 3) that s' is the best L^2-approximation to u' on I and the estimate

$$\|u' - s'\| \leq \frac{C_{\rm a}(l)}{p^l}\,|u|_{l+1} \tag{4.7.9}$$

holds for $0 \leq l \leq p$, where the constant $C_{\rm a}$ has the properties 2–4 of the statement. Integrating the error e', we get

$$\begin{aligned} e(\xi) &= \int_{\xi}^{1} (u'(t) - s'(t))\,dt \\ &= \sum_{i=p}^{\infty} a_i \int_{\xi}^{1} P_i(t)dt = -\sum_{i=p}^{\infty} \frac{a_i}{2i+1}\left(P_{i+1}(\xi) - P_{i-1}(\xi)\right)\,. \end{aligned}$$

After reordering, this is equivalently written as

$$e(\xi) = \sum_{i=p+1}^{\infty} b_i P_i(\xi) + \frac{a_p}{2p+1} P_{p-1}(\xi) + \frac{a_{p+1}}{2p+3} P_p(\xi)$$

with

$$b_i = \frac{a_{i+1}}{2i+3} - \frac{a_{i-1}}{2i-1}\,,$$

and the norm is

$$\|e\|^2 = \sum_{i=p+1}^{\infty} \frac{2}{2i+1} b_i^2 + \frac{a_p^2}{(2p+1)^2}\frac{2}{2p-1} + \frac{a_{p+1}^2}{(2p+3)^2}\frac{2}{2p+1}\,. \quad (4.7.10)$$

We apply the relation $(a-b)^2 \le 2a^2 + 2b^2$ to obtain (for $i \ge p+1$)

$$b_i^2 \le \frac{2a_{i-1}^2}{(2i-1)^2} + \frac{2a_{i+1}^2}{(2i+3)^2}\,,$$

and thus

$$\sum_{i=p+1}^{\infty} b_i^2 \frac{2}{2i+1} \le \frac{1}{2p^2}\sum_{i=p}^{\infty} a_i^2 \frac{2}{2i+1} + \frac{1}{2p^2}\sum_{i=p+2}^{\infty} a_i^2 \frac{2}{2i+1}$$

holds. Now taking into account the second and third member in the right-hand side of (4.7.10), we get

$$\|e\|^2 \le \frac{1}{p^2}\sum_{i=p}^{\infty} a_i^2 \frac{2}{2i+1}\,,$$

and hence

$$\|e\| \le \frac{1}{p}\|e'\|\,. \quad (4.7.11)$$

From (4.7.9) it then follows that

$$\|e\| \le \frac{C_{\mathrm{a}}(l)}{p^{l+1}}|u|_{l+1} \quad (4.7.12)$$

holds for $1 < l \le p$.

We conclude the local analysis by showing an orthogonality property for e. Since s' is the L^2-projection of u' on $S^{p-1}(I)$,

$$\int_{-1}^{1} (u'(\xi) - s'(\xi))\xi^m d\xi = 0 \quad (4.7.13)$$

holds for $m = 0, 1, \ldots, p-1$. We claim that $e(\xi) = \int_{\xi}^{1}(u'(t) - s'(t))dt$ is orthogonal to $S^{p-2}(I)$.

Indeed, for $m \ge 0$ we compute

$$\begin{aligned}
\int_{-1}^{1} e(\xi)\xi^m d\xi &= \int_{-1}^{1}\Big(\int_{\xi}^{1}(u'(t) - s'(t))dt\Big)\xi^m d\xi \\
&= \int_{-1}^{1}\Big((u'(t) - s'(t))\int_{-1}^{t}\xi^m d\xi\Big)dt \\
&= \frac{1}{m+1}\int_{-1}^{1}(u'(t) - s'(t))(t^{m+1}+1)dt\,,
\end{aligned}$$

which together with (4.7.13) proves that e is orthogonal to $S^{p-2}(I)$. By the back-transform $I \to \tau_i$ and summation over the elements we derive (4.7.4) for the H^1- and L^2-norms; i.e., the cases $m = 1, 0$ in (4.7.4).

It remains to prove (4.7.4) for dual norms. We apply a standard argument (cf. Schatz [109]). By definition, for $m \geq 1$,

$$\|e\|_{-m} = \sup_{v \in H^m_{(0}} \frac{(e,v)}{|v|_m}.$$

Let $P^m v \in S^m_h(\Omega)$ be the L^2-projection of $v \in H^m_{(0}$ on $S^m_h(\Omega)$. Then by orthogonality (4.7.13)

$$\|e\|_{-m} = \sup_{v \in H^m_{(0}} \frac{(e, v - P^{m-1}v)}{|v|_m}$$

holds for $1 \leq m \leq p-1$. Applying the Cauchy–Schwarz inequality and (4.7.12), we conclude for $1 \leq l \leq p$ and $1 \leq m \leq p-1$ the estimate

$$\begin{aligned} \|e\|_{-m} &\leq C_{\mathrm{a}}(l) \left(\frac{h}{2p}\right)^{l+1} |u|_{l+1}\, C_{\mathrm{a}}(m) \left(\frac{h}{2p}\right)^{m} \frac{|v|_m}{|v|_m} \\ &\leq C(l,m) \left(\frac{h}{2p}\right)^{l+m+1} |u|_{l+1}\,, \end{aligned}$$

where

$$C(l,m) = C_{\mathrm{a}}(l) C_{\mathrm{a}}(m) \leq \left(\frac{e}{2}\right)^{2p} (\pi p)^{-1/2}.$$

This completes the proof of the lemma.

Remark 4.18. We have shown that there exists an interpolant s such that $p-1$ antiderivatives of s interpolate the corresponding antiderivatives of u. These interpolants then satisfy the standard interpolation error bounds.

The present proof is given for approximation in one dimension. The error bounds in the negative norms (i.e., for the antiderivatives) are similarly obtained in two and three dimensions; see, e.g., Schatz [109].

4.7.2 Dual Stability

Besides the stability estimate (4.3.11), we will need yet another stability result, bounding lower norms of the solution by the corresponding norms of the data.

Lemma 4.19. *Let $\Omega = (0,1)$, and let $f \in L^2(\Omega)$ be a "bubble" function with the property $f^{(-i)}(0) = f^{(-i)}(1) = 0$ for $i = 1, \ldots, m$. Then the solution*

$u \in H^1_{(0}(\Omega)$ of (4.2.2) with datum f satisfies

$$|u|_1 \leq C_1 k^m \|f^{(-m)}\| + C_2 \|f^{(-1)}\|\,, \tag{4.7.14}$$

with C_1, C_2 independent of k.

Remark 4.20. The assumption on the data means that m integrals of f vanish at $x = 0$ (note that all integrals vanish at the endpoint by definition). Without this assumption we can prove only that

$$|u|_1 \leq C\ k\ \|f^{(-1)}\|\,;$$

cf. (2.4.19). In general, $|u|_1$ cannot be bounded by a term $C\|f^{(-1)}\|$ independently of k.

We give the proof for $m = 2$. The idea is to introduce a smoother kernel into the Green's function representation of $u(x)$. We define

$$K(x,s) := G(x,s) - H(x,s)\,,$$

where $H(x,s)$ is the Green's function for the auxiliary Dirichlet problem

$$\left\{ \begin{array}{l} \text{Find } u \in H^1_0(\Omega): \\ \qquad B_1(w,v) = \displaystyle\int_0^1 w'v' + k^2 \int_0^1 wv = (f,v), \qquad \forall v \in H^1_0(\Omega)\,. \end{array} \right. \tag{4.7.15}$$

Note that B_1 is a positive definite bilinear form. Hence

$$H_{xx}(x,s) - k^2\, H(x,s) = -\delta_s(x)\,, \tag{4.7.16}$$

$$G_{xx}(x,s) + k^2\, G(x,s) = -\delta_s(x)\,, \tag{4.7.17}$$

holds weakly. It follows that

$$K_{xx} = -k^2(G + H)\,.$$

Then, however, since $G, H \in H^1(\Omega)$ (as functions of x for any fixed s), we have $K \in H^3(\Omega)$. We write

$$u(x) = \int_0^1 K(x,s) f(s) ds + \int_0^1 H(x,s) f(s) ds := u_1(x) + u_2(x)\,. \tag{4.7.18}$$

The point is that K is a smooth kernel and H is the kernel for a V-elliptic problem. This problem has an inf–sup constant $\beta = 1$ in the energy norm (cf. Remark 2.18), and from the Lax-Milgram theorem it follows that

$$|u_2|_1 \leq |\!|\!| u_2 |\!|\!| \leq \|f^{(-1)}\|\,.$$

To estimate u_1, we integrate by parts to get

$$u_1'(x) = \int_0^1 K_x(x,s) f(s) ds = \int_0^1 K_{xss}(x,s) f^{(-2)}(s) ds,$$

(the boundary terms vanish due to the specific assumption on f). Thus, by the Cauchy–Schwarz inequality,

$$|u_1'(x)| \leq \|K_{xss}\| \|f^{(-2)}\|.$$

From the definition of the function K and the symmetry of the Green's functions, it follows that $K_{xss} = -k^2(G+H)_x$, and thus

$$\|K_{xss}\| \leq k^2 \left(\|G_x\| + \|H_x\| \right) .$$

It is straightforward to show that $\|G_x\|$ and $\|H_x\|$ are bounded independently of k, and we finally arrive at

$$|u_1'(x)| \leq C\, k^2 \|f^{(-2)}\| .$$

Together with the estimate of u_2, this proves the case $m = 2$.

For $m = 1$, the statement follows from the inf–sup condition. For $m \geq 3$, the proof proceeds as for $m = 2$; see [73].

4.7.3 FEM Solution Procedure. Static Condensation

In the following, we describe in detail the solution of the boundary value problem (4.2.1) by the *hp*-version of the FEM.

Step 1. Local Approximation and Static Condensation:

On any element τ_j, the trial function u and the test function v are written as scalar products of shape functions $\left\{ N_1^j, N_2^j, \ldots, N_{p+1}^j \right\}$ and the vectors of unknown coefficients $\{a^j\} = \left\{ a_1^j, a_2^j, \ldots, a_{p+1}^j \right\}^T$ and $\{b^j\} = \left\{ b_1^j, b_2^j, \ldots, b_{p+1}^j \right\}^T$, respectively. We assume a local numbering in which subscripts $1, 2$ correspond to the nodal modes, while subscripts $j \geq 3$ correspond to the bubble modes. Thus $a_1^j = u(x_{j-1}) := u_{j-1}, a_2^j = u(x_j) := u_j$, and $b_1^j = v_{j-1}, b_2^j = v_j$. The variational problem reduces locally (i.e., on interior elements τ_j) to

$$\{\bar{b}^j\}^T [B^j] \{a^j\} = \{\bar{b}^j\}^T \{r^j\} , \tag{4.7.19}$$

where the elements of the $(p+1) \times (p+1)$ square matrix $[B^j]$ are

$$b_{lm} = \int_{\tau_j} N_l^j(x)' \, N_m^j(x)' dx - k^2 \int_{\tau_j} N_l^j(x) \, N_m^j(x) dx - ikN_l^j(1) N_m^j(1) , \tag{4.7.20}$$

and the right-hand side is

$$r_l^j = \int_{\tau_j} f(x) N_l^j(x)) \, dx \, .$$

Now, decomposing

$$[B^j] = \begin{bmatrix} \left[B_{11}^j\right] & \left[B_{12}^j\right] \\ \left[B_{21}^j\right] & \left[B_{22}^j\right] \end{bmatrix}, \tag{4.7.21}$$

where $\left[B_{11}^j\right]$ is the left upper 2×2 submatrix of $[B^j]$, and assuming for the moment that $\left[B_{22}^j\right]$ is nonsingular, we define

$$[CB^j] = \left[B_{11}^j\right] - \left[B_{12}^j\right] \left[B_{22}^j\right]^{-1} \left[B_{21}^j\right]. \tag{4.7.22}$$

Then, by local variation of $\{b_3^j, \ldots, b_{p+1}^j\}^T$, we find

$$\{v_{j-1} \; v_j\} \left[CB^j\right] \begin{Bmatrix} u_{j-1} \\ u_j \end{Bmatrix} = \{v_{j-1} \; v_j\} \begin{Bmatrix} \tilde{r}_{j-1} \\ \tilde{r}_j \end{Bmatrix}, \tag{4.7.23}$$

where

$$\begin{Bmatrix} \tilde{r}_{j-1} \\ \tilde{r}_j \end{Bmatrix} = \begin{Bmatrix} r_1^j \\ r_2^j \end{Bmatrix} - \left[B_{12}^j\right] \left[B_{22}^j\right]^{-1} \begin{Bmatrix} r_3^j \\ \vdots \\ r_{p+1}^j \end{Bmatrix}. \tag{4.7.24}$$

On a uniform mesh, the local matrices $[CB^j]$ are identical on all elements and can be written in the form

$$[CB] = \begin{bmatrix} S_p(kh) & T_p(kh) \\ T_p(kh) & S_p(kh) \end{bmatrix}, \tag{4.7.25}$$

where S_p and T_p are rational polynomial functions.

Step 2. Global Assembling and Solution for u_h:

Enforcing continuity of the test functions at the nodal points of X_h leads to the usual set of linear equations

$$\mathbf{K}_p u_h^p = r_p \, .$$

The condensed stiffness matrix $\mathbf{K}_p$ is an $n \times n$ tridiagonal matrix

$$\mathbf{K}_p = \begin{bmatrix} 2S_p(kh) & T_p(kh) & & \\ T_p(kh) & 2S_p(kh) & T_p(kh) & \\ & \ddots & & \\ & T_p(kh) & 2S_p(kh) & T_p(kh) \\ & & T_p(kh) & S_p(kh) - ikh \end{bmatrix} \tag{4.7.26}$$

which is formally similar to the stiffness matrix of the h version. The vector on the right-hand side is

$$r_p = \left\{ \begin{array}{c} \tilde{r}_1 \\ \vdots \\ \tilde{r}_{j-1} + \tilde{r}_j \\ \vdots \\ \tilde{r}_n \end{array} \right\} . \tag{4.7.27}$$

Step 3. Local "Decondensation:"

The equation

$$\left[B_{22}^j\right] \left\{ \begin{array}{c} a_3^j \\ \vdots \\ a_{p+1}^j \end{array} \right\} = \left\{ \begin{array}{c} r_3^j \\ \vdots \\ r_{p+1}^j \end{array} \right\} - \left[B_{21}^j\right] \left\{ \begin{array}{c} u_{j-1} \\ u_j \end{array} \right\} \tag{4.7.28}$$

can be inverted, provided that $\left[B_{22}^j\right]$ is regular, to determine locally the bubble modes of the finite element solution. Adding together the local modes and the bubble modes, the finite element solution is obtained, completing the procedure.

Remark 4.21. The matrix $[B_{22}]$ is singular at the discrete eigenvalues of the condensation, determined by the eigenvalue problem

$$w'' + \lambda^2 w = 0\,, \qquad w(0) = w(1) = 0\,.$$

The exact solutions are $\lambda = \pi, 2\pi, \ldots$ The discrete eigenvalues are obtained if the problem is solved over the subspace of all "bubble" polynomials of order $\leq p$. In one dimension, there exist $p-1$ "bubble" polynomials ($p = 2, 3, \ldots$), and hence the eigenvalue problem has $p-1$ solutions. We list the computed and exact eigenvalues for $p = 2, \ldots, 6$ in Table 4.2. Obviously the rule $hk \leq \pi$ excludes a breakdown of condensation.

4.7.4 Dispersion Analysis and Phase Lag

The homogeneous solutions of the interior difference equations

$$T_p(kh)u_h(x_{i-1}) + 2S_p(kh)u_h(x_i) + T_p(kh)u_h(x_{i+1}) = 0$$

are, similar to (4.5.2),

$$y_{h1} = \exp(i\tilde{k}x_h), \;\; y_{h2} = \exp(-i\tilde{k}x_h)\,,$$

TABLE 4.2. Singular values λ_h of the local stiffness matrix and exact eigenvalues (lower row) of the associated eigenvalue-problem for $p = 2, \ldots, 6$.

$p\,/\,i$	1	2	3	4	5
2	3.16228	-	-	-	-
3	3.16228	6.48074	-	-	-
4	3.14612	6.48074	10.1060	-	-
5	3.14612	6.28503	10.1060	14.1597	-
6	3.14159	6.28503	9.44318	14.1597	18.7338
$i * \pi$	3.14159	6.28319	9.42478	12.5664	15.7080

where the parameter $\tilde{k}$ is determined as a function of k, h, and p by

$$\cos(\tilde{k}h) = -\frac{S_p(kh)}{T_p(kh)}\,. \tag{4.7.29}$$

In Fig. 4.21, we plot the discrete cosine $\cos(\tilde{k}h)$, computed from (4.7.29), for $p = 1, \ldots, 6$. The plot for $p = 1$ is identical with the plot of Fig. 4.7. Recall that the discrete wave number $\tilde{k}$ is complex, and hence the numerical solution is evanescent, if $|\cos(\tilde{k}h)| = |S_p(kh)/T_p(kh)| > 1$. The magnitude of the cutoff frequency grows with the increase of approximation order p. However, the discrete wave number is complex on small intervals (see, for example, $kh \approx 7$ for $p = 3$) also before it reaches the cutoff frequency. The only way to avoid this situation completely is to keep the resolution below the first cutoff frequency $kh = \sqrt{12}$.

We now show a bound for the phase difference between the exact and FE solutions for the *hp*-version.

Theorem 4.22. *Let $p \geq 1$ and denote by $\tilde{k}$ the discrete wave number defined in (4.7.29). Then, if $hk < 1$,*

$$|\tilde{k} - k| \leq k\, C \left(\frac{C_{\mathrm{a}}(p)}{2}\right)^2 \left(\frac{hk}{2p}\right)^{2p} \tag{4.7.30}$$

where k is the exact wave number, C_{a} is the approximation constant from (4.7.4), and C does not depend on k, h, or p.

The proof can be found in [73]. The key idea is to use analytic shape functions for the represention of the exact solution. These functions t_1, t_2 are found from the local boundary value problems

$$t'' + k^2 t = 0 \quad \text{on } \tau_j\,, \tag{4.7.31}$$

with inhomogeneous local Dirichlet data

$$t_1(x_{j-1}) = 1, \quad t_1(x_j) = 0\,, \tag{4.7.32}$$

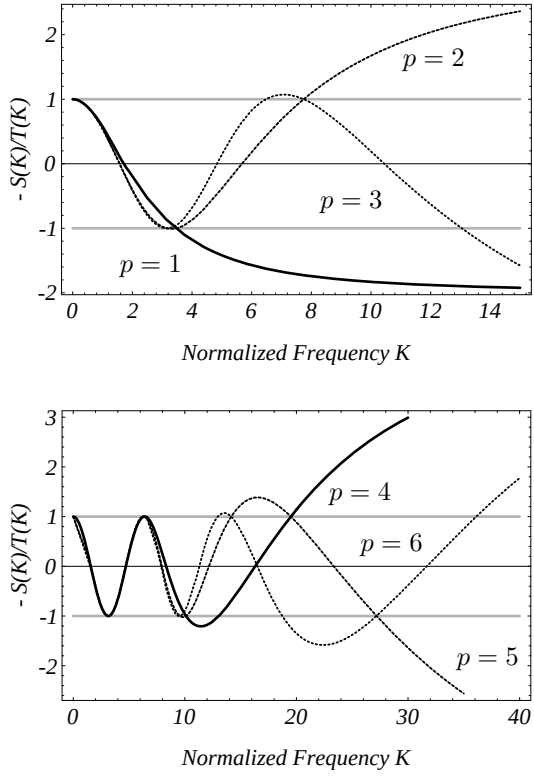


FIGURE 4.21. Cosine of the normalized discrete wave number vs. normalized frequency K.

or

$$t_2(x_{j-1}) = 0, \quad t_2(x_j) = 1\,, \tag{4.7.33}$$

respectively. Having computed t_1, t_2, the homogeneous solution of the Helmholtz equation can be written on each inner element $\tau_j \subset \Omega$ as

$$u(x) = u^1 t_1(x) + u^2 t_2(x)\,, \tag{4.7.34}$$

where $u^1 := u(x_{j-1}), u^2 := u(x_j)$ are the nodal values of u on X_h. The phase difference is then estimated by comparing the analytic with the finite element shape functions.

4.7.5 Discrete Stability

Now we have collected almost all prerequisites to showing the error estimates. The last step is to prove discrete stability estimates. It is easy to

show the following.

Lemma 4.23. *Let $u_h \in S_h^p(0,1)$ be the finite element solution to the variational problem (4.2.2) with datum $f \in L^2(\Omega)$. Assume that $hk \le \alpha < \pi$. Then*

$$\|u_h'\| \le C\|f\| \tag{4.7.35}$$

holds with a constant C independent of h,k, and p.

This is just the same statement as for $p = 1$, see Lemma 4.12. The proof can be found in [73].

We proceed to the assessment of dual discrete stability. Here we consider only data that vanishes at all points of the finite element grid. We showed in Lemma 4.17 that there exist approximating functions such that the error has this property. Hence, for integer $l \ge 0$, we define a subspace $F_0^l(\Omega) \subset L^2(\Omega)$ by

$$F_0^l(\Omega) = \left\{ f \in L^2(\Omega) \;\middle|\; f^{(-i)}|_{X_h} = 0 \quad \text{for} \quad i = 1, \dots, l \right\} \tag{4.7.36}$$

with $F_0^0(\Omega) := L^2(\Omega)$. For this data, one can locally bound the components of the condensed right-hand side in (4.7.24).

Lemma 4.24. *Consider the variational problem (4.2.2) on $S_h^p(\Omega)$ with data $f \in F_0^{p-1}(\Omega)$. Let τ_j be an arbitrary finite element and let $\{\tilde{r}_{j-1}, \tilde{r}_j\}^T$ be the condensed right-hand side vector given by (4.7.24). Assume further that the mesh size h is sufficiently small so that $hk \le \alpha < \pi$. Then*

$$|\tilde{r}_j| \le C_{\mathrm{d}}(p,m) h^{1/2} k^m \|f^{(-m)}\|_{\Delta_i} \tag{4.7.37}$$

holds for even $m = 0, 2, \dots \le p-1$ with

$$C_{\mathrm{d}}(p,0) = 1$$

and

$$C_{\mathrm{d}}(p,m) = C_1 + C_2 \alpha^{p-m}\, 2^{(m-1)/2} \frac{(p+1)!(p+1)}{((p-m+1)!)^2}, \;\; m \ge 2\,,$$

where C_1, C_2 do not depend on h, k, or p.

The proof is rather technical; the details can be found in [73]. We can prove this statement also for odd m with the additional assumption that hk is bounded from below, i.e., $0 < \beta \le hk$.

We are ready to formulate the dual stability theorem on the finite-dimensional (discrete) level.

Lemma 4.25. *Let $u_h \in S_h^p(0,1)$ be the FE solution to (4.2.2) with datum $f \in F_0^m(\Omega)$. Assume that $0 < kh \le \alpha < \pi$. Then*

$$|u_h|_1 \le C_{\mathrm{d}}(p,m)k^m\|f^{(-m)}\| + C_1\|f^{(-1)}\|\,, \tag{4.7.38}$$

where C_{d} is the constant from (4.7.37) and C_1 does not depend on $h, k,$ or p.

The proof follows the argument for the continuous case; see Lemma 4.19. For details, we refer again to [73].

4.7.6 Error Estimates

We formulate two theorems. The first error estimate is similar to the *h*-version estimate (4.5.14).

Theorem 4.26. *For $1 \le l \le p$ let $u \in H^1_{(0}(0,1) \cap H^{l+1}(0,1)$ and $u_h \in S_h^p(0,1)$ be the solution and the FE solution to the variational problem (4.2.2), respectively. Assume that $hk \le \alpha < \pi$. Then the error $e := u - u_h$ satisfies the bound*

$$|e|_1 \le C_{\mathrm{a}}(l)\left(1 + C_1 k\left(\frac{kh}{2p}\right)\right)\left(\frac{h}{2p}\right)^l |u|_{l+1}\,, \tag{4.7.39}$$

where C_1 does not depend on $h, k,$ or p, and $C_{\mathrm{a}}(l)$ is the constant in the approximation property (4.7.4).

The proof proceeds as for $p = 1$, using the interpolant s from Lemma 4.17. We write $e = u - s + z$ and observe that $z := s - u_h$ solves (4.2.2) with data $f = k^2(u - s)$. Thus $|e|_1 \le Ck^2\|u - s\| + |u - s|_1$, and the statement follows with the approximation property (4.7.4). This proves (4.7.39).

But (4.7.4) admits error bounds also with respect to dual norms of the data $k^2(u-s) \in F_0^{p-1}(\Omega)$. Using also the corresponding dual stability properties from Lemma 4.25, we directly arrive at the following result.

Theorem 4.27. *Let $1 \le l \le p$ and $0 \le m \le p$, m even, with $p \ge 2$. Let $u \in H^1_{(0}(0,1) \cap H^{(l+1)}(0,1)$ be the solution to the variational problem (4.2.2) with data $f \in H^{(l-1)}(0,1)$, and let $u_h \in S_h^p(0,1)$ be the finite element solution to this problem. Assume further that the mesh size h is such that $hk \le \alpha < \pi$. Then*

$$|e|_1 \le C_{\mathrm{a}}(l)\left[1 + C_1\left(\frac{kh}{2p}\right)^2 + kC_{\mathrm{d}}(p,m)C_{\mathrm{a}}(m)\left(\frac{kh}{2p}\right)^{m+1}\right]\left(\frac{h}{2p}\right)^l |u|_{l+1} \tag{4.7.40}$$

holds with C_1 independent on $k, h,$ and p.

The order m is related to the specific approximation properties of the interpolant s, allowing the estimation of $p+1$ antiderivatives in the dual norm. As usual in the *hp*-version, the error is bounded by higher-order derivatives of the exact solution.

Assuming again oscillatory behavior with $|u|_{l+1}/|u|_1 = O(k^l)$ and taking $l = p$, we can estimate the relative error after (4.7.40) as

$$\tilde{e}_1 \leq C_1 \left(\frac{hk}{2p}\right)^p + C_2\, k \left(\frac{hk}{2p}\right)^{2p} . \tag{4.7.41}$$

The first term in this estimate is the approximation error, while the second term represents numerical pollution. As for the *h*-version, this term is of the same order as the phase lag. We conclude that the pollution effect for $p \geq 2$ is significantly reduced if the mesh is fine enough such that

$$\theta = \frac{kh}{2p} < 1 .$$

Since this expression is taken to the power $2p$ in the pollution term, the wave number k must be rather large to render a significant pollution effect.

Remark 4.28. The order of convergence depends on the regularity of the solution. A function over a one-dimensional domain $\Omega \subset \mathbf{R}$ has regularity $l+1$ if its weak derivatives $\partial^j u$, $j \leq l+1$ exist in Ω. For solutions u of exterior Helmholtz problems, the Wilcox expansion theorem states that u is analytic, i.e., of infinite regularity. In that case, the order of convergence is always p after estimate (4.7.41). However, there are applications such as diffraction from binary gratings (cf. Elschner and Schmidt [51]), where one maximally has regularity 2. Solving such problems with *hp* FEM, one can expect only the convergence rate 1 in the energy norm. On the other hand, since the pollution error is estimated using negative norms, the estimate for $l \leq p$ is still

$$\tilde{e}_1 \leq C_1 \left(\frac{hk}{2p}\right)^l + c_2\, k \left(\frac{hk}{2p}\right)^{l+p} .$$

Hence, the application of *p*-version elements always leads to a reduction of the pollution error. Relative to the approximation error, this reduction is even more significant for low regularity.

Remark 4.29. Setting

$$\theta = \left(\frac{hk}{p}\right)^p ,$$

we can rewrite (4.7.41) as

$$\tilde{e}_1 \leq C_1\theta + C_2\theta^2 .$$

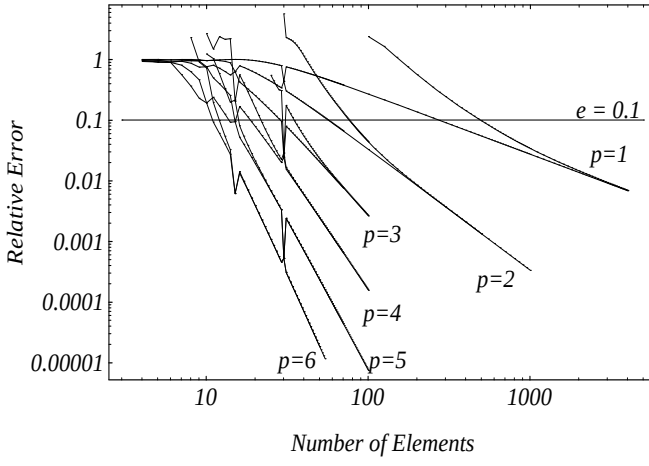


FIGURE 4.22. Relative error of the finite element solution versus best approximation error for $k = 30\pi$ and $p = 1, 2, \ldots, 6$.

Thus the estimate is similar for all p, including the standard h-version with $p = 1$; cf. (4.5.15). As a practical matter, we conclude that the error will, in general, be the same for any p if θ is kept constant (i.e., if higher approximation is applied on coarser mesh).

4.7.7 Numerical Results

We illustrate our analysis with some results from computational experiments. The following computations were carried out on Model Problem I with data $f = 1$ in (4.2.2). In Fig. 4.22, we plot the error of the best approximation and the FE error in the H^1-norm for $k = 30\pi$ and $p = 1, \ldots, 6$. We observe the theoretically predicted asymptotic convergence rates h^p and the reduction of the pollution effect for $hk < 2p$. We also see that roughly the same pollution (look at the distance between the BA and the FE error curves along the horizontal mark $e = 0.1$) occurs for constant ratio θ as defined in the remark above.

Remark 4.30. The "bumps" in the error lines in Fig. 4.22 are an artifact of the one-dimensional model. For h such that $kh = l\pi, l = 1, 2, \ldots$, we locally approximate only one of the functions $\sin x$ or $\cos x$, respectively. From the Taylor expansions of these functions it is easy to understand the "even–odd" effect in the approximation with polynomials of order p.

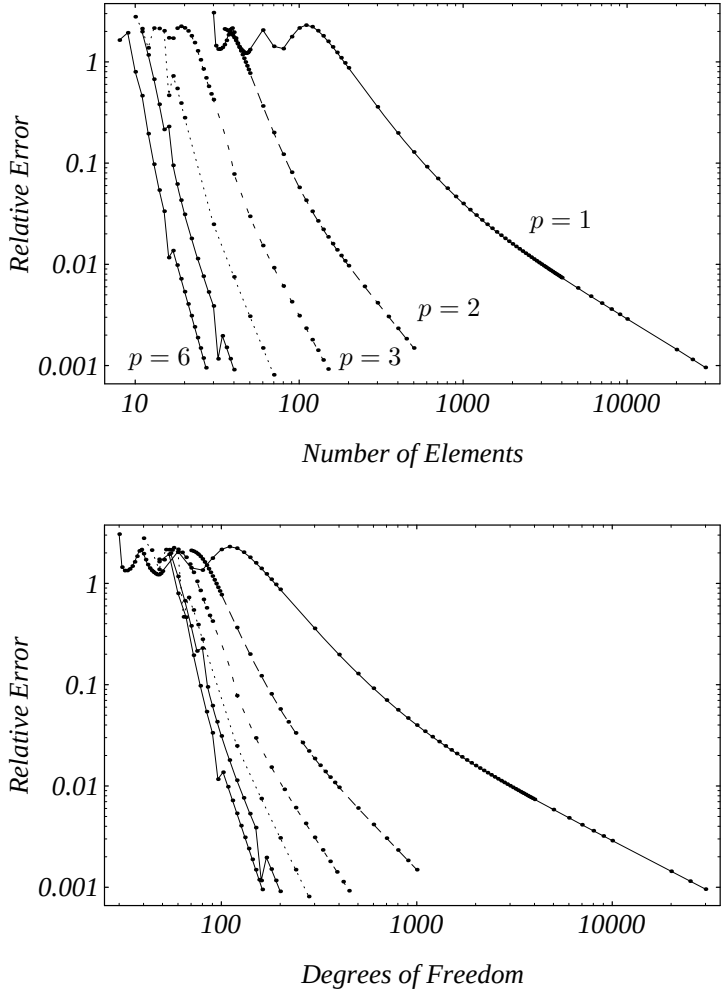


FIGURE 4.23. Error of the finite element solution as a function of the number of nodal points N (upper plot) and as a function of the number of DOF (lower plot) for $k = 100$ and $p = 1, \ldots, 6$.

TABLE 4.3. Number of elements to achieve accuracy (relative error in H^1-seminorm) of ε. Parameters: $k = 30\pi$, $p = 1, \ldots, 6$; n: number of elements; DOF=$n * p$: degrees of freedom; nmd: computational cost measured in numbers of multiplications and divisions.

p	1	2	3	4	5	6	ε
$\tilde{e}_{fe}$	0.49795	0.51962	0.5470	0.5582	0.5851	0.7208	
n	211	48	25	16	12	10	
DOF	211	96	75	64	60	60	0.5
nmd	1051	284	321	508	824	1296	
$\tilde{e}_{fe}$	0.099947	0.09956	0.09683	0.0911	0.0829	0.09833	
n	491	76	35	22	16	12	
DOF	491	152	105	88	80	72	0.1
nmd	2451	452	451	700	1100	1556	
$\tilde{e}_{fe}$	0.01000	0.0.01055	0.01013	0.01000	0.01026	0.00983	
n	2813	180	64	35	23	17	
DOF	2813	360	192	140	115	102	0.01
nmd	14061	1076	828	1116	1583	2206	

The theoretical analysis and the numerical tests show that the quality and reliablity of the FE solution grow with p. In computational practice, this (theoretically unbounded) growth is inhibited by the decrease in numerical stability due to ill conditioning of the stiffness matrix for higher p. We thus conclude this section by a short discussion of computational efficiency of the FEM implementation for different p. Consider first the plots in Fig. 4.23, where the relative error of the FE solution is plotted, respectively, against the number of mesh points and the number of DOF.

Increasing p, one significantly reduces the number of elements needed to stay within some preset tolerance. The same is true if accuracy is related to the number of DOF, though the gain for $p \geq 4$ becomes less significant. A comparison of numerical effort is made by the count of the number of multiplications and divisions (nmd). In the given one-dimensional case, condensation involves computing the inverse of a $(p-1) \times (p-1)$ matrix, which requires $(p-1)^3$ operations [5, p. 515]. Generally (on a non-uniform mesh) this has to be performed on each element. The solution of the resulting tridiagonal system then requires $5n - 4$ operations [5, p. 528]. The overall number of multiplications and divisions is thus

$$\text{nmd} = 5n - 4 + n(p-1)^3.$$

In Table 4.3, we tabulate the nmd needed to achieve a relative error of the finite element solution in H^1-seminorm of 0.1%, 0.5%, or 0.01%, respectively. We observe a significant payoff in computational effort in passing from $p = 1$ to $p = 2$ or $p = 3$. As usual in the *hp*-method, the optimal relation between h and p depends on the required accuracy; generally, the higher the accuracy the bigger the payoff by higher-order elements.

4.8 Generalized FEM for Helmholtz Problems

The pollution effect at high wave numbers is inherent in the standard Galerkin FEM solutions for Helmholtz problems. As previously, we speak of a nondimensional wave number that models a high frequency or a large computational domain. As both cases frequently occur in practical applications, the question arises of how to reduce the pollution effect.

Our analysis has shown that the pollution in the FEM error is due to the deterioration of stability. This is a specific property of the Helmholtz variational form. Accordingly, one may attempt a reduction of pollution by a modification of the Helmholtz operator in such a way that the modified operator has better stability properties. This approach is called stabilization. On the other hand, the practical interest is not stabilization per se but error control, i.e., convergence. Recalling the conclusions from Ceá's lemma, convergence can also be improved by raising the order of approximation. In the *hp*-version of the FEM, higher-order polynomials are used as shape functions. The next step is to use analytic functions for approximation. These functions should incorporate knowledge on the operator (e.g., be partial solutions) to have a priori good approximation properties. Methodically, such an approach can be viewed as a limiting case $p \to \infty$ of the *hp*-version of the FEM.

In this section, we review stabilized methods and methods with specific approximation properties. The approaches are investigated analytically and in numerical experiments on Model Problems I, II. The idea of stabilization by Galerkin–least-squares FEM after Harari and Hughes [63] is explained on the one-dimensional example. We also consider a so-called quasi-stabilized method that gives in one dimension the same nodal values as a FEM with analytic shape functions. All of these methods eliminate the pollution effect for Model Problem I.

This elimination is not possible using any generalized FEM in two dimensions. Here, the stabilizing effect of the GLS–FEM is sensitive to the direction of the exact solution. The error can be significantly reduced in certain "preferred" directions, but the stabilized approach, in general, has little effect if the exact solution does not have dominant components in one of these directions. As an alternative, we review the quasi-stabilized FEM (QSFEM) after Babuška and Sauter [11, 17]. In this method, the error is nowhere completely eliminated. Rather, the pollution error is minimized for all possible directional components of the exact solution.

4.8.1 Generalized FEM in One Dimension

To present the idea of the generalized FEM, we first consider Model Problem I with uniform mesh X_h.

Stabilization:

The Helmholtz variational form is indefinite due to the term $-k^2(u,v)$. In the Galerkin–least-squares (GLS) method proposed by Harari and Hughes [63], the variational form is modified to

$$b_{\mathrm{GLS}}(u,v) = b(u,v) + \tau\,(\mathcal{L}u, \mathcal{L}v)_{\tilde{\Omega}}, \qquad u, v \in V_h\,, \tag{4.8.1}$$

and the right-hand side is

$$f_{\mathrm{GLS}}(v) = f(v) + \tau\,(f, \mathcal{L}v)_{\tilde{\Omega}}. \tag{4.8.2}$$

Here, $\mathcal{L}$ is the Helmholtz differential operator, τ is a parameter yet to be determined, and $(\cdot,\cdot)_{\tilde{\Omega}}$ is the reduced L^2 inner product, where integration is carried out only on the element interiors (i.e., the singularities at inter-element boundaries are suppressed in the reduced inner product).

The goal is to make, by appropriate choice of the parameter τ, the form b_{GLS} "unconditionally stable" and thus to "circumvent the Babuška–Brezzi condition," i.e., to avoid the stability problems of the form b as quantified by the inf–sup condition.

The optimal τ is found from discrete dispersion analysis. A typical stencil of the standard FEM stiffness matrix in one dimension is (cf. (4.5.1))

$$(2\alpha_{\mathrm{G}} + 1)u_{j-1} + 2(4\alpha_{\mathrm{G}} - 1)u_j + (2\alpha_{\mathrm{G}} + 1)u_{j+1}\,, \tag{4.8.3}$$

with $\alpha_{\mathrm{G}} = (kh)^2/12$. The GLS matrix is similarly defined, but with α_{G} replaced by

$$\alpha_{\mathrm{GLS}} = \alpha_{\mathrm{G}}(1 - \tau k^2)\,. \tag{4.8.4}$$

The discrete wave number of the FE solution is determined by (4.5.4), which we here write in the form

$$\cos \tilde{k}h = \frac{1 - 4\alpha_{\mathrm{G}}}{1 + 2\alpha_{\mathrm{G}}}.$$

Thus we seek τ such that

$$\cos kh = \frac{1 - 4\alpha_{\mathrm{GLS}}}{1 + 2\alpha_{\mathrm{GLS}}},$$

i.e., the GLS solution has the exact wave number. Solving this equation for α_{GLS} and equating the result to $\alpha_{\mathrm{G}}(1 - \tau k^2)$, we find

$$\tau = \frac{1}{k^2}\left(1 - \frac{6}{k^2h^2}\,\frac{1 - \cos kh}{2 + \cos kh}\right)\,. \tag{4.8.5}$$

It is shown in numerical experiments on Model Problem I that this choice of τ leads indeed to solutions with no phase lag [63]. Thus the error in the GLS method is an interpolation error that is free of pollution.

A generalized FEM with the same property on nonuniform meshes is developed in Babuška and Sauter [17]. In this quasi-stabilized FEM (QSFEM), the nodal values of the approximate solution u_h are computed from the algebraic system

$$\mathbf{G}^{\text{stab}}\mathbf{u}_h = Q^{\text{stab}}(f)\,, \tag{4.8.6}$$

where $\mathbf{u}_h$ is the vector of nodal values of the function u_h on X_h, the FE stiffness matrix $\mathbf{G}^{\text{stab}}$ is the tridiagonal matrix defined by

$$G_{ij} = \frac{k^2 h}{2\tan\frac{kh}{2}} \begin{cases} \dfrac{\sin k(x_{i+1}-x_{i-1})}{\sin k(x_{i+1}-x_i)\sin k(x_i-x_{i-1})} & \text{if} \quad i=j\,, \\ -\dfrac{1}{\sin k|x_i-x_j|} & \text{if} \quad |i-j|=1\,, \\ 0 & \text{otherwise}\,, \end{cases} \tag{4.8.7}$$

and the mapping Q^{stab} is defined by

$$\left(Q^{\text{stab}}f\right)_i = \frac{h}{2\tan\frac{kh}{2}} \sum_{m=i}^{i+1} \frac{\tan\frac{k(x_m-x_{m-1})}{2}}{x_m-x_{m-1}} \frac{\int_{x_{m-1}}^{x_m} f(x)dx}{x_m-x_{m-1}}\,. \tag{4.8.8}$$

In [17], it is proven that the solution u_h obtained from (4.8.6) is nodally exact for piecewise constant data and that it is pollution-free for any data $f \in H^1(\Omega)$.

Approximation with Analytic Shape Functions:

The nodal values of the QSFEM are equivalently obtained from a Galerkin FEM with analytic shape functions. Consider the nodal functions Φ_i of the form

$$\Phi_i = \begin{cases} t_1^{(i)} & \text{on} \quad \Delta_i\,, \\ t_2^{(i-1)} & \text{on} \quad \Delta_{i-1}\,, \\ 0 & \text{otherwise,} \end{cases} \tag{4.8.9}$$

where the shape functions t_1, t_2 are computed from the local boundary value problems (4.7.31) and (4.7.32) or (4.7.33), respectively. On the master element $\xi \in (-1,1)$, the shape functions are explicitly

$$t_1(\xi) = -\frac{\sin K\xi}{2\sin K} + \frac{\cos K\xi}{2\cos K}\,, \tag{4.8.10}$$

$$t_2(\xi) = \frac{\sin K\xi}{2\sin K} + \frac{\cos K\xi}{2\cos K}\,, \tag{4.8.11}$$

with $K = kh/2$; see Fig. 4.24.

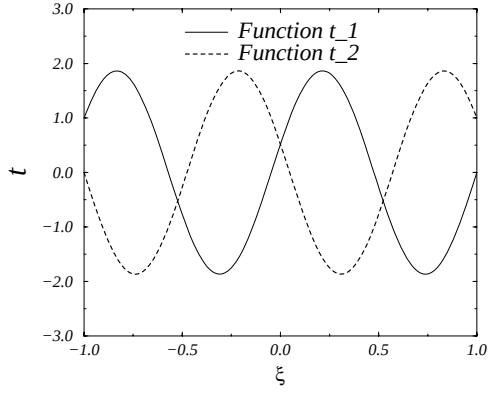


FIGURE 4.24. Analytic shape functions on master element.

The trial functions for the Galerkin FEM are written in the standard way as a linear combination of nodal functions

$$u_h(x) = \sum_{i=1}^{N} u_i \Phi_i(x)\,, \tag{4.8.12}$$

where u_i are the unknown nodal values of the function u_h. Then the matrix coefficient G_{ii-1} is

$$\begin{aligned} G_{ii-1} &= \frac{2}{x_i - x_{i-1}} \int_{-1}^{1} t_1'(\xi) t_2'(\xi) d\xi - \frac{x_i - x_{i-1}}{2}\, k^2 \int_{-1}^{1} t_1(\xi) t_2(\xi) d\xi \\ &= -\frac{k}{\sin k(x_i - x_{i-1})}\,. \end{aligned}$$

Similarly,

$$\begin{aligned} G_{ii} &= \frac{2}{x_i - x_{i-1}} \int_{-1}^{1} (t_2')^2 - \frac{x_i - x_{i-1}}{2}\, k^2 \int_{-1}^{1} (t_2)^2 \\ &\quad + \frac{2}{x_{i+1} - x_i} \int_{-1}^{1} (t_1')^2 - \frac{x_{i+1} - x_i}{2}\, k^2 \int_{-1}^{1} (t_1)^2 \\ &= \frac{k \,\sin k(x_{i+1} - x_{i-1})}{\sin k(x_i - x_{i-1}) \sin k(x_{i+1} - x_i)}\,. \end{aligned}$$

The expression on the right-hand side is exact for piecewise constant functions. In this case, we integrate:

$$\begin{aligned} &\frac{2}{x_i - x_{i-1}} \int_{-1}^{1} f t_1 + \frac{2}{x_{i+1} - x_i} \int_{-1}^{1} f t_2 \\ &= \frac{2}{k} \left(\frac{\tan k(x_i - x_{i-1})}{x_i - x_{i-1}} \frac{\int_{x_{i-1}}^{x_i} f}{x_i - x_{i-1}} + \frac{\tan k(x_{i+1} - x_i)}{x_{i+1} - x_i} \frac{\int_{x_i}^{x_{i+1}} f}{x_{i+1} - x_i} \right). \end{aligned}$$

Equations (4.8.7), (4.8.8) are then obtained if we multiply by a scaling factor. For a general right-hand side, the integration over the elements is equivalent to local averaging of the data. Since (4.8.12) represents in fact the *exact* solution of (4.2.1), it is clear that the generalized FEM (which uses linear shape functions but has nodal values from analytic shape functions) is exact at the nodal points. Since the homogeneous generalized FEM has no phase lag, it can be shown [17] that no pollution occurs in the error for all right-hand sides $f \in H^1(\Omega)$.

Conclusions from the 1-D example:

In the one-dimensional case, it is possible to eliminate the phase lag of the FE solution and, accordingly, the pollution term in the error. The salient feature of the Galerkin–least-squares FEM and the quasi-stabilized FEM is high nodal approximation and elimination of the phase lag rather than stabilization of the underlying variational form. For small values of kh, the GLS parameter τ is negative and thus does not stabilize the Helmholtz variational form. Indeed, expanding (4.8.5) for small kh leads to

$$\tau = -\frac{h^2}{12} + O(k^2h^4);$$

see also Fig. 4.25.

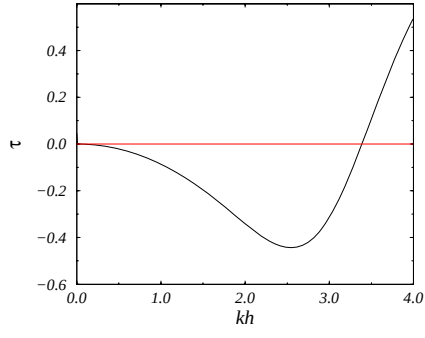


FIGURE 4.25. The function $\tau(kh)$ for $k = 100$.

4.8.2 Generalized FEM in Two Dimensions

Unlike the one-dimensional model case, Helmholtz problems in higher-dimensional applications possess an infinite number of linearly independent particular solutions. For instance, all plane waves $u(x, y) = \exp(i\mathbf{k}\cdot\mathbf{x})$ with $|\mathbf{k}| = k$ are solutions of $\triangle u + k^2u = 0$ in $\mathbf{R}^n$, $n \geq 2$. We now investigate how this essential difference influences the performance of the generalized FEM.

Galerkin–Least-Squares FEM:

We consider the numerical solution of Model Problem II with bilinear FEM on a uniform square mesh. The shape functions are the nodal modes

$$N_1(\xi,\eta) = \phi_1(\xi)\phi_1(\eta),\ \ldots,\ N_4(\xi,\eta) = \phi_2(\xi)\phi_2(\eta)\,.$$

Inserting the bilinear approximation into the GLS variational equality

$$b(u_h, v_h) + \tau(\mathcal{L}u_h, \mathcal{L}v_h) = 0$$

on a uniform interior element patch as shown in Fig. 4.26 leads to the discrete equation

$$\sum_{k=i-1}^{i+1}\sum_{l=j-1}^{j+1} \left(S_{ijkl} - k^2(1-\tau k^2)M_{ijkl}\right)\, u_h(x_k, y_l) = 0\,,$$

where $S_{ijkl},, M_{ijkl}$ are the coefficients of the assembled stiffness and mass matrix, respectively. Thus, at every interior point of the mesh, the FEM

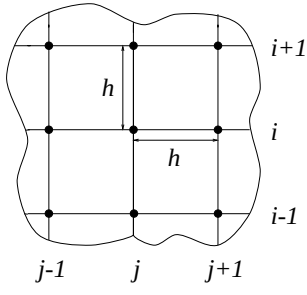


FIGURE 4.26. Patch in uniform 2-D mesh.

equations can be represented by the 9-point difference star

$$A^{\text{interior}} = \begin{bmatrix} A_2 & A_1 & A_2 \\ A_1 & A_0 & A_1 \\ A_2 & A_1 & A_2 \end{bmatrix}\,. \tag{4.8.13}$$

Assuming now a solution in the form of a plane wave with (known) direction θ and (unknown) discrete wave number $\tilde{k} = \tilde{k}(\theta)$,

$$u_h(x,y) = e^{i\tilde{k}(x\cos\theta + y\sin\theta)}\,,$$

we get the dispersion relation

$$A_0 + 2A_1(\cos\tilde{\xi}_1 + \cos\tilde{\xi}_2) + 4A_2(\cos\tilde{\xi}_1\cos\tilde{\xi}_2) = 0\,, \tag{4.8.14}$$

with

$$\tilde{\xi}_1 := \tilde{k}h\cos\theta\,, \qquad \tilde{\xi}_2 := \tilde{k}h\sin\theta\,.$$

For the Galerkin method with exact integration, we have the standard coefficients of the stiffness and mass matrix

$$A_0 = \frac{8}{3} - 4\alpha_{\mathrm{G}}\,, \quad A_1 = -\frac{1}{3} - \frac{\alpha_{\mathrm{G}}}{4}\,, \quad A_2 = -\frac{1}{3} - \alpha_{\mathrm{G}}\,,$$

where now $\alpha_{\mathrm{G}} := \dfrac{(kh)^2}{9}$. The same coefficients are obtained for the GLS–FEM, with α_{G} replaced by $\alpha_{\mathrm{GLS}} = \alpha_{\mathrm{G}}(1 - \tau k^2)$. Inserting this stencil into the dispersion relation (4.8.14), the optimal value of $\tau(\theta, k, h)$ is computed by setting $\tilde{k} = k$, i.e., replacing $\tilde{\xi}_1, \tilde{\xi}_2$ by

$$\xi_1 = kh\cos\theta\,, \quad \xi_2 = kh\sin\theta\,,$$

to obtain

$$\tau = \frac{1}{k^2}\left(1 - 6\frac{4 - \cos\xi_1 - \cos\xi_2 - 2\cos\xi_1\cos\xi_2}{(2 + \cos\xi_1)(2 + \cos\xi_2)k^2h^2}\right). \tag{4.8.15}$$

Inserting this value of τ into the GLS variational equality, we get an FE solution that has no phase lag if the exact solution is a plane wave in direction θ.

Example 4.31. Assuming $\theta = 0$ we get with $\cos\xi_1 = \cos kh, \cos\xi_2 = 1$, the 1-D dispersion relation (cf. (4.8.5))

$$\tau = \frac{1}{k^2}\left(1 - \frac{6}{k^2h^2}\,\frac{1 - \cos kh}{2 + \cos kh}\right).$$

However, a general signal consists of plane waves going in an infinite number of directions. Even if there are directionally prevalent components in this decomposition, they are not necessarily known a priori. It is not clear if the GLS leads to improved approximation of a wave that is not dominant in the preferred direction. In fact, numerical tests (cf. Thompson and Pinsky [114, Fig. 12]; see also the paragraph on numerical results below) show that the GLS–FEM has the same error as the standard Galerkin FEM if the direction of the exact solution is different from the direction chosen for the factor τ.

Minimization of Pollution in the Quasi-Stabilized FEM:

The discretization of the Helmholtz equation by the Galerkin or the Galerkin–least-squares FEM renders on a uniform square mesh the uniform difference stencils (4.8.13) with dispersion relation (4.8.14). On the other hand, the exact dispersion relation of a plane wave $u = \exp(\xi_1 x + \xi_2 y)\,, \xi_1 = kh\cos\theta\,, \xi_2 = kh\sin\theta\,$, is the circle

$$\xi_1^2 + \xi_2^2 - k^2h^2 = 0\,.$$

The phase lag, and thus the pollution of the discrete solution, is determined by the distance between the curves described by the exact and the discrete dispersion relations. The maximal distance of the curves is

$$d(k,h,k_h) = \max_{0\le\theta\le 2\pi} \left\| \begin{pmatrix} \tilde{\xi}_1(\theta) \\ \tilde{\xi}_2(\theta) \end{pmatrix} - \begin{pmatrix} \xi_1(\theta) \\ \xi_2(\theta) \end{pmatrix} \right\|$$

in vector norm, where $\tilde{\xi}_1(\theta), \tilde{\xi}_2(\theta)$ are the roots of the dispersion relation (4.8.14). Recalling the definitions of $\tilde{\xi}_i$ and ξ_i, we see that

$$d = \max_{\theta} |\tilde{k}(\theta) - k| h\,. \tag{4.8.16}$$

Since $\tilde{k}$ is computed from the dispersion relation (4.8.14) as a function of A_0, A_1, A_2, we may ask for the *minimal* distance d over all possible coefficients of the stencil (4.8.13). The coefficients are linearly dependent through the dispersion relation; hence we effectively have two degrees of freedom for a modification of the stencil. Assuming that the phase lag, as in the one-dimensional case, can be expanded as a series with *odd powers* of kh,

$$(\tilde{k} - k)h = c_1(kh)^3 + c_2(kh)^5 + c_3(kh)^7 + \cdots,$$

where the c_i depend on A_0, A_1, A_2 via $\tilde{k}$, one can choose A_0, A_1, A_2 such that $c_1 = c_2 = 0$, whence

$$\tilde{k}_{\text{opt}} - k = k^7 h^6 + O\left(k^9 h^8\right)\,. \tag{4.8.17}$$

Example 4.32. The stencil

$$A_{\text{opt}}^{\text{interior}} = \sum_{m=0}^{3} (kh)^m A_m^{\text{interior}}$$

has the optimal property (4.8.17) if

$$A_0^{\text{interior}} = \begin{bmatrix} -\frac{1}{5} & -\frac{4}{5} & -\frac{1}{5} \\ -\frac{4}{5} & 4 & -\frac{4}{5} \\ -\frac{1}{5} & -\frac{4}{5} & -\frac{1}{5} \end{bmatrix}, \qquad A_1^{\text{interior}} = -\frac{1}{250} \begin{bmatrix} 17 & 58 & 17 \\ 58 & 0 & 58 \\ 17 & 58 & 17 \end{bmatrix},$$

$$A_2^{\text{interior}} = \frac{-1}{50000} \begin{bmatrix} 801 & 2549 & 801 \\ 2549 & 0 & 2549 \\ 801 & 2549 & 801 \end{bmatrix},$$

$$A_3^{\text{interior}} = \frac{-1}{4.5 \times 10^7} \begin{bmatrix} 152626 & 473849 & 152626 \\ 473849 & 0 & 473849 \\ 152626 & 473849 & 152626 \end{bmatrix}.$$

A rigorous outline of (4.8.17) can be found in Babuška et al. [17, 11]. There it is also shown that for any 9-point difference stencil, one can find some

exact solution of the Helmholtz equation such that its approximation on the stencil is polluted. Hence, in two dimensions, there is no generalized FEM (GFEM) with piecewise linear shape functions that is pollution-free for all possible loads. On the other hand, it is possible to construct stencils with minimal pollution error. The phase lag of the discrete solutions from these stencils is of order k^7h^6.

Remark 4.33. In the GLS–FEM, the modified discrete operator stems from a modified ("stabilized") variational form. This form is subsequently discretized by the standard FEM approach, i.e., the modifications are introduced on the functional level. The modifications of the discrete operator in the QSFEM are imposed on the algebraic level directly into the stiffness matrix. Thus the QSFEM is rather a finite difference than a finite element method. On the other hand, the conclusions from the analysis apply to any generalized FEM that leads to nine-point difference stencils. The idea has recently been generalized to rectilinear stencils (Elschner and Schmidt [51]).

Numerical Results:

Let us first illustrate the dispersion relation (4.8.14). We plot this relation for $kh = 2.5$, with the coefficients computed by the Galerkin method, the stabilized method after Pinsky and Thompson, and the quasi-stabilized method. The results are shown in Fig. 4.27, all in comparison to the exact dispersion curve (3.3.11).[4] We see that the quasi-stabilized relation is virtually identical with the exact relation, whereas the distance $d(\theta)$ of the stabilized curve is zero for some θ but is large for others. The distance between the Galerkin curve and the exact curve is large everywhere.

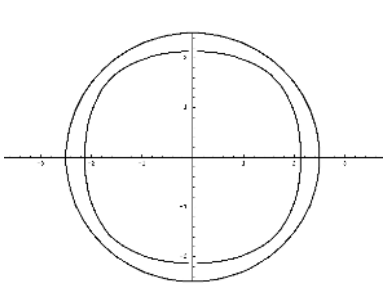
Second, we have solved Model Problem II with the methods discussed above. The boundary conditions of the computational example have been formulated in such a way that the exact solution is $u = \exp(i\mathbf{k} \cdot \mathbf{x})$.

Let us compare the standard Galerkin FEM with the stabilized methods. In Fig. 4.28, we show the difference of the FE and BA errors $|e|_1 - |e_{\mathrm{ba}}|_1$ for mesh sizes $h(k)$ such that $kh = 1.5$ (i.e., on coarse grids with roughly four elements per wavelength). We consider wave vectors $\mathbf{k} = k(\cos\theta, \sin\theta)$ with

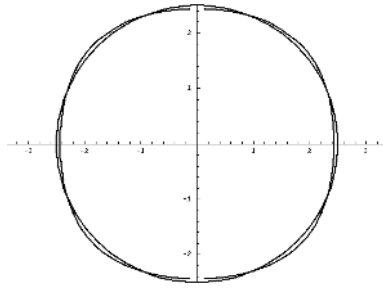
$$k = 30, 100, 150, \quad \theta = 0, \frac{\pi}{16}, \frac{\pi}{8}, \frac{3\pi}{16}, \frac{\pi}{4}, \frac{3\pi}{8}, \frac{\pi}{2}\,.$$

For each of these combinations (k, θ), the numerical solution is computed with Galerkin FEM, GLS–FEM, and the QSFEM. The preferred direction of the GLS–FEM is $\theta_0 = \frac{\pi}{8}$. Again, the directional sensitivity of the GLS-FEM is evident. While the numerical error is very small if the exact solution is dominant in the preferred direction, the error is not much reduced (with

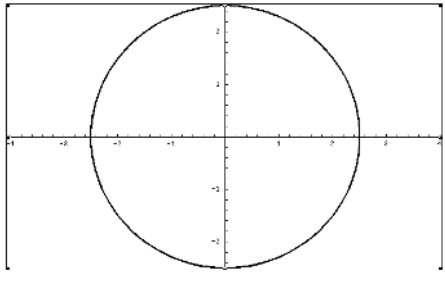
[4] Thanks to Stefan Sauter for the graphics routine that plots the 2-D dispersion curves.



(a) Galerkin



(b) Stabilized



(c) Q-stabilized

FIGURE 4.27. Dispersion curves for $k = 100$.

respect to the standard Galerkin FEM) if the exact solution differs from the preferred direction. On the other hand, the error reduction by QSFEM is directionally independent.

Analytic Approximation and Partition of Unity Method:

In one dimension, any solution of the homogeneous Helmholtz equation can be written as a linear combination of the functions $\exp(ikx), \exp(-ikx)$. Solutions of Helmholtz problems on two-dimensional bounded domains can be approximated with arbitrarily small error using plane waves. More precisely, the set

$$W = \left\{ \left\{ \exp(ik(x\cos\frac{2\pi m}{n} + y\sin\frac{2\pi m}{n})), m = 0, 1, \ldots, n-1 \right\}, n = 1, 2, \ldots \right\} \tag{4.8.18}$$

is dense in the H^1-norm on a bounded domain Ω (cf. Melenk [91, p.127], Herrera [68]).

Thus analytic shape functions can be constructed on a rectilinear mesh as tensor products of the one-dimensional functions $t_i(k_x x)t_j(k_y y), i, j = 1, 2$.

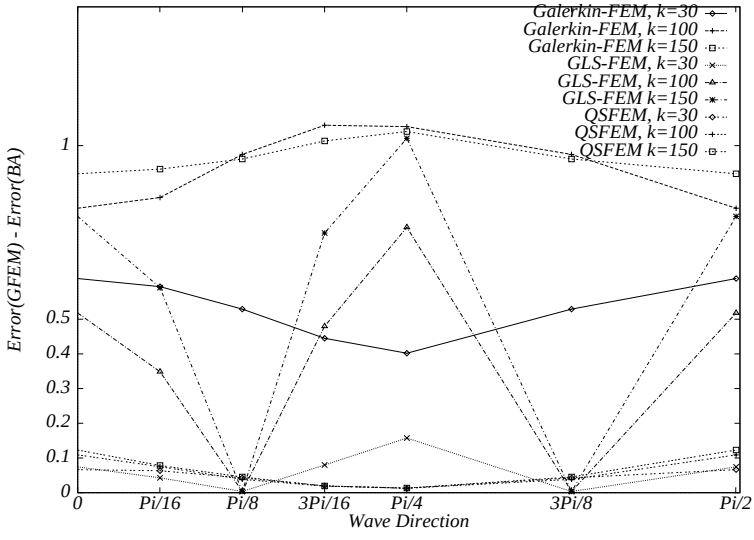


FIGURE 4.28. Dependency of the H^1-error on the wave direction θ for $kh = 1.5$ and $k = 30, 100, 150$.

Directional enrichment can be achieved by superposing several products with $k_x^2 + k_y^2 = k^2$. However, like the GLS–FEM and QSFEM, this approach lacks generality, since it is restricted to regular rectilinear grids. This restriction results from the fact that the analytic functions are used both for approximation and as shape functions on the FEM mesh, i.e., the functions have to satisfy the usual local boundary conditions on the elements.

This restriction is removed in a new generalized FEM, the partition of unity FEM (PUM) proposed by Melenk and Babuška [91, 92]. The conceptual idea of this new method makes it possible to incorporate analytical knowledge of the exact solution on a general mesh or even using a mesh-free approach of discretization. The key idea is to employ analytic functions for approximation without imposing any mesh-dependent boundary conditions directly on these functions. Rather, the shape functions are obtained by multiplying the approximating functions with patch functions that have compact support and globally form a partition of unity on the domain Ω. Thus a typical PUM basis function is written as

$$\Phi_i(x, y) = \varphi_i(x, y) v_i(x, y)\,, \tag{4.8.19}$$

where the $\{\varphi_i(x, y)\}_{i=1}^N$ form a partition of unity on Ω,

$$\sum_{i=1}^{N} \varphi_i(x, y) \equiv 1 \qquad \text{on } \Omega\,, \tag{4.8.20}$$

and

$$v_i(x,y) = \sum_{j=1}^{m(i)} v_i^{(j)}(x,y) \tag{4.8.21}$$

are the analytic approximating functions. For the Helmholtz equation, these functions can be chosen from (4.8.18).

An partition of unity is given, for example, by the piecewise linear basis functions ("hat functions") on any FE partition of Ω. In Fig. 4.29, a typical local approximation space of the PUM on a patch ω_j is depicted. The patch function φ_i is obtained from the piecewise linear shape functions. The approximating function v_i is a superposition of plane waves from W, being directed in angles $\theta_1, \ldots, \theta_4$. In a finite element context, one expects that

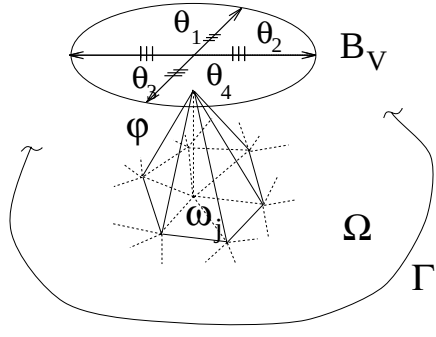


FIGURE 4.29. PUM basis function and approximation by plane waves.

by incorporation of functions from a local space W into the FEM trial space the quality of approximation is increased significantly. This is confirmed by Melenk and Babuška [92], who prove the approximation estimate

$$\|u_H - u_m\| \leq C(\Omega, k, s) \left(\frac{\ln^2 m}{m}\right)^s \|u_H\|_s.$$

Here, m is the number of basis functions from W that are used in (4.8.21). The approximated function u_H is any homogeneous solution of the Helmholtz equation on $\Omega \subset \mathbf{R}^2$. While this is the same convergence rate as for the p-version, the increase from m to $m+1$ adds only one DOF, whereas an ever increasing number of DOF is added if advancing by degree in the p-version with polynomial approximation. Also the algebraic rate of convergence is theoretically unlimited if the exact solution is analytic, hence one expects exponential convergence with respect to m. Indeed, the first numerical tests of the PUM for the Helmholtz equation show high convergence rates. The tests are performed on Model Problem II. Table 4.4 (taken from [91, p. 148]) shows the potential advantage in efficiency com-

pared to the FEM (h version) and QSFEM. In the table, N is the number of subdivisions of the sides of the square domain ($N = N_x = N_y$).

TABLE 4.4. DOF necessary in different methods to stay below the indicated tolerance for the relative error in H^1-norm, with wave number $k = 32$.

ϵ	15%	7%	1%
FEM	9400	26000^a	$800,000^a$
QSFEM	4096	16384	$80,000^a$
PUM, $N = 1$	88	88	104
PUM, $N = 8$	468	810	810

[a] estimated

4.9 The Influence of Damped Resonance in Fluid–Solid Interaction

In this section, we discuss some numerical aspects of Model Problem III, which describes fluid–solid interaction. The coupled problem of acoustic fluid–solid interaction is well-defined for all frequencies ω. Due to the radiation damping, the system does not have eigensolutions in the form of standing waves. The vibrations of the elastic structure are damped by the acoustic medium. The magnitude of the damping coefficient depends on the ratio of the fluid and solid material constants. If the fluid has small density, the system behaves almost as the elastic solid, vibrating in vacuo. Thus a loss of stability is expected near the interior (i.e., of the elastic obstacle) eigenfrequencies.

4.9.1 Analysis and Parameter Discussion

With respect to the wave number k, the coupled problem has generally the same stability properties as the uncoupled problem. For a load $\mathcal{F}$ that is a square-integrable function, there exists a unique solution $\mathcal{U}$ of the variational problem (4.2.21). The solution is twice weakly differentiable and depends on the load as (Babuška et al. [90])

$$\|\mathcal{U}_{,xx}\|_0 \leq C_{\mathrm{s}}(1+k)\|\mathcal{F}\|_0\,, \tag{4.9.1}$$

where $\|\cdot\|_0$ is the L^2-norm that is induced by the inner product (4.2.20).

Analyzing the convergence of the FE solution leads to the asymptotic error estimate

$$\|\mathcal{U}-\mathcal{U}_h\|_{1,V} \leq C_{\mathrm{opt}} \inf_{\Phi\in\mathcal{S}_h} \|\mathcal{U}-\Phi\|_{1,V}\,, \tag{4.9.2}$$

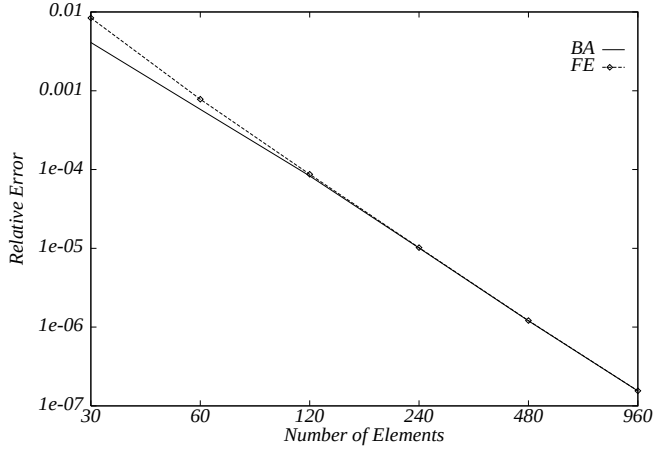


FIGURE 4.30. Relative error of best approximation (BA) and FE solution in H^1-norm as a function of the number of elements for $k = 10$ and $p = 2$. The FE error is marked by $\diamond$.

where $\mathcal{S}_h$ is the subspace of FEM approximation. Here, the H^1-norm is defined as

$$\|\mathcal{U}\|_{1,V} = \left((\mathcal{U}',\mathcal{U}')^2 + k^2(\mathcal{U},\mathcal{U})^2\right)^{1/2} . \tag{4.9.3}$$

The estimate holds with C_{opt} independent of k, h, provided that $k^2 h \ll 1$.

Thus both the exact and the discrete problems are well-posed. The analysis is based on the Gårding inequality, and hence does not provide preasymptotic estimates. The stability constants depend on the coupling parameters a and n, as defined in Section 4.2.3. These coupling parameters characterize the ratio of the solid and fluid material constants. For water and steel, typical values are

$$E = 210 \times 10^9 \text{N/m}^2, \qquad c_{\text{f}} = 1.5 \times 10^3 \text{m/s}, \qquad c_{\text{s}} = 5.1 \times 10^3 \text{m/s},$$
$$\rho_{\text{f}} = 1.0 \times 10^3 \text{kg/m}^3, \quad \rho_{\text{s}} = 7.8 \times 10^3 \text{kg/m}^3.$$

4.9.2 Numerical Evaluation

In our numerical experiments for the coupled problem we pursue two objectives: first, to investigate the preasymptotic error behavior of the coupled problem; second, to address the sensitivity of the coupled problem to the material properties of the media. Consider Model Problem III on a domain $\Omega = \cup\Omega_i$ with $\Omega_1 = (0,3)$, $\Omega_2 = (3,6)$, $\Omega_3 = (6,9)$. In the left fluid region, a source is present in the form of a step-function

$$g_1(x) = \begin{cases} 1\,, & x \in [1,2], \\ 0\,, & \text{otherwise}\,, \end{cases}$$

and no load is given in the solid or in the right fluid region: $f = g_2 = 0$. The material parameters for steel and water are assumed.

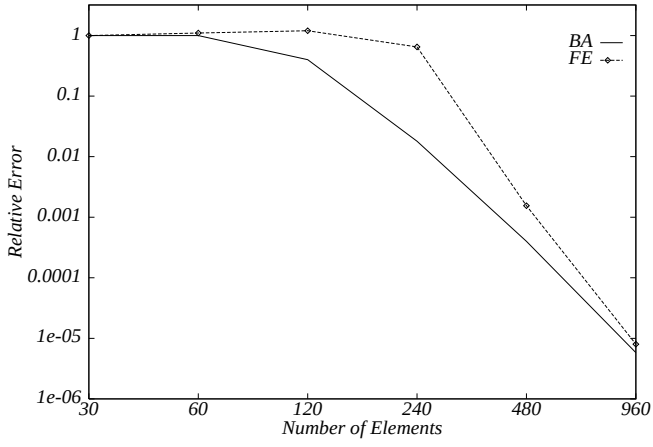


FIGURE 4.31. Relative error of best approximation (BA) and FE solution in H^1-norm as a function of the number of elements for $k = 500$ and $p = 5$. The FE error is marked by $\diamond$.

The nondimensional problem is discretized using an *hp*-version of the FEM on uniform meshes with $30, 60, 120, \ldots, 960$ elements in the whole fluid–solid region. The error of the FEM solution is again compared with the error of the best approximation.

Representative results for the FE and BA errors are shown in Fig. 4.31. Generally, higher p have to be chosen for higher k in order to get errors below 100%.

We see that the coupled problem generally shows the same numerical effects as the uncoupled one.

The more interesting feature of the coupled problem is its sensitivity to the ratio of the material properties fluid–solid. This is investigated in the following experiments. We now consider an infinite coupled domain $\Omega = \Omega_1 \times \Omega_2$, where the solid is given on $\Omega_1 = (0, 1)$ and the fluid on $\Omega_2 = (1, \infty)$. A Dirichlet boundary condition $u(0) = 0$ is assumed, and no load is given in the fluid, $g_2(x) = 0$. This is the problem of forced vibrations of a Dirichlet-fixed rod interacting with a fluid at $x = 1$. Via a DtN condition at $x = 1$, this problem can be reduced to the boundary value problem (cf. Example 1.6)

$$\begin{aligned} -u'' - \kappa k^2 u &= f \qquad \text{in } (0,1)\,, \\ u(0) &= 0\,, \\ u'(1) - ik\alpha u(1) &= 0\,, \end{aligned} \tag{4.9.4}$$

where $\kappa = (c_\mathrm{f}/c_\mathrm{s})^2$ and $\alpha = \rho_\mathrm{f} c_\mathrm{f}^2 / \rho_\mathrm{s} c_\mathrm{s}^2$. For $\alpha \to 0$ the problem is ill conditioned at the eigenvalues

$$\sqrt{\kappa}\, k = \left(\frac{\pi}{2} + m\pi\right), \quad m = 1, 2, \ldots\,.$$

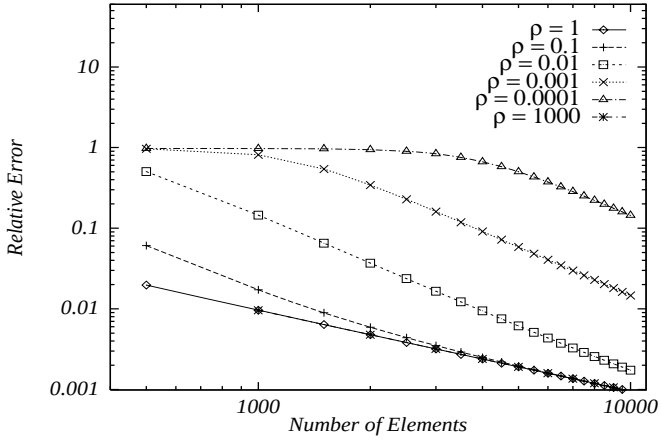


FIGURE 4.32. Relative error in the H^1-seminorm for $k\kappa^{1/2} = 10.5\pi$ and different values of ρ_{f}.

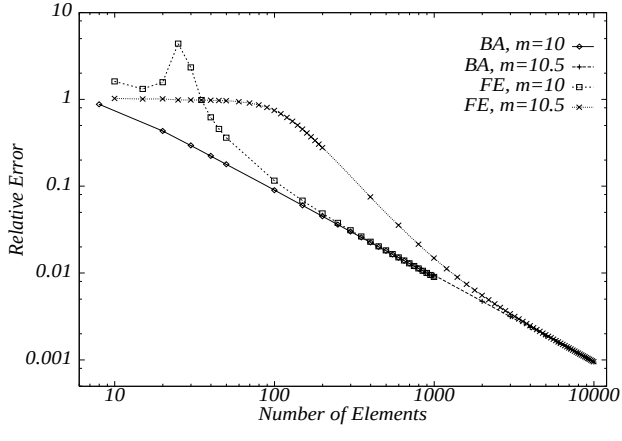


FIGURE 4.33. Relative errors for best approximation (BA) and FE solution, computed in the H^1-seminorm: comparison of $k\kappa^{1/2} = 10.5\pi$ (exact eigenvalue) to $k\kappa^{1/2} = 10\pi$ (between eigenvalues).

Figure 4.32 shows the error of the FE solution (with data $f = 1$, piecewise linear approximation) near the critical values for $\sqrt{\kappa}k$.

The relative error in the H^1-seminorm is plotted for different values of a (we formally let $a \to 0$ by $\rho_f \to 0$). The solution converges almost optimally for $\rho_f = 1000$. This corresponds to the propagation of a wave through a homogeneous medium as considered in Model Problem I. No significant pollution is observed on $[0, 1]$, since the wave number is small. However, the pollution effect grows considerably as $a \to 0$.

To simulate in the same plot the FEM convergence between eigenvalues, consider the line for $\rho = 10^3$. For large a^{-1}, the Robin boundary condition tends to the Dirichlet condition. This boundary value problem has exact eigenvalues at $m\pi,\ m = 1, 2, \ldots$, and hence the assumed value of $m = 10.5$ lies between two eigenvalues. This solution converges almost optimally.

This effect is also illustrated in Fig. 4.33. Again we compute with the material constants for steel and water. We see the significant increase of the pollution effect when the wave number is moved toward an interior eigenfrequency (from 10π to 10.5π). The mesh must be very fine in order to achieve a reliable resolution of the dynamic instability.

Resuming the experiments, we conclude that unlike the uncoupled problem, the FE solution of the coupled problem may be considerably polluted also for small k. The pollution occurs close to eigenvalues of the interior problem. In the general case, these eigenvalues are not known a priori. The reliability of the solution deteriorates close to the interior frequencies, in particular, if the damping by the acoustic medium is small.

4.10 A Posteriori Error Analysis

The goal of a posteriori error analysis is twofold (cf. Verfürth [116, p. 1]): first, one needs to identify (and resolve adaptively) *local concentrations* of the errors due to singularities. Second, the *a posteriori* error estimator should give a reliable indication of the *global quality* of the computed solution. Here, we focus on the second aspect. We analyze the residual estimator after Babuška and Miller [16]. In computational experiments, we also investigate the averaging method after Zienkiewicz and Zhu [122]. The question is how the indefiniteness of the variational form and the resulting pollution in the FE solution influence the efficiency of *a posteriori* error estimation. The investigation is carried out on Model Problem I.

4.10.1 Notation

Let $u_h \in V_h$ be the FE solution of (4.2.2). Here, we consider only the h-version of the FEM; hence $V_h =: S_h(0, 1) \subset H_{(0}(0, 1)$ denotes the space of piecewise linear functions that are defined on the unit interval and satisfy

a Dirichlet condition at $x = 0$. The function

$$r_i := \left(f + \frac{d^2 u_h}{dx^2} + k^2 u_h \right) \Big|_{\tau_i} \tag{4.10.1}$$

is called the interior residual in the element τ_i. We define the element *residual indicator function* $\hat{e}_i \in H_0^1(\tau_i)$ as the solution of the variational problem:

$$\left\{ \begin{array}{l} \text{Find } \hat{e}_i \in H_0^1(\tau_i) : \\ \qquad b_i(\hat{e}_i, \hat{v}) = (r_i, \hat{v})_i \qquad \forall\, \hat{v} \in H_0^1(\tau_i)\,, \end{array} \right. \tag{4.10.2}$$

where $H_0^1(\tau_i)$ denotes the subspace of H^1-functions that vanish on the element boundaries,

$$b_i(\hat{e}_i, \hat{v}) = \int_{\tau_i} \left(\frac{d\hat{e}_i}{dx} \frac{d\overline{\hat{v}}}{dx} - k^2 \hat{e}_i \overline{\hat{v}} \right) dx$$

is the reduction of the Helmholtz variational form to the element τ_i, and $(\cdot, \cdot)_i$ denotes the local L^2 inner product. The *element error indicators* are given by

$$\eta_i := |\hat{e}_i|_{1,i} = \left(\int_{\tau_i} \left| \frac{d\hat{e}_i}{dx} \right|^2 dx \right)^{1/2} . \tag{4.10.3}$$

Finally, we define the *global estimator* of the H^1-norm of the FE error $|e|_{1,\Omega}$ as

$$\mathcal{E} := \left(\sum_{i=1}^{N} \eta_i^2 \right)^{1/2} . \tag{4.10.4}$$

The efficiency of this estimator is measured by the *global effectivity index*

$$\kappa := \frac{\mathcal{E}}{|e|_{1,\Omega}} . \tag{4.10.5}$$

Remark 4.34. The error estimators based on the solution of local Dirichlet problems are equivalent (cf. Verfürth [116, Section 1.3]) to the original residual estimators of Babuška and Miller [16].

4.10.2 Bounds for the Effectivity Index

Theorem 4.35. *Assume $kh < \pi$. Then*

$$\frac{1}{\left(1 + \dfrac{k^2 h}{\pi}\right) \left(1 + \left(\dfrac{kh}{\pi}\right)^2\right)} \leq \kappa \leq \frac{1 + C(1+k)\dfrac{k^2 h^2}{\pi}}{\left(1 - \left(\dfrac{kh}{\pi}\right)^2\right)} \tag{4.10.6}$$

holds with

$$C = \frac{2}{\left(1 - 2(1+k)\dfrac{k^2h^2}{\pi^2}\right)\pi}\,, \tag{4.10.7}$$

provided that k, h are such that C is positive.

The proof of the theorem is based on the observation that the local indicators measure an interpolation error, while on the other hand, we know that the FE solution $u_h \in V_h$ is also shifted with respect to the exact solution. This shift is a global effect of dispersion (which again is related to the global effect of operator instability) and thus cannot be indicated locally.

To quantify these observations, we define a function $\tilde{u} \in H^1_{(0}(0,1)$ in such a way that the FE solution is the H^1-projection of $\tilde{u}$. Then there will be no phase difference between $\tilde{u}$ and u_h. Let the function $\tilde{u}$ be defined as the solution of the variational problem

$$\left\{ \begin{array}{l} \text{Find } \tilde{u} \in H^1_{(0}(0,1) : \\ \qquad (\tilde{u}', v') = (f, v) + k^2(u_h, v) + \langle u_h, v\rangle, \qquad \forall v \in H^1_{(0}(0,1)\,, \end{array} \right. \tag{4.10.8}$$

where we define

$$\langle u_h, v\rangle := iku_h(1)\bar{v}(1)\,.$$

We will call $\tilde{u}$ the shifted solution. By definition, the FE solution of Model Problem I is the projection of $\tilde{u}$ in the H^1-seminorm. Since we solve a one-dimensional problem, the projection u_h is in fact the nodal interpolant of $\tilde{u}$ (cf. Section 4.4.1). An illustration of the definition is given in Fig. 4.34, where we plot the exact, the shifted, and the finite element solutions.

Now we are ready to formulate two lemmas that lead directly to the proof of the theorem.

Lemma 4.36. *Let $u, \tilde{u}, u_h$ be the exact, the shifted, and the finite element solutions of (4.2.2), respectively. Let $e = u - u_h$, $\tilde{e} = \tilde{u} - u_h$. Then*

$$\left(1 + C(1+k)\frac{k^2h^2}{\pi}\right)^{-1} |\tilde{e}|_1 \;\le\; |e|_1 \;\le\; \left(1 + \frac{k^2h}{\pi}\right)|\tilde{e}|_1 \tag{4.10.9}$$

holds with C from (4.10.7), provided that h, k are such that C is positive.

For the proof we observe that the exact solution satisfies

$$(u', v') = (f, v) + k^2(u, v) + \langle u, v\rangle \qquad \forall\, v \in V\,.$$

Subtracting (4.10.8), we have

$$(\mu', v') = k^2(e, v) + \langle e, v\rangle \qquad \forall\, v \in V\,,$$

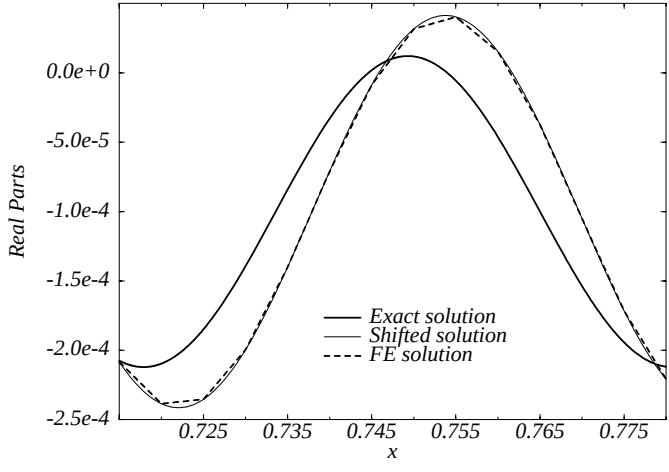


FIGURE 4.34. Real part of the exact, finite element, and shifted solutions for $k = 100$ on a mesh with resolution $\lambda/h \approx 13$.

where we define $\mu := u - \tilde{u} \in H^1_{(0}(0,1)$. Adding $-k^2((u-\tilde{u}),v) - \langle u - \tilde{u}, v\rangle$ to both sides, we find that μ is the solution of the problem

$$\left\{ \begin{array}{l} \text{Find } \mu \in H^1_{(0}(0,1) : \\ \qquad b(\mu, v) = k^2(\tilde{e}, v) \qquad \forall v \in H^1_{(0}(0,1)\,. \end{array} \right. \tag{4.10.10}$$

(Note that $\tilde{e}(1) = 0$ because u_h is the nodal interpolant of $\tilde{u}$.) Thus μ can be written as

$$\mu(x) = k^2 \int_0^1 G(x,s)\tilde{e}(s)\,ds\,,$$

where G is the Green's function given in (4.2.5). Differentiating in x and applying the Cauchy–Schwarz inequality, we see that

$$|\mu|_1 \le k^2 \|G_{,x}\|\|\tilde{e}\|\,.$$

Using the approximation property $(4.4.2)_3$, we have

$$|\mu|_1 \le \frac{k^2 h}{\pi}|\tilde{e}|_1\,.$$

Then, writing $e = u - u_h = \mu - \tilde{e}$, the upper bound of (4.10.9) follows by the triangle inequality.

For the lower bound, we write the variational equality of (4.10.8) as

$$(e' - \tilde{e}', v') = k^2(e,v) + \langle e, v\rangle\,.$$

Choosing $v = -\tilde{e}$, we get

$$|\tilde{e}|_1^2 = (e', \tilde{e}') - k^2(e, \tilde{e}) \leq |e|_1|\tilde{e}|_1 + k^2\|e\|\|\tilde{e}\|$$

by the Cauchy–Schwarz inequality. In the proof of the asymptotic error estimate (4.4.7) it is shown that

$$\|e\| \leq C(1+k)h|e|_1 \tag{4.10.11}$$

holds with C from (4.10.7). Hence

$$|\tilde{e}|_1^2 \leq |e|_1|\tilde{e}|_1 \ + \ k^2C(1+k)h|e|_1\frac{h}{\pi}\,|\tilde{e}|_1\,,$$

where we have also used the interpolation estimate (4.4.2). Dividing now both sides of the estimate by $|\tilde{e}|_1$, we obtain the lower bound, proving the lemma.

Remark 4.37. The assumptions on k, h in the lower bound are quite restrictive. In many cases, instead of (4.10.11), the simpler relation

$$|\tilde{e}|_1 \leq |e|_1 \tag{4.10.12}$$

holds, since $\tilde{e}$ is an interpolation error (of the shifted solution), whereas e is polluted. The lower bound (which is the numerator in the upper bound of the theorem) holds then with the constant 1. We observed this behavior, in particular, in the numerical evaluation. However, (4.10.12) is not generally true. A counterexample is given in Babuška et al. [12].

Lemma 4.38. *Assume $kh < \pi$. Then*

$$\left(1 + \frac{k^2h^2}{\pi^2}\right)^{-1} |\tilde{e}|_{1,i} \leq |\hat{e}_i|_{1,i} \leq \left(1 - \frac{k^2h^2}{\pi^2}\right)^{-1} |\tilde{e}|_{1,i}\,. \tag{4.10.13}$$

For the proof, let us fix an arbitrary interior element τ_i and use in (4.10.8) a test function $\hat{v} \in H_0^1(\tau_i)$ that is extended by zero outside of τ_i. We get

$$(\tilde{e}', \hat{v}')_i = (f + k^2u_h, \hat{v})_i = (r_i, \hat{v})\,,$$

where the latter equality follows from the definition of the interior residual (note that $u_h'' = 0$ in the element interiors). Now from the definition of the residual indicator function (4.10.2) we have

$$b_i(\hat{e}_i, \hat{v}) = (\tilde{e}', \hat{v}')_i\,. \tag{4.10.14}$$

Taking again $\hat{v} = \hat{e}_i$, we obtain

$$(\tilde{e}', \hat{e}_i')_i = |\hat{e}_i|_{1,i}^2 - k^2\|\hat{e}_i\|_i^2\,,$$

and hence

$$|\hat{e}_i|_{1,i}^2 = (\tilde{e}', \hat{e}_i')_i + k^2 \|\hat{e}_i\|_i^2 \leq |\tilde{e}|_{1,i} |\hat{e}_i|_{1,i} + \frac{k^2 h^2}{\pi^2} |\hat{e}_i|_{1,i}^2 ,$$

where we have used the interpolation estimate (4.4.2). Reordering and dividing by $|\hat{e}_i|_{1,i}$, we get

$$\left(1 - \frac{k^2 h^2}{\pi^2}\right) |\hat{e}_i|_{1,i} \leq |\tilde{e}|_{1,i} ,$$

and the upper bound of the statement readily follows.

To prove the lower bound, we take $\hat{v} = \tilde{e}$ in (4.10.14):

$$\begin{aligned} |\tilde{e}|_{1,i}^2 &= b_i(\hat{e}_i, \tilde{e}) \\ &= \left|(\hat{e}_i', \tilde{e}')_i - k^2 (\hat{e}_i, \tilde{e})_i\right| \leq |(\hat{e}_i', \tilde{e}')_i| + k^2 |(\hat{e}_i, \tilde{e})_i| \\ &\leq |\hat{e}_i|_{1,i} |\tilde{e}|_{1,i} + k^2 \|\hat{e}_i\|_i \|\tilde{e}\|_i \leq \left(1 + \frac{k^2 h^2}{\pi^2}\right) |\hat{e}_i|_{1,i} |\tilde{e}|_{1,i} . \end{aligned}$$

Canceling $|\tilde{e}|_{1,i}$ on both sides, we obtain the lower bound for $|\hat{e}_i|_{1,i}$, and the statement is shown.

Remark 4.39. The lemma indicates what the local indicators really measure, namely, the difference between the shifted solution and the FE solution (rather than the difference between the exact and the FE solutions). Indeed, we may assume that $kh \ll 1$ in practice, and hence the local effectivity index of the indicator with respect to the error function $\tilde{e}$ is close to 1 on each element. On the other hand, considering the upper bound in Lemma 4.37 we see that the global error of the FE solution may be significantly larger than the error with respect to the shifted solution ($k^2 h \gg 1$ for large k on meshes with $kh =$ constant). Using Lemma 4.38 we can identify the error of the shifted solution with the local indicator.

The estimate (4.10.6) of Theorem 4.35 is a direct consequence of the estimates (4.10.9) and (4.10.13). We state two conclusions. First, the error estimator is asymptotically exact; i.e., the effectivity index tends to 1 as $h \to 0$. Second, the lower bound in (4.10.6) indicates that the error may be significantly underestimated if $k^2 h$ is *not small*.

We now illustrate these conclusions in numerical experiments.

4.10.3 Numerical Results

In Fig. 4.35, we plot the error function e, the error with respect to the shifted solution $\tilde{e}$, and the local residual indicator function $\hat{e}$ for $f \equiv 1$, $k = 100$, and $h = 1/300$. *We see that the indicator $\hat{e}$ closely measures the*

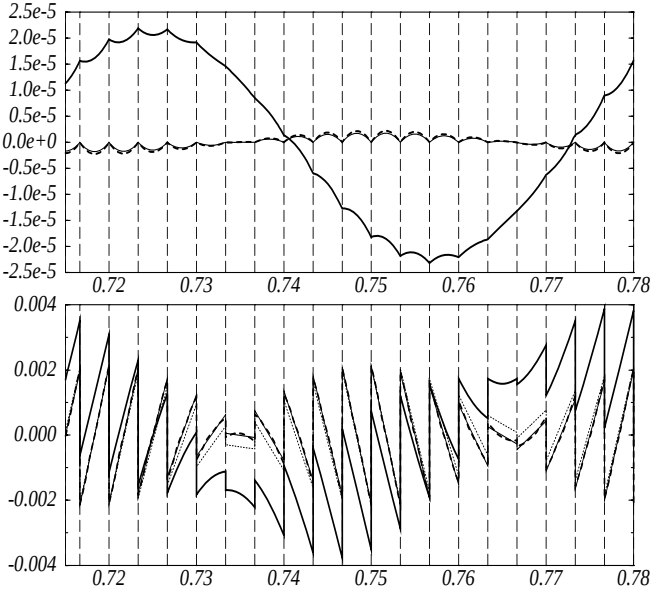


FIGURE 4.35. Real parts of exact error $e(x)$ (thick solid line), error with respect to the shifted solution $\tilde{e}(x)$ (thick dashed), and the residual error indicator function $\hat{e}(x)$ (thin solid). The derivates are shown in the lower plot, where also the (Z–Z)-estimator is plotted (dotted). Dashed vertical lines indicate nodes.

TABLE 4.5. Quality of the residual estimator: comparison of exact and estimated errors, magnitudes in percentage of H^1-norm of exact solution.

hk	k	Estimated, %	Exact, %
	10	16.73	19.58
	100	29.29	128.58
0.6	200	25.38	199.71
	500	17.43	133.85
	10	8.50	8.95
	100	10.10	28.88
0.3	200	11.27	66.40
	500	7.29	123.18
	10	2.79	2.81
	100	2.92	3.99
0.1	200	2.98	7.10
	500	2.93	18.09

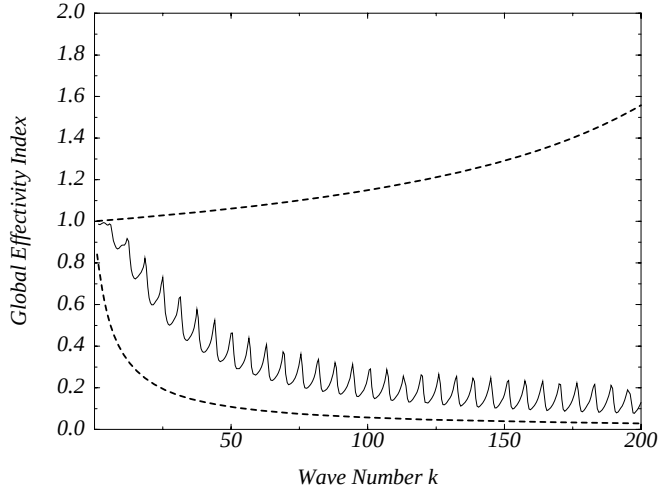


FIGURE 4.36. Numerical evaluation of bounds for the effectivity index. The solid line shows the measured index as a function of the wave number, the dashed lines show the predicted upper and lower bounds, respectively.

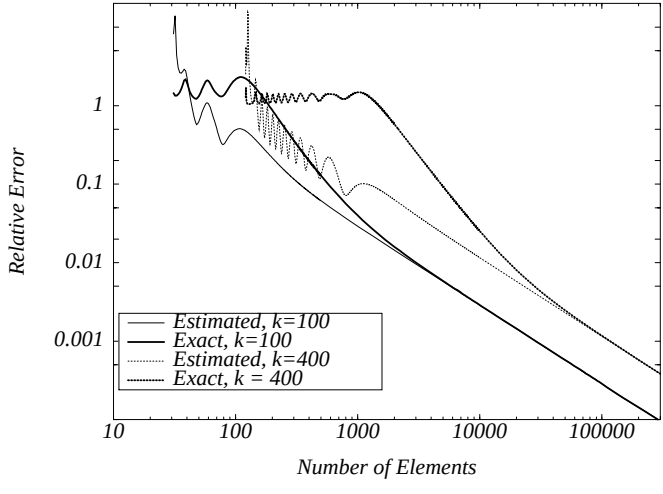


FIGURE 4.37. Convergence $h \to 0$ of exact and estimated error for $k = 100$ and $k = 400$.

error $\tilde{e}$ but does not reflect the true behavior of the FE error. The amplitude of the error is significantly underestimated by the local indicators.

The following computations further illustrate the bounds in (4.10.6). To show that the lower bound indeed reflects the behavior of the effectivity index for meshes with kh =constant, we plot in Fig. 4.36 the numerically computed effectivity index for $kh = 0.6$. This magnitude of kh corresponds to the "rule of thumb" to use 10 elements per wavelength. We compare the effectivity index of the FE computation (continuous line) to the theoretical upper and lower bound (dashed lines) in (4.10.6). Clearly, the lower bound of (4.10.6) reflects the actual behavior of the computed effectivity index.

We now evaluate the estimator $\mathcal{E}$ using linear elements. For this particular case, the second derivatives of u_h vanish in all element interiors, and hence the residuals r_i are given by $r_i = (f + k^2 u_h)|_{\tau_i}$. The element residual problems (4.10.2) were solved using quadratic approximations for trial and test functions $\hat{e}_i$ and $\hat{v}$.

In Fig. 4.37, we plot the exact relative error $\dfrac{|e|_{1,I}}{|u|_{1,I}}$ and the estimated relative error $\dfrac{\mathcal{E}}{|u|_{1,I}}$ for $k = 100$ and $k = 400$. Again we see that on coarse meshes the estimator $\mathcal{E}$ underestimates the error $|e|_{1,I}$. Note that for $k = 400$ the true error is 100% at a mesh size where the estimator predicts an error of only 10%. Additional numbers are given in Table 4.5 for meshes with $hk = 0.6, 0.3$, and 0.1. Recall that $hk = 0.6$ corresponds to the "rule of thumb" (4.4.1), whereas $hk = 0.3, hk = 0.1$ are overrefined meshes. Note that the estimators predict errors of a constant order of magnitude for all wave numbers. Comparing this observation with the *a priori* investigation (in particular, Fig. 4.9 and corresponding remarks), and also comparing Fig. 4.37 with Figures 4.11 and 4.12 from Section 4.5., we see that the estimator shows the typical convergence behavior of a best approximation — yet another illustration of the fact that the residual estimator measures the interpolation error.

To show that this behavior is typical for local error indicators in general, the computations were carried out also for the estimator $\mathcal{E}_{\mathrm{ZZ}}$ based on the so-called ZZ element error indicators, as defined in Zienkiewicz and Zhu [122] by

$$\eta_i^{\mathrm{ZZ}} := \left(\int_{\tau_i} \left| \hat{\sigma}_i^{\mathrm{ZZ}}(x) - \frac{du_h}{dx}(x) \right|^2 dx \right)^{\frac{1}{2}}, \tag{4.10.15}$$

where

$$\hat{\sigma}_i^{\mathrm{ZZ}}(x) = \frac{x - x_{i-1}}{x_i - x_{i-1}} \hat{s}_i^{\mathrm{ZZ}} + \frac{x_i - x}{x_i - x_{i-1}} \hat{s}_{i-1}^{\mathrm{ZZ}}, \qquad x \in \tau_i\,, \tag{4.10.16}$$

with

$$\hat{s}_i^{\mathrm{ZZ}} = \frac{1}{2} \frac{du_h}{dx}(\bar{x}_i) + \frac{1}{2} \frac{du_h}{dx}(\bar{x}_{i+1})\,, \qquad \bar{x}_i = \left(\frac{x_{i-1} + x_i}{2} \right),$$

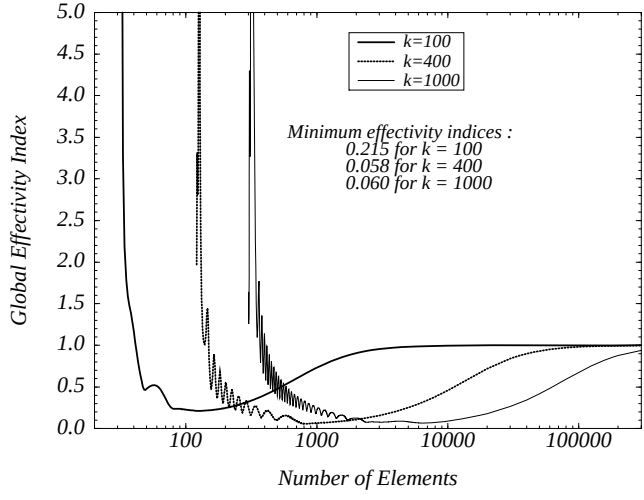


FIGURE 4.38. Convergence of the global effectivity index for the residual estimator.

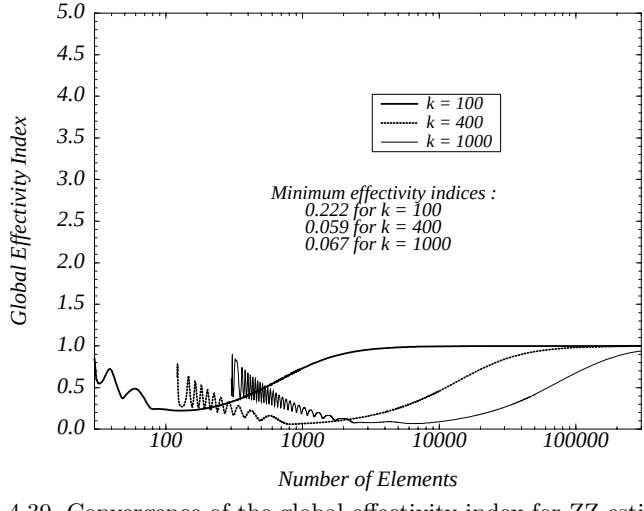


FIGURE 4.39. Convergence of the global effectivity index for ZZ estimator.

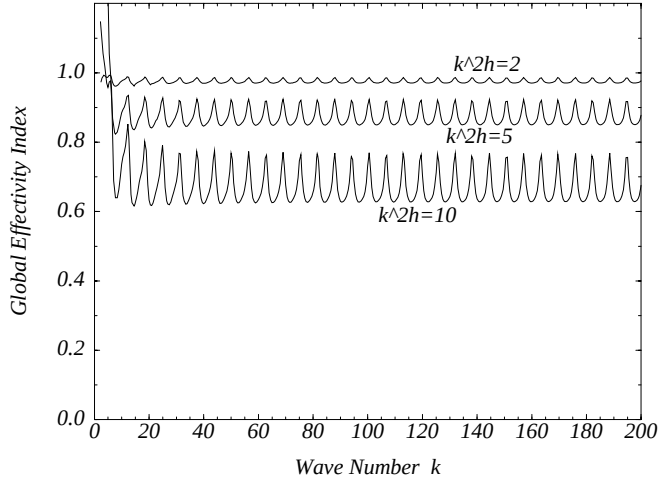


FIGURE 4.40. Global effectivity index on sequences of meshes with k^2h =constant. The effectivity indices have been computed from the residual estimator.

for each interior node i, $1 \leq i \leq N-1$. The estimator and effectivity indices are computed as in (4.10.4), (4.10.5).

In Figures 4.38 and 4.39, we plot the effectivity indices κ from the residual and ZZ methods, respectively, as functions of the number of elements for different k. We observe that in both cases the indices converge monotonically to 1 as $h \to 0$. Thus, for sufficiently fine meshes, $\mathcal{E}$ is an efficient estimator of $|e|_{1,\Omega}$. However, the asymptotic behavior of the estimator does not reflect the quality of engineering computations, where it is a common practice to employ meshes with a specified number of elements per wavelength. The plots show again how the true FE error is underestimated on meshes that correspond to the "rule of thumb" kh=constant.

The ZZ estimator and the residual estimator behave similarly, differing only on very coarse grids, where the residual estimator overestimates the error, while the ZZ estimator underestimates the error.

Finally, we expect from (4.10.6) that the global estimator $\mathcal{E}$ is efficient if the size of k^2h is restricted (as k grows), whereas it is unbounded if the meshes are designed by the "rule of thumb" (4.4.1). This is confirmed in Figures 4.40 and 4.41. The plots show how the effectivity index depends on k if the meshes for the FE computation satisfy constraints on k^2h and kh, respectively. Obviously, the estimators are efficient and reliable for all k if $hk^2 = \beta$. On the other hand, the quality of the estimators deteriorates if kh = constant.

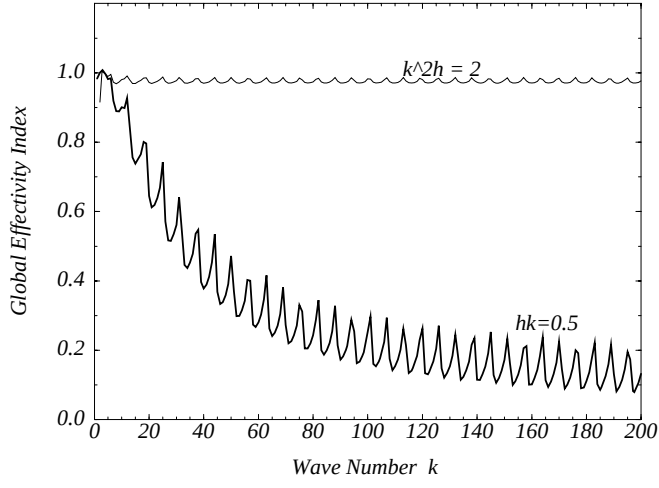


FIGURE 4.41. Variation of the global effectivity index with wave number. The upper line shows the effectivity indices on a sequence of fine meshes for which $k^2h = 2$. The lower line shows the deterioration of the effectivity index on meshes with $kh = 0.5$. The effectivity indeces have been computed from the the ZZ estimator.

4.11 Summary and Conclusions for Computational Application

The application of discrete methods to Helmholtz problems generally requires the resolution of different scales. For exterior problems, the FEM is applied in the near field. The far-field behavior has to be mapped onto the near field in such a way that the reduced problem is well-posed and the truncation error is controlled.

The FE model in the near field resolves the size of the computational domain, as well as the wavelength of the incident signal. These scales may differ in order of magnitude, leading to a large nondimensional wave number in a scaled model. The solutions are then highly oscillatory (rough). The variational operator is indefinite, and the stability of the model deteriorates with large wave number.

In the case of elastic scattering, the wave propagates through an inhomogeneous medium. Here, the numerical model suffers additional stability problems at the (discrete) eigenfrequencies of the interior problem.

We have investigated the reliability of finite element methods for large wave number. From our numerical results, the term "large wave number"

for the homogeneous case can be quantified roughly by the relation

$$\frac{L}{\lambda} \approx 10\,,$$

that is, the computational domain is ten times larger than the length of the incident wave. In our discussion, a wave number k means the nondimensional number

$$k = 2\pi\frac{L}{\lambda}.$$

Our investigation leads to the following conclusions.

Error Control for the h-Version of the FEM:

There is a principal difference between asymptotic and preasymptotic error behavior. The asymptotic estimates do not hold in the typical range of engineering computations. On meshes with kh =constant, the FE error has a pollution term the size of which corresponds to the phase error of the numerical solution. The pollution term, and thus the error, is unbounded for large wave numbers on meshes designed by the "rule of thumb" kh =constant.

For piecewise linear approximation, the error can be controlled by a rule k^3h^2=constant. That is, the number of elements computed from the "rule of thumb" should be augmented by $\sqrt{k}$, yielding a resolution

$$\frac{\lambda}{h} \approx 6\sqrt{k}\,. \tag{4.11.1}$$

This relation also controls the error in the L^2-norm. Computational experiments show that these rules, which are concluded from a priori error analysis, lead to reliable FE solutions. This has been confirmed in experiments on the original one-dimensional model problems, as well as in benchmark computations for higher-dimensional applications.

Regarding *a posteriori* error estimation, one cannot, in general, rely on the standard local estimators and indicators to control the error in the FE solution of Helmholtz problems. These errors effectively estimate a quasi-interpolating numerical solution and do not take into account the phase error. The true error is underestimated if the wave number is large.

Convergence Speedup:

Raising the order of approximation p leads to significant speedup of the h convergence, both in the interpolation and the pollution error. Taking into account the computational cost, a selection of $p = 2, 3$, or 4 has been found optimal in the one-dimensional model problem. The scale of the mesh for the *hp* version is

$$\theta = \left(\frac{hk}{p}\right)^p\,,$$

meaning that reliable results can also be expected on coarse meshes. It is possible to create specialized FEM, incorporating analytical information about the Helmholtz operator into the trial space. We have discussed the quasistabilized FEM and the partition of unity FEM.

Error Norm:

Most of the estimates are given in the H^1-norm (energy norm). Usually the interest in acoustic computations is in the primary variable (pressure p) rather than its derivative. Still, the H^1-norm is relevant for exterior problems if one is interested in the far-field response. In many FE approaches, this response is computed analytically from the Helmholtz boundary integral equation using the FE data on some "collecting surface" within Ω_a. Since this integral equation involves both p and its normal derivative, the H^1-norm is appropriate for error control in the near field. The situation is different for interior problems. Here, the L^2-norm may be more appropriate for error control. The numerical analysis of the convergence behavior and *a posteriori* error control in the L^2-norm is a matter of ongoing research.

Error control for Fluid–Solid Interaction Problems:

For elastic scattering, the numerical error is significantly polluted also for small wave numbers if the corresponding frequency ω is close to an eigenfrequency of the interior problem. Efficient error control in the vicinity of eigenfrequencies is an open problem.

Finite Element Modeling of Exterior Problems:

Since the pollution of the finite element error depends on both the frequency and the size of the computational domain, it is favorable to keep this domain as small as possible. This means that the artificial surface should be close to the obstacle.

4.12 Bibliographical Remarks

Numerous textbooks are available for an introduction to the FEM, for instance Akin [2], Bathe [18], or Hughes [71]. Since we placed special emphasis on the *hp*-version, we mostly follow the book by Szabó and Babuška [112]. The foundations of the FEM for elliptic partial differential equations can be found in a number of textbooks such as Breass [30], Brenner–Scott [31], Carey–Oden [33], or Schatz [109]. The classical monograph by Ciarlet [34] has been reedited in [35].

Regarding indefinite forms, the fundamental inf–sup condition was proven by Babuška in 1973 [7]. Schatz published his results on forms satisfying a Gårding inequality in 1974 [108]. The results are covered in Hackbusch's textbook [62]; see also Braess [30]. The application to Helmholtz forms is

more recent; see Aziz, Kellog and Stephens [6] and Douglas et al. [50]. Our proof of the asymptotic quasioptimality of the FE error is based on [50].

Several papers have been devoted to the dispersion analysis of the discrete Helmholtz operators; cf. Harari and Hughes [63, 64], Thompson and Pinsky [113, 114]. The preasymptotic estimates in this chapter were first published in joint papers with Babuška [72, 73, 74]. Our elaborations on numerical pollution are closely related to Wahlbin's article [117]. Compare, for example, our estimates of the pollution term for higher p to the following from [117, p. 377].

> "If a function is appreciably smaller in a negative norm than in, say, the L^2 norm, this is frequently due to oscillations. As an example, the reader may contemplate the functions $v_n(x) = \sin(nx)$ ... ."

Since the pollution in the discrete oscillatory solutions of the Helmholtz equation is caused by a global (namely, in the Helmholtz differential operator) rather than a local (geometric) singularity, our error estimates are global, too. Other than that, our main estimates (4.7.39) and (4.7.40) are similar to Wahlbin's "basic local estimates" [117, Theorems 9.1, 9.2].

On the generalized FEM for Helmholtz problems, see Babuška and Sauter [17]; the idea of the QSFEM is outlined in Babuška et al. [11]. The partition of unity FEM was first proposed in the PhD thesis of Melenk [91] and has since been developed in a series of papers by Melenk and Babuška [92, 93]. This method can also be viewed and implemented as a meshless method; see Babuška and Melenk [15].

The material of Section 4.10 is taken from our joint work with Babuška, Strouboulis and Gangeraj [12].

5

Computational Simulation of Elastic Scattering

We use the finite–infinite element program SONAX[1] for the computational simulation of elastic scattering. We first perform computational experiments with an elastic sphere where the exact solution is known and then proceed to the simulation of physical experiments for an elongated cylindrical shell embedded in water.

5.1 Elastic Scattering from a Sphere

5.1.1 Implementation of a Coupled Finite–Infinite Element Method for Axisymmetric Problems

The computer program SONAX computes the numerical solution for elastic scattering from rotationally symmetric obstacles by a Fourier finite–infinite element (FE–IE) method. It is especially geared towards the approximation of elongated scatterers, since the artificial boundary is, in general, a prolate spheroid. This allows the user to keep the FE region small, reducing the computational effort and the numerical pollution at large kL (here L is the principal length of the finite element domain around the obstacle).

The obstacle is modeled in cylindrical coordinates (r, ϕ, z); see Fig. 5.1. With the assumption of rotational symmetry, all structural and acoustic

[1]The program SONAX was created by Dr. Joseph Shirron at the Naval Research Laboratory, in Washington, D.C.

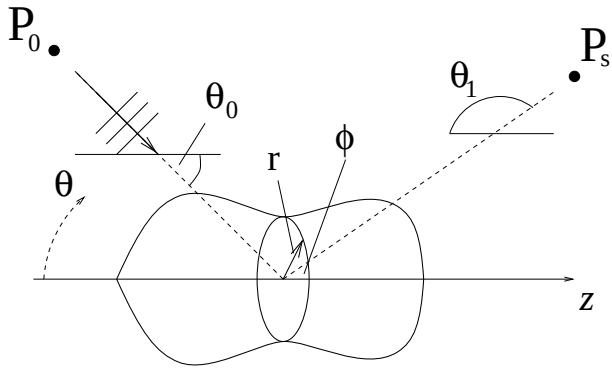


FIGURE 5.1. Rotationally symmetric obstacle insonified by a plane wave.

unknowns of the full three-dimensional model can be expanded in a Fourier series in the azimuthal angle ϕ. Then the degrees of freedom for the numerical model are introduced in the two-dimensional cross-section of the three-dimensional model. If the load is also symmetric, only the zeroth component of the expansion need be computed. The assumption of rotational symmetry is made only for the structure.

If the load is nonsymmetric, then the boundary data $g(r, z, \phi)$ is decomposed as

$$g(r, z, \phi) = \sum_{n=-\infty}^{\infty} g_n(r, z) e^{in\phi} .$$

A finite number of problems with rotationally symmetric loads $g_n(r, z)$ are solved, and the three-dimensional solution p is obtained as the superposition of the solutions $p_n e^{in\phi}$.

The load can be given as a combination of plane waves in arbitrary directions or as a combination of point sources (spherical waves). The source point P_0 is the far-field point determined by the polar angle θ_0. The load is given by user input in the form of a sweep over a frequency range $k = k_0, \ldots, k_1$ in steps of size Δk. For each frequency in the sweep, the scattered pressure p_s can be computed at arbitrary far-field points P_s. We thus define, for some given far-field point P_s, a frequency response function (FRF)

$$F(k) := p_s(P_s, k) .$$

The program SONAX features the possibilities of monostatic and bistatic frequency sweeps. In monostatic scattering, the locations of the source and receiver are fixed. If the source and receiver are at the same location ($\theta_0 = \theta_s$), the monostatic sweep returns the backscattered field. In bistatic scattering, the source is fixed at θ_0, but there are many receivers at various locations $\theta_s(i), i = 1, 2, \ldots$.

The far-field response is computed in several steps: First, we solve the FE–IE problem to obtain the numerical solution $p_h^N \approx p$ in the near field.

We then compute the normal derivative $\partial_\nu p_h^N$ on the wet surface. Finally, the integral

$$p(r) \approx \int_\Gamma \left[p_h^N(r') \partial_\nu g(r, r') - \partial_\nu p_h^N(r') g(r, r') \right] \, dS' \tag{5.1.1}$$

is computed. Since P_s is a far-field point, we use the asymptotic expression (2.1.29) for the free-space Green's function.

The finite elements in the three-dimensional elastic and fluid media are based on the hierarchical Legendre-type shape functions (cf. Babuška–Szabó [112]). The far-field behavior is resolved by infinite elements using the unconjugated Bubnov–Galerkin formulation (3.5.12). The test and trial spaces are identical, both being based on (3.5.2). The resulting system matrix is symmetric. In general, the artificial surface is a prolate spheroid, given in a bipolar coordinate system. On this surface, the infinite elements are constructed by the same principle as described for the sphere in Section 3.5. For details, see Burnett [32].

5.1.2 Model Problem

We analyze scattering from an elastic sphere. The exact solution for this problem is given in Section 2.1. It has been found to be in good agreement with experimental results; cf. Junger and Feit [81, pp. 352–356].

This model is chosen as a benchmark for SONAX. In particular, we are interested in the convergence behavior of the FE–IEM with respect to the numerical parameters h, p (finite elements), and N (infinite elements). We consider three different spheres, S05, S15, and S25, with uniform thicknesses $t = 0.05\,\mathrm{m}$, $t = 0.1\,\mathrm{m}$, $t = 0.25\,\mathrm{m}$, respectively. All spheres have the same midsurface radius $a = 5\,\mathrm{m}$. The material parameters for all spheres are given in Table 5.1.

TABLE 5.1. Material parameters.

E	=	$2.07E + 11\,\mathrm{P}$	Young's modulus
ρ_s	=	$7669\,\mathrm{kg/m^3}$	solid density
ρ_f	=	$1000\,\mathrm{kg/m^3}$	fluid density
ν	=	0.3	Poisson's ratio
c_f	=	$1524\,\mathrm{m/s}$	fluid speed of sound

A plane wave of unit amplitude is incident on the shells, and the far-field pattern of the backscattered pressure is computed.

In Fig. 5.2, we show the exact frequency response function for the three shells, i.e., the absolute value of the far-field pattern in the backscattered direction. We observe that no resonance peaks are formed at the higher

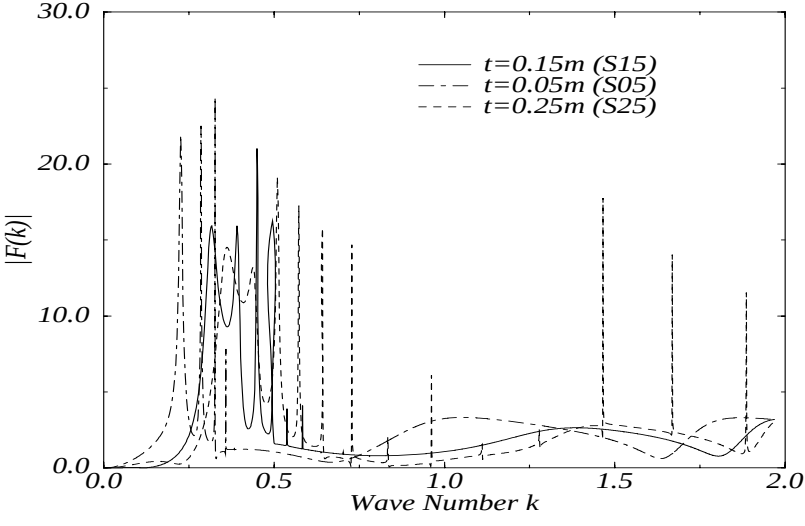


FIGURE 5.2. Far-field pattern of backscattered pressure for elastic spherical shells S05, S15, S25.

TABLE 5.2. Wave numbers corresponding to eigenfrequencies of free vibration.

1	2	3	4	5	5	7	8
0.501	0.596	0.639	0.669	0.701	0.741	0.794	0.862

eigenmodes for the shells S05 and S15, whereas they are still visible for the "thick" shell S25. For S05, only the first four eigenmodes are visible in the FRF plot, whereas the first six peaks can be distinguished for S15. The eigenfrequencies for the free vibrations of the shell in vacuum are the zeros of the mechanical impedance; cf. the denominator in (2.1.20). We list the wave numbers corresponding to the first eight frequencies in Table 5.2.

The in-fluid damped eigenvalues are shifted to the left from their positions in free vibration. The shift is rather significant since the acoustic impedance z_n is complex with a nonvanishing real part; cf. the denominator in (2.1.19). For illustration, see in Fig. 5.3 the damped solution (with ρ_s, ρ_f as in Table 5.1) compared to the weak damping $\rho_f = 1.0\,\mathrm{kg/m^3}$. In the FRF for the weakly damped case, the peaks occur precisely at the eigenvalues in Table 5.2.

We now turn to the evaluation of the numerical discretizations. The computational experiments with the FE–IE model are performed for the shell S15 only. We are interested in wave numbers $k = 0.25, \ldots, 2\,\mathrm{m}^{-1}$, which correspond to the nondimensional frequency range $ka = 1.25, \ldots, 10$.

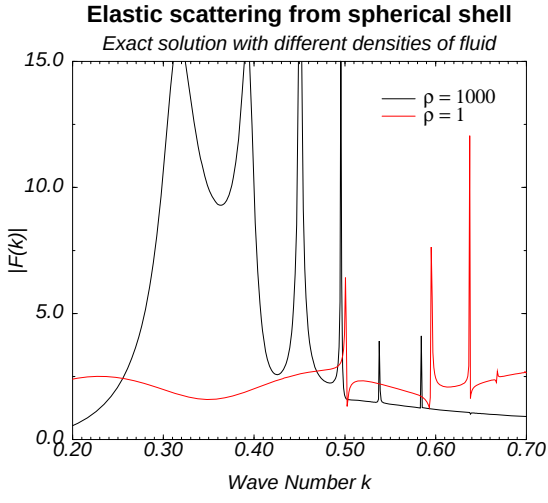


FIGURE 5.3. Scattering from elastic shell with different densities of acoustic medium.

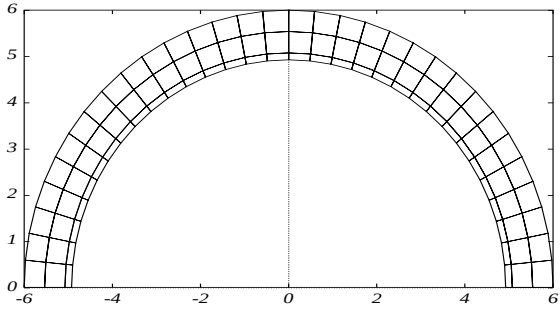


FIGURE 5.4. Spherical shell with mesh a32-r2-R600.

Relating the wavenumber to the size of the computational domain, we get the nondimensional wave number $K = ka\pi = 0.8, \ldots, 32$.

In Fig. 5.4, we show a typical mesh for the finite element discretization of the model (after reduction to two dimensions by Fourier expansion in the azimuthal angle ϕ). The thin-walled elastic medium is partitioned into one layer of finite elements. The shell is enclosed in a spherical artificial boundary of radius R_a. We choose a uniform angular discretization and introduce a number of element layers in the radial direction. In the plot, we show the mesh for the shell S15 with 32 elements in the angular direction and 2 element layers in the fluid. The artificial boundary is located at $R_a = 6\,\mathrm{m}$. For this mesh, the mesh sizes in the angular and radial directions,

respectively, are

$$h_\theta = \frac{\pi a}{32} = 0.491\,\mathrm{m}, \qquad h_r = \frac{R - a + t/2}{2} = 0.4625\,\mathrm{m},$$

and hence the magnitude of kh varies between 0.1 and 1. That is, for the largest k the mesh stays below the minimal resolution $kh = \pi/2$ for piecewise linear approximation. The recommended mesh size by the rule $\lambda/h = 10$ corresponds to a magnitude $kh = \pi/5 \approx 0.62$. For the mesh as shown, this resolution is matched for $k = 1.28\,\mathrm{m}^{-1}$, i.e., roughly in the middle of the frequency sweep. Convergence studies are carried out on a series of meshes shown in Table 5.3.

TABLE 5.3. Meshes for convergence tests.

·	·	a16-r2-R600	·	·
a32-r2-R525	a32-r2-R550	a32-r2-R600	a32-r4-R600	a32-r8-R600
·	·	a64-r2-R600	·	·

In the table, a and r denote the angular and radial partitions, respectively, and R is the radius of the artificial boundary. On these meshes, we will test the effect of polynomial enrichment as well as angular and radial h-refinement. The mesh a32-r2-R600 is the pivot mesh. The meshes for angular refinement are given in the vertical sequence of the meshes in Table 5.3, while the meshes for radial refinement are given to the right of the pivot. Finally, we will use the meshes a32-r2-R525 and a32-r2-R550 to test the sensitivity of the FE–IE coupling to the size of the FE domain and the number of DOF in the infinite elements.

5.1.3 Computational Results

Polynomial Enrichment:

In the following, we adopt the notation of p- or q-convergence, respectively, for the solution behavior with respect to polynomial enrichment in the solid (degree p) or in the fluid (degree q). We first consider pq-convergence on the coarse mesh a16-r1-R600 (16 elements in the angular and one in the radial direction). In Fig. 5.5 we show q-convergence with piecewise linear elements in the shell. Due to the underresolution in the shell, the model converges to the wrong solution. For $p = 2$ the shell is sufficiently well resolved; see Fig. 5.6. Still the fluid is underresolved for $q = 1$, causing a shift in the lower damped eigenfrequencies and a deviation from the exact solution for higher wave numbers. The same effect is observed for $p = 3$ and $p = 4$; we show here $p = 4$ in Fig. 5.7. Clearly, the minimal degree of approximation on the coarse mesh is $p = q = 2$, that is, quadratics both in the shell and in the fluid. This conclusion is also supported by the plots of

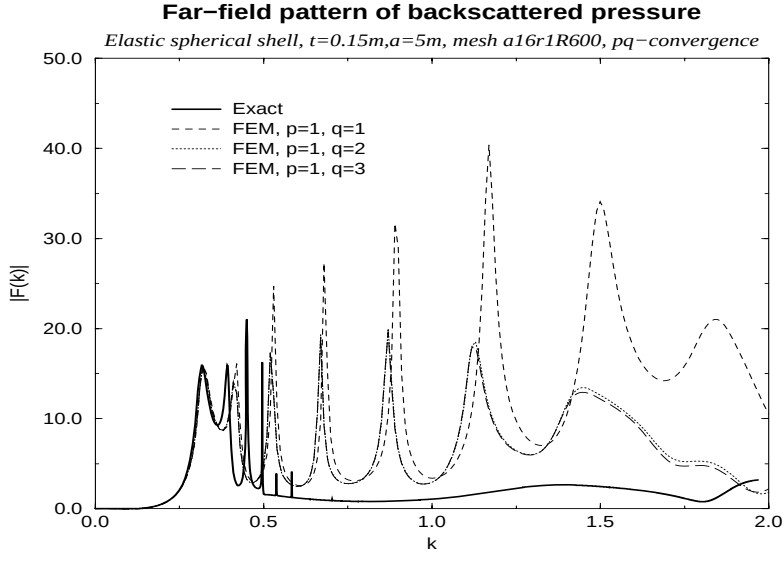


FIGURE 5.5. q-Convergence on coarse mesh with $p = 1$.

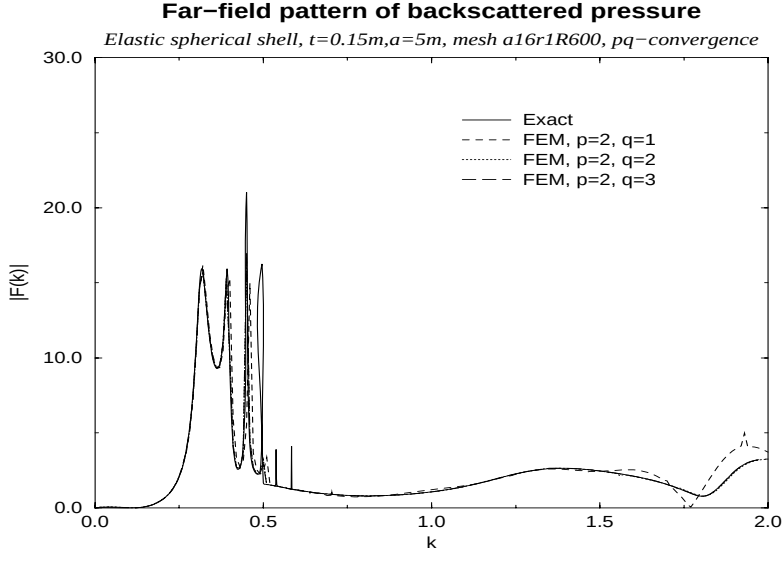


FIGURE 5.6. q-Convergence on coarse mesh with $p = 2$.

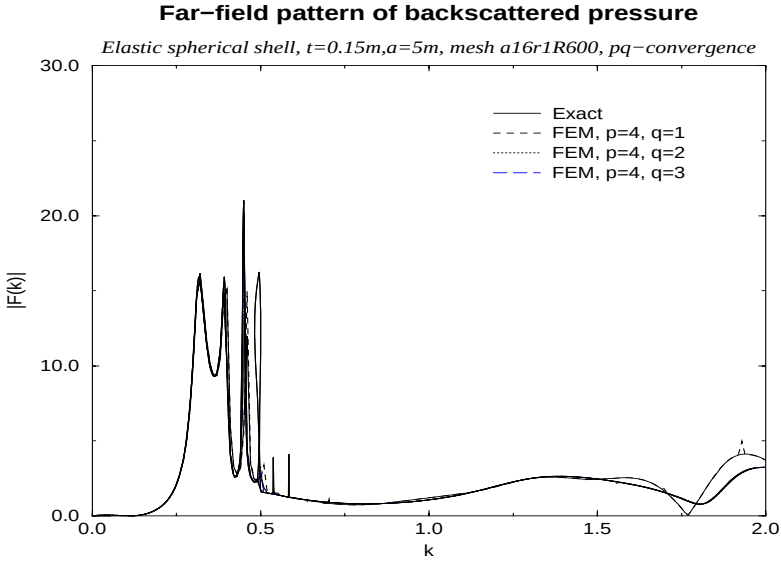


FIGURE 5.7. q-Convergence on coarse mesh with $p = 4$.

p-convergence with fixed q. In Fig. 5.8, we plot the curves with $p = 1, 2, 3$ for $q = 2$. We get very good agreement if $p = 2$ or $p = 3$. For $p = 3$ and $q = 2$, the FEM model also resolves the higher eigenfrequencies (5th and 6th), provided the step size of the frequency sweep is chosen sufficiently small, as shown in the detail of Fig. 5.9.

Let us relate these observations for the FRF to the error analysis in Chapter 4. For the one-dimensional model problem, the relative error in the H^1-norm at $K = 10$ is about 20% for the best approximation, compared to 25% for the finite element solution. Thus there is no significant pollution for wave number, but the relative error in the H^1-norm is large due to insufficient mesh refinement.[2] The nondimensional wave number $K = 10$ corresponds to $k = 0.6\,\mathrm{m}^{-1}$ in the plot of the frequency response. Hence, for $q = 1$, the fluid part is computed with considerable error, contributing to the shift in the higher eigenvalues. On the other hand, the estimate of the relative error for quadratic approximation in the fluid is of order

$$\left(\frac{kh}{4}\right)^2 \approx 2\%$$

(pollution is negligible). Correspondingly, we see very good agreement in the frequency response curves for higher approximation.

[2]Note that both p and its first derivative are used to compute the far-field response by (5.1.1), hence the H^1-norm is the appropriate error norm for far-field computations.

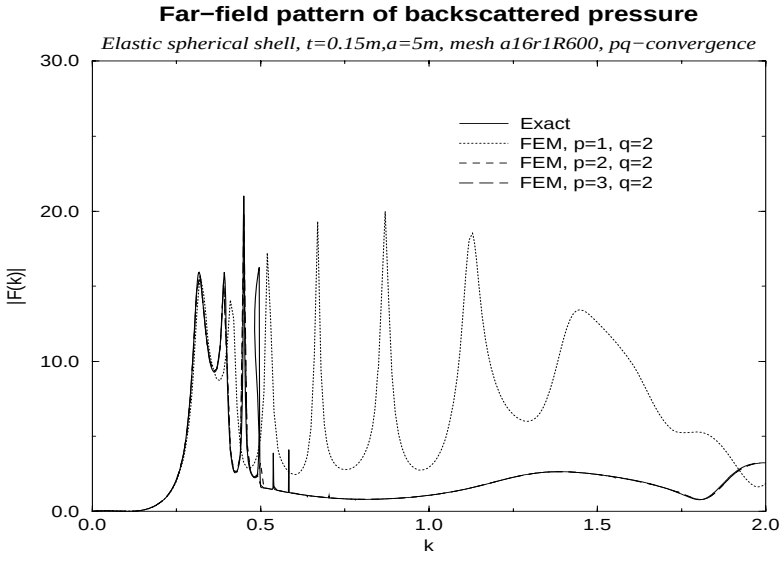


FIGURE 5.8. p-Convergence on coarse mesh with $q = 2$.

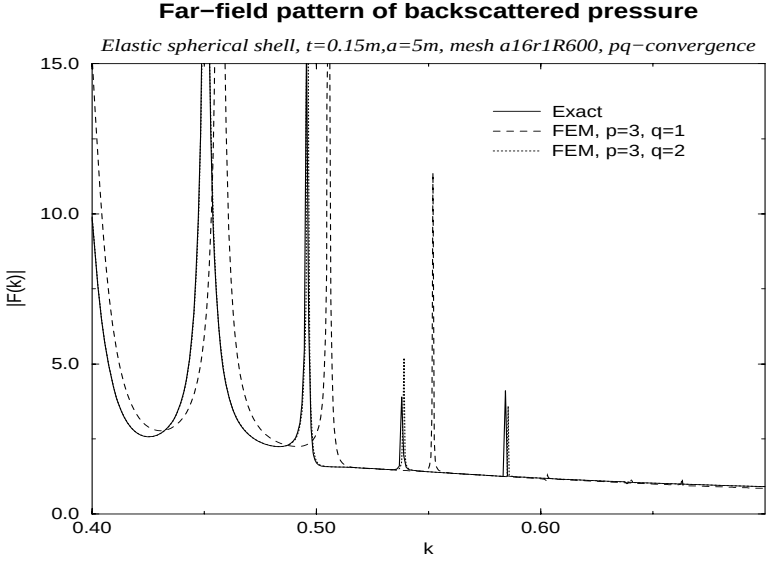


FIGURE 5.9. q-Convergence on coarse mesh with $p = 3$; details for higher eigenfrequencies.

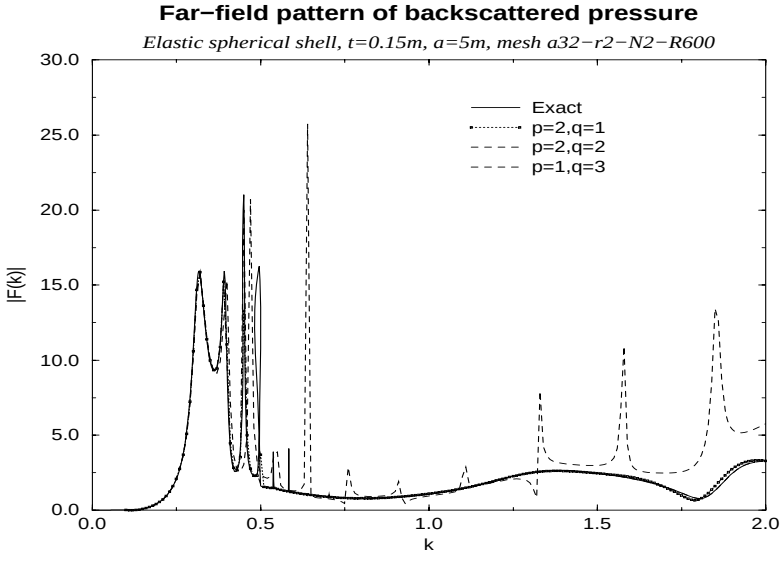


FIGURE 5.10. pq-Convergence on mesh with 32 angular and 2 radial (fluid) elements.

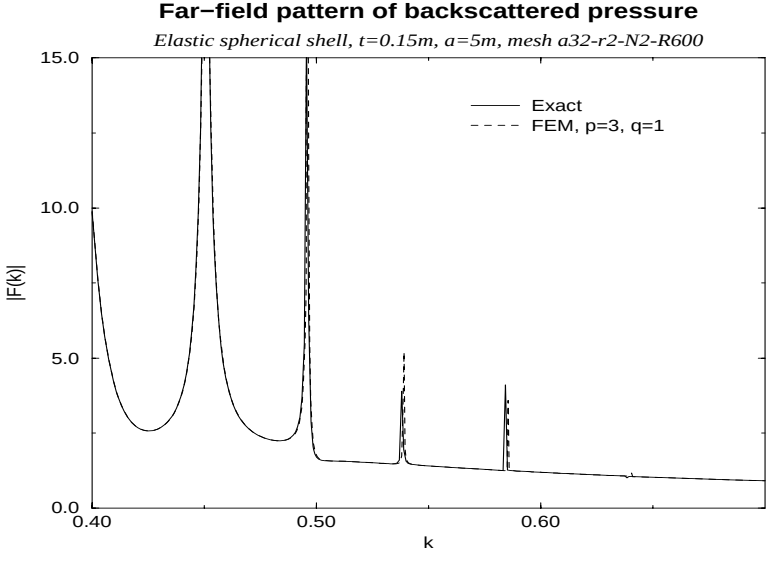


FIGURE 5.11. q-Convergence on mesh a32r2, $p = 3, q = 1, 2$; details for higher eigenfrequencies.

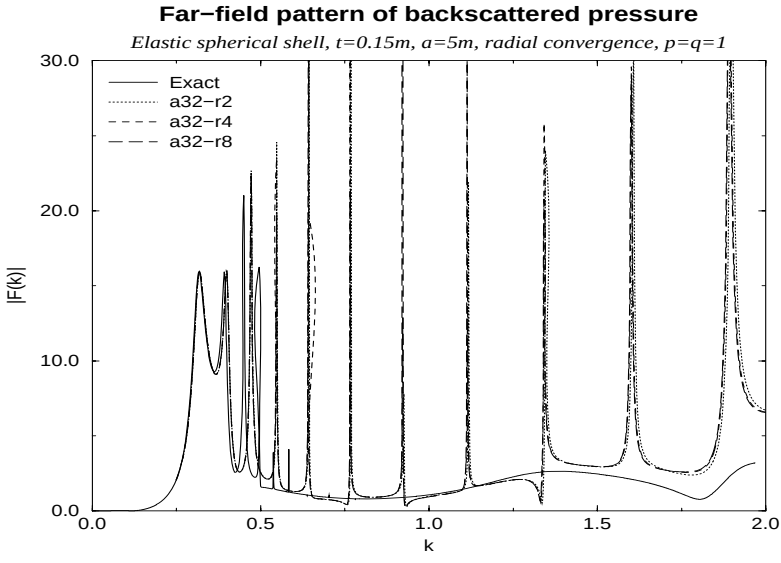


FIGURE 5.12. Radial convergence on mesh with 32 angular elements, $p = q = 1$.

Mesh Refinement:

In Fig. 5.10 we show results from the finer mesh with 32 angular elements and 2 radial fluid elements. A pq-convergence study leads to similar conclusions as far as the solid resolution is concerned. Also for this mesh, the application of piecewise linear elements in the solid implies q-convergence to the wrong solution. On the other hand, with $p = 2$ we now see good agreement in the higher eigenmodes also for $q = 1$, indicating the effect of h-convergence in the fluid. This is shown in detail in Fig. 5.11.

In Fig. 5.12, we show the FRF on radially refined meshes with $p = q = 1$. Again we observe convergence to the wrong solution. The radial refinement in the fluid is, obviously, no remedy for the insufficient resolution of the shell.

Angular refinement, which is performed simultaneously in the shell and in the fluid, improves the quality of the numerical FRF. On a mesh with 64 angular linear elements, we achieve about the same agreement with the exact FRF as on the mesh with 16 elements and quadratic elements in the shell; cf. Fig. 5.13. However, also on the finest mesh we observe small peaks in the numerical solution. Taking quadratic elements in the shell and linear elements in the fluid, angular refinement from 16 to 32 elements yields very good agreement (not shown here).

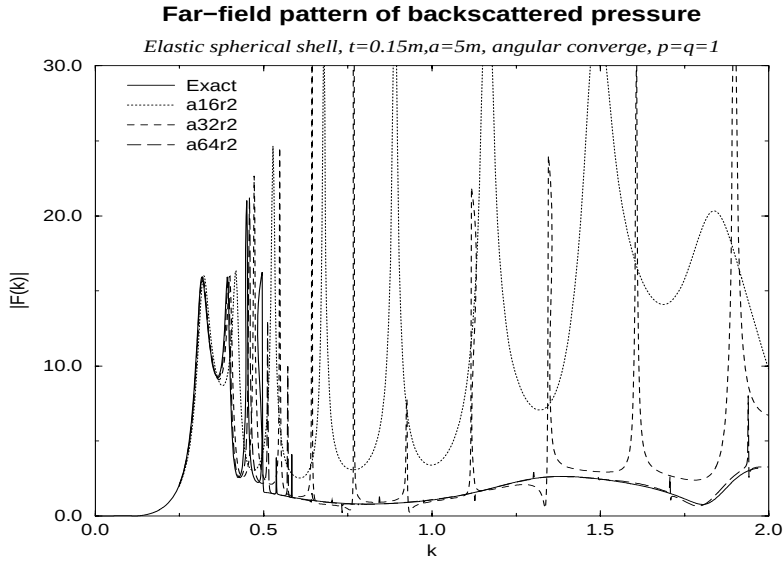


FIGURE 5.13. Angular convergence on mesh with 2 radial (fluid) elements, $p = q = 1$.

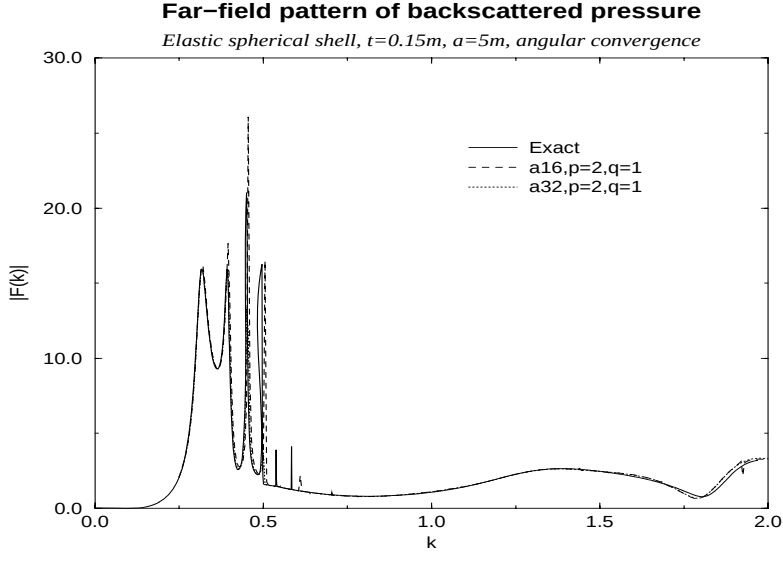


FIGURE 5.14. Angular convergence on mesh with 2 radial (fluid) elements, $p = 2, q = 1$.

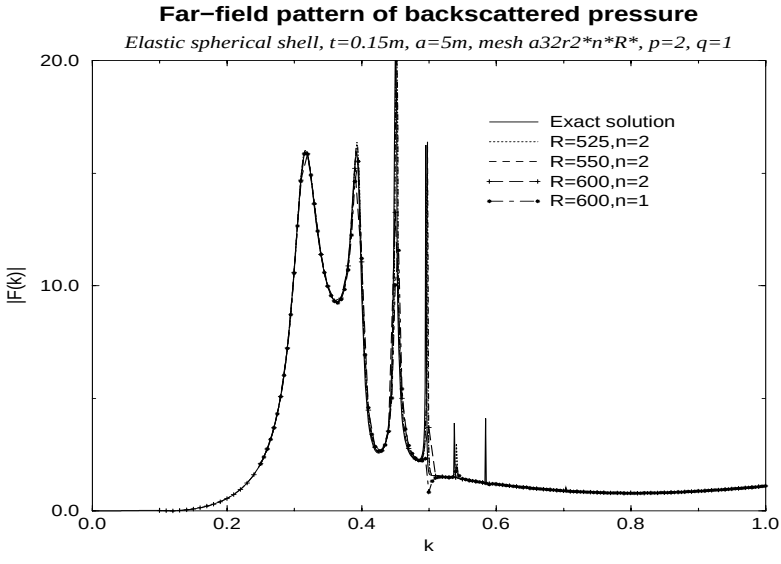


FIGURE 5.15. Influence of far-field resolution.

Resolution of the Far Field:

Finally, we consider the influence of far-field resolution. In Fig. 5.15, we plot results with $R_a < 600$ (i.e., the artificial boundary is moved closer to the obstacle). For the same number of DOF in the infinite elements, the agreement between the FRF curves does not visibly deteriorate as the size of the FE region is decreased. Even reducing the number of exterior DOF to the minimal $N = 1$ does not cause major disagreement with the exact response. Only a slight perturbation of the FRF is observed at the fourth eigenvalue.

5.1.4 Conclusions

The quality of the computed frequency response is more sensitive to the numerical resolution of the solid than the numerical resolution of the fluid. Though the positions of the frequency peaks are influenced both by the mechanical and by the acoustic impedance, underrefinement in the solid causes a larger error in the position of the frequency peaks than underresolution in the fluid.

The approximation of the far-field behavior with infinite elements influences the quality of the numerical FRF less significantly. In particular, the size of the computational domain may be rather small, and for the wave numbers considered, only a small number of radial shape functions is needed.

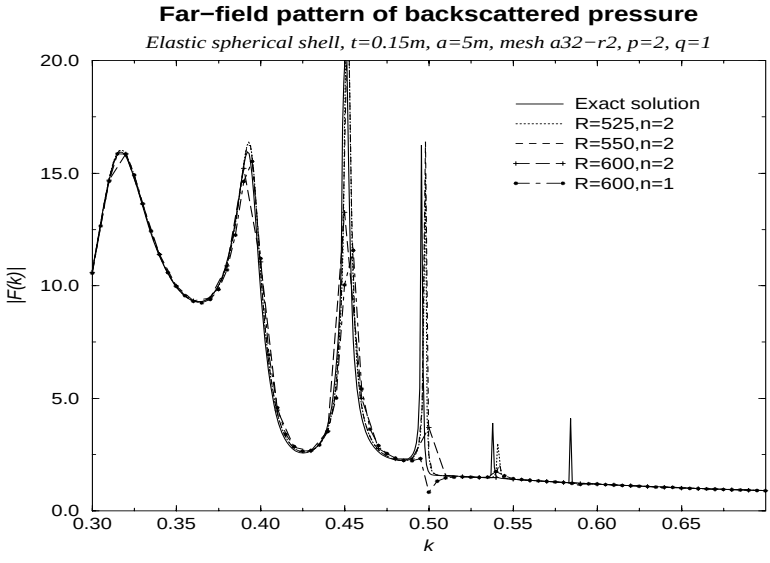


FIGURE 5.16. Influence of far-field resolution, detailed plot.

5.2 Elastic Scattering from a Cylinder with Spherical Endcaps

5.2.1 Model Parameters

We consider a hemispherically capped cylinder, called mock shell; see Fig. 5.17.

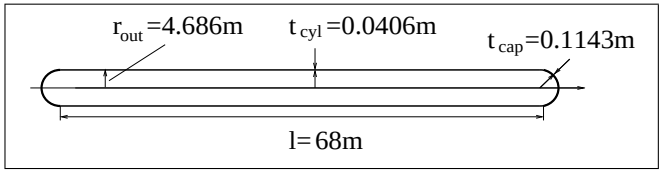


FIGURE 5.17. Cylindrical shell with spherical endcaps.

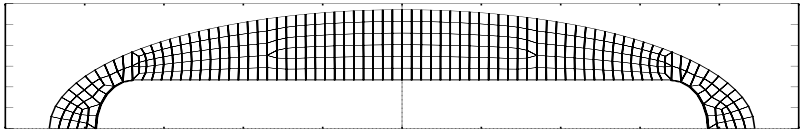
The outer radius of the shell is 4.6863 m, the length of the circular cylinder is 67.9958 m, the cylinder thickness is 0.04064 m, and the endcap thickness is 0.1143 m. The material parameters are given in Table 5.4.

We will use two different meshes; see Fig. 5.18. The size of the coarse mesh is $H = 1.25$ m in both directions, and the size of the fine mesh is $h = H/2$. We will typically compute a frequency sweep of $ka = 2.5, \ldots, 20$, which corresponds to

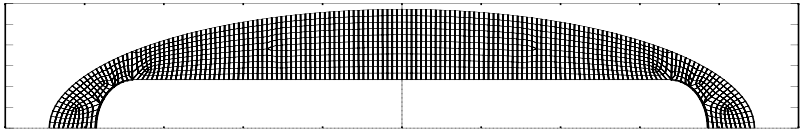
$$\frac{\lambda}{H} = 10, \ldots, 1; \qquad \frac{\lambda}{h} = 20, \ldots, 2.$$

TABLE 5.4. Material parameters of capped shell.

E	=	$2.0E+11\,\mathrm{P}$	Young's modulus
ρ_s	=	$7908.5\,\mathrm{kg/m^3}$	solid density
ρ_f	=	$1000\,\mathrm{kg/m^3}$	fluid density
ν	=	0.29	Poisson's ratio
c_f	=	$1482\,\mathrm{m/s}$	fluid speed of sound



(a) Coarse mesh H



(b) Fine mesh h

FIGURE 5.18. Mock shell in fluid; coarse and fine mesh for computations.

That is, the coarse mesh is for all wave numbers of the sweep below the recommended resolution for piecewise linear approximation. The fine mesh has the recommended resolution of 10 in the middle of the sweep. It is for all k finer than the minimal resolution $n = 2$.

To get an estimate of the pollution effect, we have to consider a nondimensional wave number that reflects the size of the computational domain. In this elongated structure with $L \gg a$, the nondimensional wave number is $K = kL$, where $L = 90$ m measures the long principal axis of the spheroid. The frequency sweep is then $K = 45, \ldots, 450$; hence we have to expect significant pollution of the numerical results at the higher end of the sweep.

5.2.2 Convergence Tests

We first perform convergence tests to establish confidence in the computational simulation. In Fig. 5.19, we illustrate the pq-convergence on the coarse mesh for $k = 1, \ldots, 3$. On the abscissa we show the magnitude of

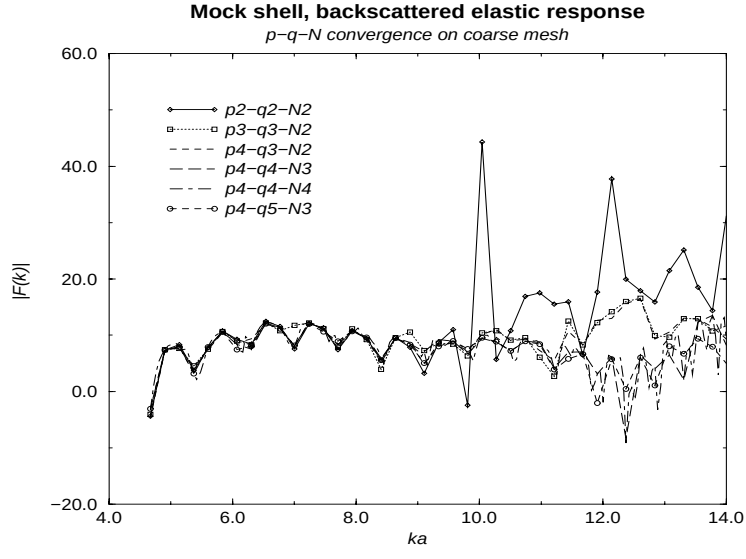


FIGURE 5.19. Far-field pattern of backscattered pressure for the mock shell. *pqN*-Convergence on coarse mesh.

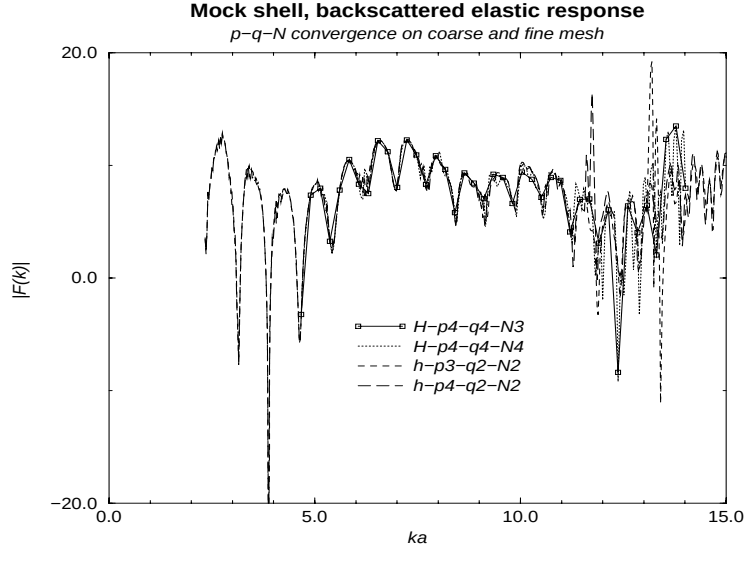


FIGURE 5.20. Far-field pattern of backscattered pressure for the mock shell. *pq*-Convergence on coarse and fine mesh.

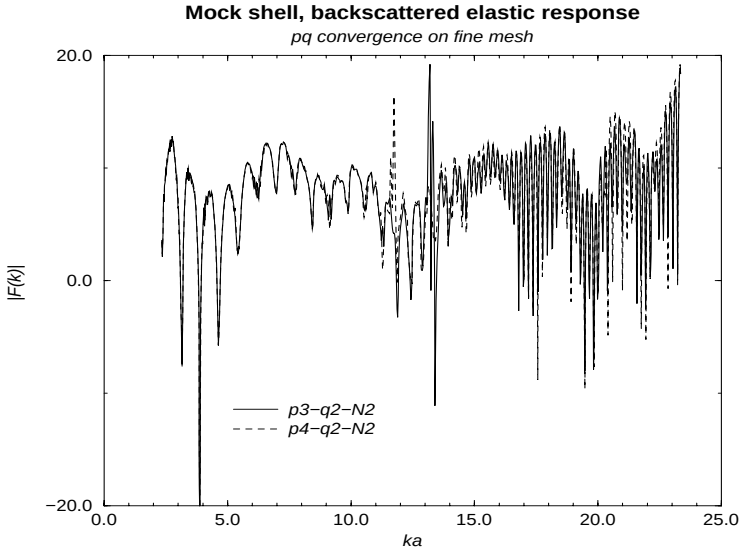


FIGURE 5.21. Far-field pattern of backscattered pressure for the mock shell. pq-Convergence on fine mesh.

ka, whereas the far-field pattern of the scattered pressure is measured on the vertical axis (see Remark 5.1 below for the scaling of the data and the unit of measurement). A coarse step size Δk has been taken in the frequency sweep. The legend shows the values for the numerical parameters p, q (polynomial degrees in the solid and fluid mesh, respectively), and N (number of radial shape functions in the infinite elements).

For $p, q \leq 3$, convergence is observed only in the range $k = 1, \ldots, 2$. At the higher end of the sweep, good mutual agreement is found only between the curves with higher polynomial resolution. We thus assume that these latter results are convergent. This assumption is validated in the next plot, Fig. 5.20, where we compare the results on the coarse mesh with data computed on the fine mesh. Except for the peaks[3] at $ka = 12$ in the $p3$-$q2$-$N2$ line and at $ka = 13$ in the $p4$-$q2$-$N2$ line, the results on the fine mesh are in good mutual agreement. Both curves are closely fitted by the coarse mesh data up to $ka \approx 13$. We thus have enough confidence to recompute the fine mesh data with a fine resolution of the frequency sweep. For the curves as shown in Fig. 5.21, the sweep was $k = 0.5 : 0.005 : 4.29$, consisting of 759 separate FE–IE computations. In our implementation of SONAX on RISC workstations, this sweep took up 52 minutes of CPU time for the case $p3$-$q2$-$N2$ on the fine mesh. This time can, of course, be significantly

[3]Up to now we have no explanation for these peaks.

reduced by using a frequency-adaptive approach, where a coarse mesh with low pq is used at the beginning of the sweep, and the numerical resolution is increased in several steps during the sweep.

Remark 5.1. The frequency response is given as the target strength, which is precisely defined as $\mathrm{TS} = 20 \log_{10} |F(k)|$, where the far-field pattern in the computation of $F(k)$ has been scaled to 1 meter. The appropriate unit for the TS is decibel (dB). In the plots, we write $|F(k)|$ for brevity.

5.2.3 Comparison with Experiments

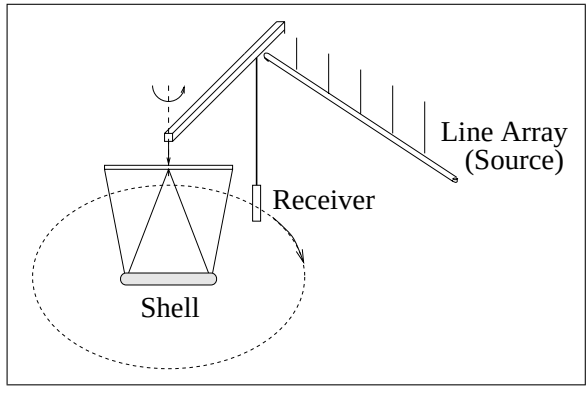


FIGURE 5.22. Setup of experiments.

Experimental studies on the scattering from the mock shell were carried out by Dr. Brian Houston at the Naval Research Laboratory. The setup of the experiments is sketched in Fig. 5.22. The test model is a shell that has been precisely fabricated with ratio 1:50 of the original model. The experimental results are scaled back to the true scale, where the computations with the FE-IE model are performed. The shell is embedded in a large tank containing water. A plane wave is simulated by a line array of point sources. The response is measured by a receiver. The source is fixed. Both the receiver and the shell can be rotated independently. To measure backscattered data, the receiver is aligned with the source (monostatic measurements). The center of the shell is located at a distance $l = 3$ m from the center of the line array. The reproducibility of the experimental data has been carefully checked by repeated measurements. The curves that will be shown here are the averaged representations of a series of measurements.

In Fig. 5.23, we show the comparison between the experiment and the simulation for the backscattered elastic response at bow incidence. The shell is positioned such that its long axis is perpendicular to the line array. The

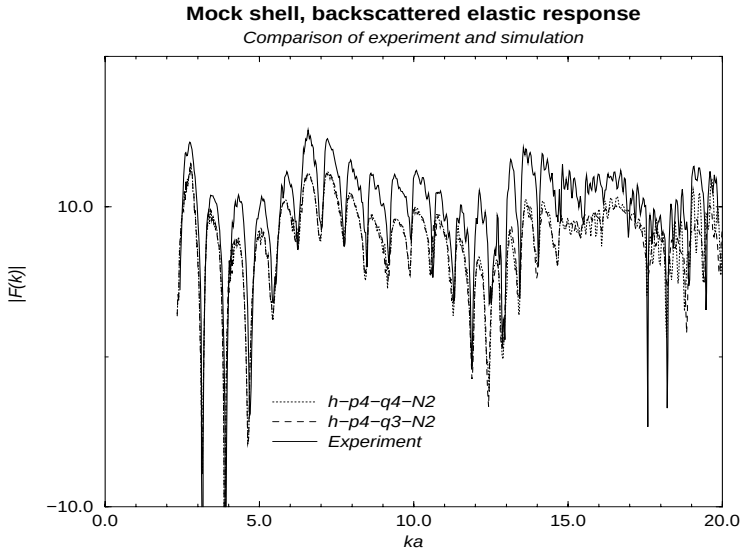


FIGURE 5.23. Far-field pattern of backscattered pressure for the mock shell. Comparison of high-resolution simulation and experiment. Plane wave insonification.

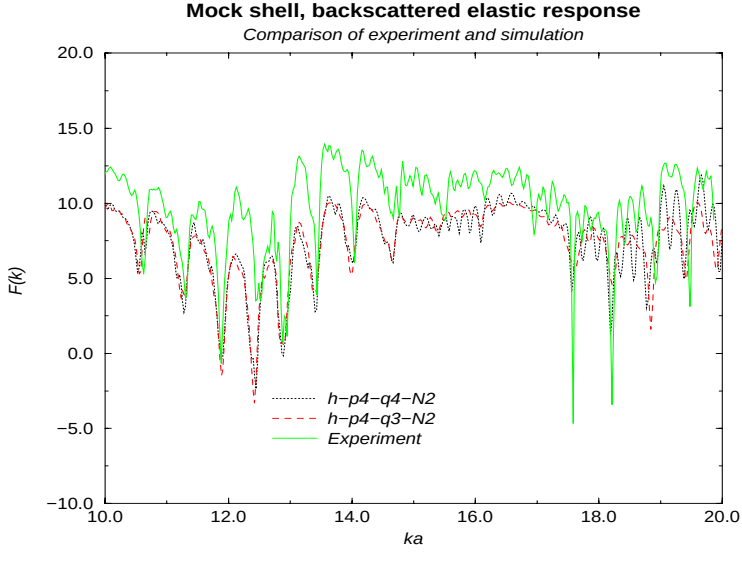


FIGURE 5.24. Far-field pattern of backscattered pressure for the mock shell. Detail at the higher end of the frequency sweep.

simulation very well reflects the k-dependent pattern of the experiment, but the agreement in the amplitude is not satisfactory. The oscillations at the high end of the sweep are better reproduced with degree $q = 4$; see also the detailed plot in Fig. 5.24.

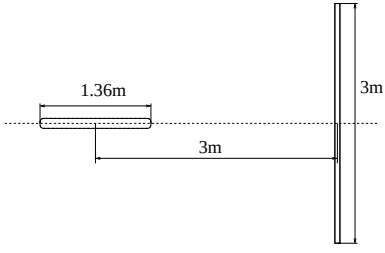


FIGURE 5.25. Situation for end-on insonification.

Still the amplitudes are not in good agreement. Obviously, this is not due to the lack of convergence in the simulation. Reconsidering the experimental setup, we were led to the hypothesis that the shell, at bow incidence of the wave, responds to a point source rather than a plane wave (see a sketch of the situation in Fig. 5.25). Hence the computations were repeated with a point source as load input. The results are shown in Fig. 5.26. We now observe a very good agreement between computational and experimental data throughout the range of the frequency sweep.

In a final numerical experiment (here the load is again a plane wave), we investigate the sensitivity of the numerical model with respect to coarsening of the grid. Recall that the numerical parameters of the high-fidelity computations had been determined, based both on our *a priori* studies of error behavior and on the convergence tests described above. The minimal resolution for high confidence of the simulation had been established to be $p = 4$, $q = 3$, and $N = 2$ (on the fine mesh of size h), and this was validated by comparison with the experimental data.

In Fig. 5.27, we compare the experimental data with the results from computations with smaller numerical resolution. Lowering the polynomial degree on the fine mesh, we compute a sweep that becomes blurred and unstructured at $ka \approx 13$. On the coarse mesh with the same polynomial degree of approximation, the computational sweep is in good agreement only up to $ka \approx 9$ (except for an unexplained peak at $ka \approx 5.6$). This is better than expected, taking into account the pollution effect, which should be significant on the coarse mesh. We assume that we are dealing with the effect of smoothening of the near-field data, which takes place when the far-field data are computed from the integral representation of the exterior solution.

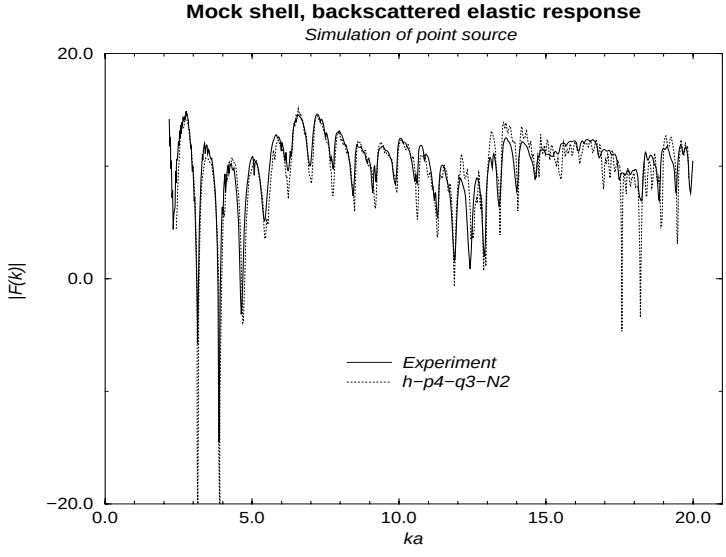


FIGURE 5.26. Far-field pattern of backscattered pressure for the mock shell. Point source insonification.

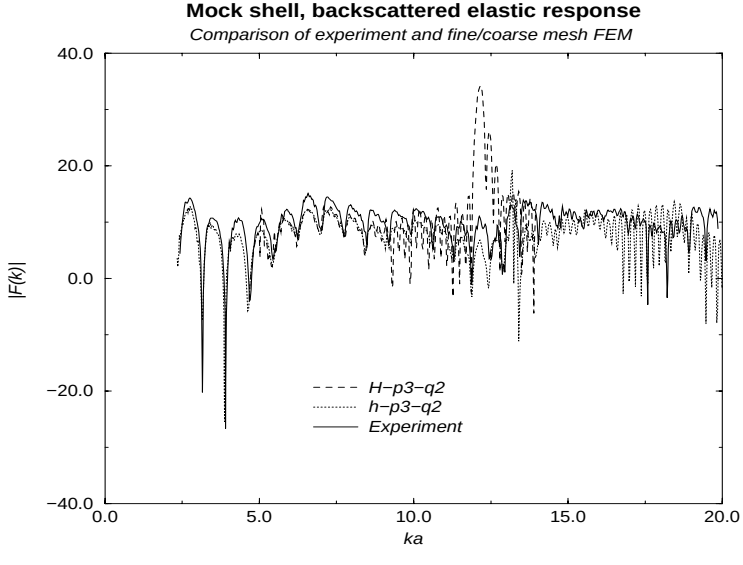


FIGURE 5.27. Far-field pattern of backscattered pressure for mock shell. Comparison of simulation on fine and coarse mesh with experiment. Plane wave insonification.

5.3 Summary

We report on our experience from computational simulations of elastic scattering using finite and infinite elements. The computations were performed with the *hp* FE–IE program SONAX, written by Dr. Joseph Shirron, Naval Research Laboratory, Washington D.C.

The testing is carried out in several steps. First, we analyze an obstacle for which the exact solution can be found analytically. The main purpose is to establish confidence in the method and the implementation and to test the influence of the computational parameters on the accuracy of the discrete solution. We then turn to the application, where the exact solution is not known. Here, we test *hp*-convergence of the discrete solution. These tests are carried out, based on the study of the error behavior in Chapter 4. After a sufficient resolution has been found, the results for this model are compared to experimental data. The measured and computed frequency response functions are found to be in good agreement.

The implementation shows that the FEM can be used as an effective and convenient numerical tool for the computational analysis of acoustic fluid–structure interaction. The high efficiency is due partly to the application of the Fourier FEM with the corresponding dimensional reduction (at the expense of a restriction to axially symmetric objects). In the present computations, we considered symmetric (with respect to angle ϕ) loads only. In related tests, we also obtained good agreement for angular incidence, performing computations at $\phi = 45°$ and $\phi = 70°$. In this case, we need to superimpose several Fourier modes. To the best of our knowledge, an analysis of the Fourier FEM for the Helmholtz equation is still lacking. For the Laplace equation, see Heinrich [67].

Regarding the numerical parameters, we make the following observations: first, the coupled model is more sensitive to insufficient numerical resolution of the structure, compared to the resolution in the fluid (here we mean only the resolution of the near field that is discretized by FEM). Second, the degree of approximation in the far-field part of the fluid (discretized by infinite elements) has only little influence on the accuracy of the far-field response. We recall that this response is computed analytically, using the Helmholtz integral equation, from the near-field FE data. In the example considered, even several successive reductions of the FE domain did not significantly influence the final result. Third, results obtained by polynomial enrichment on a coarse mesh were generally of higher quality than results of the *h*-version with approximation $p = 1$. This especially concerns the discretization of the shell.

References

[1] M. Abramovitz and I.A. Stegun, *Handbook of Mathematical Functions*, National Bureau of Standards, 1964.

[2] J.E. Akin, *Finite Elements for Analysis and Design*, Academic Press, London, 1994.

[3] H.W. Alt, *Lineare Funktionalanalysis*, 2. Auflage, Springer-Verlag, Berlin, 1992.

[4] R.J. Astley, G.J. Macaulay, J.-P. Coyette, and L. Cremers, Three-dimensional wave-envelope elements of variable order for acoustic radiation and scattering. Part I. Formulation in the frequency domain, *J. Acoust. Soc. Am.* 103 (1) 1998, 49-63.

[5] K.E. Atkinson, *An Introduction to Numerical Analysis*, second edition, J. Wiley, 1989.

[6] A.K. Aziz, R.B. Kellogg, and A.B. Stephens, A two-point boundary value problem with a rapidly oscillating solution, *Numer. Math.* 53, 107–121 (1988).

[7] I. Babuška and A.K. Aziz, The mathematical foundations of the finite element method. In: A.K. Aziz (ed.), *The mathematical foundations of the finite element method with applications to partial differential equations*, Academic Press, New York, 1972, 5–359.

[8] I. Babuška, A. Craig, J. Mandel, and J. Pitkäranta, Efficient preconditioning for the p-version finite element method in two dimensions, *SIAM J. Numer. Anal.* 28 (1991), 3, 624–661.

[9] I. Babuška and H. Elman, Performance of the hp-version of the FEM with various elements, *Int. J. Numer. Meth. Eng.*, 36, 2503–2523 (1993).

[10] I. Babuška, M. Griebel, and J. Pitkäranta, The problem of selecting the shape functions for a p-type finite element, *Int. J. Numer. Meth. Eng.*, 28, 1891–1908 (1989).

[11] I. Babuška, F. Ihlenburg, E. Paik, and S. Sauter, A generalized finite element method for solving the Helmholtz equation in two dimensions with minimal pollution; *Comp. Methods Appl. Mech. Eng.*, 128 (1995) 325–359.

[12] I. Babuška, F. Ihlenburg, T. Strouboulis, and S. Gangaraj, A posteriori error estimation for FEM solutions of Helmholtz's equation — Part I: the quality of local indicators and estimators, *Int. J. Numer. Meth. Eng.* 40, 3443–3462, 1997.

[13] I. Babuška, F. Ihlenburg, T. Strouboulis, and S. Gangaraj, A posteriori error estimation for FEM solutions of Helmholtz's equation — part II: estimation of the pollution error, *Int. J. Numer. Meth. Eng.* 40, 3883–3900, 1997.

[14] I. Babuška, I.N. Katz and B.S. Szabó, *Finite element analysis in one dimension*, Lecture Notes (unpublished), 1985.

[15] I. Babuška and J.M. Melenk, The partition of unity method, *Int. J. Numer. Meth. Eng.*, 40, 727–758, 1997.

[16] I. Babuška and A. Miller, A Feedback Finite Element Method with A-posteriori Error Estimation — Part 1: The Finite Element Method and Some Basic Properties of the A Posteriori Error Estimator, *Comp. Methods Appl. Mech. Eng.*, 61, 1–40, 1987.

[17] I. Babuška and S. Sauter, Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers, *SIAM J. Numer. Anal.* 34 (6), 2392–2423, 1997.

[18] K.J. Bathe, *Finite Element Procedures in Engineering Analysis*, Prentice Hall, Englewood Cliffs, 1982.

[19] A. Bayliss, C. Goldstein, and E. Turkel, On accuracy conditions for the numerical computation of waves, *J. Comp. Physics* 59 (1985), 396–404.

[20] A. Bayliss, M. Gunzburger, and E. Turkel, Boundary conditions for the numerical solution of elliptic equations in exterior regions, *SIAM J. Appl. Math.*, Vol. 42, 2 (1982) 430–451.

[21] A. Bedford and D.S. Drumheller, *Elastic Wave Propagation*, J. Wiley 1994.

[22] J. Berenger, A perfectly matched layer for the absorption of electromagnetic waves, *J. Comput. Physics* 114 (1994), 185–200.

[23] J. Berenger, Perfectly matched layer for the FDTD solution of wave-structure interaction problems, *IEEE Trans. Antennas Propagat.* 44 (1996), 110–117.

[24] P. Bettess and O.C. Zienkiewicz, Diffraction and refraction of surface waves using finite and infinite elements, *Int. J. Numer. Methods Eng.*, 11, 1271–1296 (1977).

[25] P. Bettess, *Infinite Elements*, Penshaw Press, Cleadon 1992.

[26] Ph. Bouillard, J.-F. Allard, and G. Warzee, Accuracy control for acoustic finite element analysis, *ACUSTICA* 1, 1996, p. S156.

[27] Ph. Bouillard, J.-F. Allard, and G. Warzee, Superconvergent patch recovery technique for the finite element method in acoustics, *Comm. Numer. Methods Eng.*, 12, 581–594 (1996).

[28] Ph. Bouillard and F. Ihlenburg, Error control for finite element solutions of Helmholtz equation. In: Proceedings, 15th IMACS World Congress 1997 on Scientific Computation, Modelling and Applied Mathematics.

[29] F. Bowman, *Introduction to Bessel Functions*, Dover N.Y. 1958.

[30] D. Braess, *Finite Elemente*, 2. Auflage, Springer Verlag Berlin 1997; English edition: *Finite Elements: Theory, Fast Solvers and Applications in Solid Mechanics*, Cambridge University Press, Cambridge, 1997.

[31] S.C. Brenner and L.R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.

[32] D. Burnett, A 3-D acoustic infinite element based on a generalized multipole expansion, *J. Acoust. Soc. Am.* 96 (1994), 5, 2798–2816.

[33] G.F. Carey and J.T. Oden, *Finite Elements : A Second Course*, Volume II, Prentice-Hall, 1983.

[34] P.G. Ciarlet, *The Finite Element Method for Elliptic Equations*, North Holland, Amsterdam, 1978.

[35] P.G. Ciarlet and J.L. Lions (eds.), *Handbook of Numerical Analysis*, Vol. II, North Holland, Amsterdam, 1991.

[36] Y.-C. Chang and L. Demkowicz, Vibrations of a spherical shell. Comparison of 3D elasticity and Kirchhoff shell theory results, *Comp. Assist. Mech. Eng. Sci.*, 187–206 (1995).

[37] Y.-C. Chang and L. Demkowicz, Scattering of a spherical shell. Comparison of 3D elasticity and Kirchhoff shell theory results, *Comp. Assist. Mech. Eng. Sci.*, 207–229 (1995).

[38] F. Collino and P. Monk, The perfectly matched layer in curvilinear coordinates, to appear in *SIAM J. Sci. Comp.*

[39] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, Springer-Verlag Berlin, Heidelberg, New York, 1992.

[40] M. Costabel, M. Dauge: Singularités des équations de Maxwell dans un polyèdre. *C. R. Acad. Sci. Paris Série I* **324**, 1997, 1005–1010.

[41] R. Dautray, and L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol. 1, Springer-Verlag, Berlin, 1990.

[42] L. Demkowicz, Asymptotic convergence in Finite and Boundary element methods — Part 1: Theoretical results, *Comput. Math. Appli.* Vol. 27, No. 12, 69–84.

[43] L. Demkowicz, Asymptotic convergence in Finite and Boundary element methods — Part 2: The LBB constant for Rigid and elastic scattering problems, *Comput. Math. Appli.* Vol. 27, No. 12, 69–84, Vol. 28, 6, 93–109 (1994).

[44] L. Demkowicz, W. Rachowicz, K. Banaś, and J. Kucwaj, 2D adaptive package (2DhpAP), Technical Report, Cracow, April 1992.

[45] L. Demkowicz and F. Ihlenburg, Analysis of a Coupled Finite–Infinite Element Method for Exterior Helmholtz Problems, TICAM Technical Report, 11-96.

[46] L. Demkowicz and K. Gerdes, Convergence of the infinite element methods for the Helmholtz equation, TICAM-Report 6/96, to appear in *Numer. Math.*.

[47] L. Demkowicz, A. Karafiat, and J.T. Oden, Solution of elastic scattering problems in linear acoustics using *hp* boundary element method, *Comp. Methods Appl. Mech. Eng.* 101 (1992) 251–282.

[48] L. Demkowicz and J.T. Oden, Recent Progress on application of *hp*-adaptive BE/FE methods to elastic scattering, *Int. J. Numer. Meth. Eng.*, 37, 2893–2910 (1994).

[49] L. Demkowicz and J.T. Oden, Application of *hp*-adaptive BE/FE methods to elastic scattering, TICAM Report 94–15, 1994.

[50] J. Douglas Jr., J.E. Santos, D. Sheen, L. Schreyer, Frequency domain treatment of one-dimensional scalar waves, *Mathematical Models and Methods in Applied Sciences*, Vol. 3, (1993) 171–194.

[51] J. Elschner and G. Schmidt, Diffraction in periodic structures and optimal design of binary gratings I. Direct problems and gradient formulas, to appear in *Math. Methods Appl. Sci.*

[52] B. Enquist and A. Majda, Absorbing boundary conditions for the numerical simulation of waves, *Math. Comp.* 31, No. 139, 629–651 (1977).

[53] B. Enquist and A. Majda, Radiation boundary conditions for acoustic and elastic wave calculations, *Comm. Pure Appl. Math.*, Vol. 32, 313–357, 1979.

[54] Feng Kang, Finite element method and natural boundary reduction, *Proceedings of the International Congress of Mathematicians*, Warsaw, 1983, 1439–1453.

[55] K. Gerdes, The conjugated vs. the unconjugated infinite element method for the Helmholtz equation in exterior domains, Research Report 96–11, Seminar Angewandte Mathematik, ETH Zürich, to appear in *Comp. Methods Appl. Mech. Eng.*

[56] K. Gerdes and L. Demkowicz, Solutions of 3D-Laplace and Helmholtz equations in exterior domains using hp infinite elements, *Comp. Methods Appl. Mech. Eng.* 137 (1996) 239–273.

[57] K. Gerdes and F. Ihlenburg, The pollution effect in FE solutions of the 3D Helmholtz equation with large wavenumber, to appear in *Comp. Methods Appl. Mech. Eng.*

[58] M.J. Grote and J.B. Keller, On nonreflecting boundary conditions, *J. Comp. Physics* 122 (1995), 231–243.

[59] D. Givoli, Non-reflecting boundary conditions, *J. Comp. Physics* 94 (1991), 1–29.

[60] D. Givoli, *Numerical Methods for Problems in Infinite Domains*, Elsevier, Amsterdam, 1992.

[61] D. Givoli, and J.B. Keller, Special finite elements for use with high-order boundary conditions, *Comp. Methods Appl. Mech. Eng.* 119 (1994), 199–213.

[62] W. Hackbusch, *Theorie und Numerik elliptischer Differentialgleichungen*, Teubner Verlag 1996; English edition: *Elliptic Partial Differential Equations, Theory and Numerical Treatment*, Springer-Verlag Berlin, Heidelberg, New York, 1992.

[63] I. Harari and T.J.R. Hughes, Finite element method for the Helmholtz equation in an exterior domain: model problems, *Comp. Methods Appl. Mech. Eng.* 87 (1991), 59–96.

[64] I. Harari and T.J.R. Hughes, A cost comparison of boundary element and finite element methods for problems of time-harmonic acoustics, *Comp. Methods Appl. Mech. Eng.* 97 (1992), 77–102.

[65] I. Harari and T.J.R. Hughes, Galerkin/least squares finite element methods for the reduced wave equation with non-reflecting boundary conditions in unbounded domains, *Comp. Methods Appl. Mech. Eng.* 98 (1992) 411–454.

[66] I. Harari and T.J.R. Hughes, Analysis of continuous formulations underlying the computation of time-harmonic acoustics in exterior domains, Comp. Methods Appl. Mech. Eng.., 97, 103–124 (1992).

[67] B. Heinrich, The Fourier FEM for Poisson's equation in axisymmetric domains with edges, *SIAM J. Numer. Anal.*, 33, 1885–1911, 1996.

[68] I. Herrera, *Boundary Methods: An Algebraic Theory*, Pitman, Boston, 1984.

[69] G.C. Hsiao, The coupling of boundary element and finite element methods, *ZAMM* 70 (1990) 6, T493–T503.

[70] G.C. Hsiao and R.E.Kleinmann, Error analysis in numerical solution of acoustic integral equations, *Int. J. Numer. Meth. Eng.*, 37, 2921–2933.

[71] T. J. R. Hughes, *The Finite Element Method*, Prentice Hall, Englewood Cliffs, 1987.

[72] F. Ihlenburg and I. Babuška, Finite element solution to the Helmholtz equation with high wave numbers, *Comput. Math. Appli.*, Vol. 30, No. 9, 9–37, 1995.

[73] F. Ihlenburg and I. Babuška, Finite element solution to the Helmholtz equation with high wave numbers - Part II: the h-p-version of the FEM, *SIAM J. Numer. Anal.*, Vol. 34, , No. 1, 315–358, 1997.

[74] F. Ihlenburg and I. Babuška, Dispersion analysis and error estimation of Galerkin finite element methods for the numerical computation of waves, *Int. J. Numer. Meth. Eng.*, 38 (1995) 3745–3774.

[75] F. Ihlenburg and Ch. Makridakis, Error estimates of Galerkin FEM for a system of coupled Helmholtz equations in one dimension. In: W. Hackbusch and G. Wittum (eds.), *Numerical treatment of coupled systems*, Vieweg, 1995, 96–105.

[76] F. Ihlenburg and J. Shirron,. Numerical and analytical resolution of different scales for efficient high fidelity simulation of acoustic scattering problems. To appear in *Proccedings of Kieler GAMM Symposium 1997*, Vieweg, 1998

[77] D.J. Inman, *Engineering Vibration*, Prentice Hall, Englewood Cliffs, 1996.

[78] C. Johnson and J.C. Nedelec, On the coupling of boundary integral and finite element methods, *Math. Comp.* 35, 152, 1063–1079, 1980.

[79] F. John, *Partial Differential Equations*, Fourth ed., Springer New York, 1982.

[80] W.B. Jones, W.J. Thron, *Continued Fractions, Analytic Theory and Applications*, Addison-Wesley, Reading, Mass., 1980.

[81] M.C. Junger and D. Feit, *Sound, Structures and Their Interaction*, 2nd ed., MIT Press, Cambridge MA, 1986.

[82] A. Karp, A convergent far field expansion for two-dimensional radiation functions, *Comm. Pure Appl. Math.* 14, 427 (1961).

[83] J.B. Keller and D. Givoli, Exact non-reflecting boundary conditions, *J. Comp. Physics* 82 (1989) 172–192.

[84] F. Klein, *Vorlesungen über die Entwicklung der Mathematik im 19. Jahrhundert*, Teil 1, R. Courant and O. Neugebauer (eds.), Springer-Verlag, Berlin, 1926.

[85] R. Kress, *Integral Equations*, Springer-Verlag Berlin, Heidelberg, New York, 1990.

[86] L.D. Landau and E.M. Lifshitz, *Fluid Mechanics*, Pergamon Press, Oxford, 1987.

[87] R. Leis, *Initial Boundary Value Problems in Mathematical Physics*, J. Wiley & Teubner Verlag, Stuttgart, 1986.

[88] J. Mackerle, Finite element and boundary element techniques in acoustics — a bibliography (1990–92), *Finite Elem. Anal. Design*, 15 (1994), 263–272.

[89] M. Malhotra, *Iterative solution methods for large-scale finite element models in structural acoustics*, SUDAM report No. 92-2 (Ph.D. dissertation), Div. of Applied Mechanics, Stanford University, 1996.

[90] Ch. Makridakis, F. Ihlenburg, and I. Babuška, Analysis and finite element methods for a fluid–solid interaction problem in one dimension, *Math. Methods and Models in Appl. Sciences*, Vol. 6, No. 8 (1996), 1119–1141.

[91] J.M. Melenk, *On Generalized Finite Element Methods*, Ph.D. thesis, University of Maryland at College Park, 1995.

[92] J.M. Melenk and I. Babuška, The partition of unity method: basic theory and applications, *Comp. Methods Appl. Mech. Eng.*, 139, 289–314, 1996.

[93] J.M. Melenk and I. Babuška, Approximation with harmonic and generalized harmonic polynomials in the partition of unity method, *Comp. Assist. Mech. Eng. Sci.*, Vol. 4, 607–632, 1997.

[94] S.G. Michlin and K.L. Smolickij, *Approximate Methods for Solution of Differential an Integral Equations*, Elsevier, 1967.

[95] P.M. Morse, H. Feshbach, *Methods of Theoretical Physics*, McGraw–Hill Book Comp. New York, 1953.

[96] C. Müller, *Foundations of the Mathematical Theory of Electromagnetic Waves*, Springer Verlag, 1969.

[97] A. Nachman, A brief perspective on computational electromagnetics, *J. Comp. Physics* 126, 237–239 (1996).

[98] J. Nečas, *Les methodes directes en theorie des equations elliptiques*, Prague, Academia, 1967.

[99] S. Nettel, *Wave Physics, Oscillations- Solitons-Chaos*, second ed., Springer-Verlag, 1995.

[100] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes in FORTRAN*, second edition, Cambridge University Press, 1992.

[101] H. J.P. Morand and R. Ohayon, *Fluid–Structure Interaction*, J. Wiley, New York, 1995.

[102] J. T. Oden and L.F. Demkowicz, *Applied Functional Analysis*, CRC Press, 1996.

[103] F.W.J. Olver, *Asymptotics and Special Functions*, Academic Press, 1974.

[104] S.P. Parker (ed.), *Acoustics Source Book*, McGraw–Hill 1987.

[105] F. Riesz and B. Sz.-Nagy, *Functional Analysis*, Dover Publications, New York, 1990.

[106] A.A. Samarskii, *Introduction to the theory of difference schemes*, [Russian], Moscow, Nauka edition, 1971.

[107] J. Sanchez Hubert and E. Sanchez Palencia, *Vibration and Coupling of Continuous Systems. Asymptotic Methods*, Springer Verlag, Berlin, Heidelberg, 1989.

[108] A. Schatz, An observation concerning Ritz–Galerkin methods with indefinite bilinear forms, *Math. Comp.* 28 (1974), 959–962.

[109] A.H. Schatz, An analysis of the finite element method for second order elliptic boundary value problems. In: A.H. Schatz, V.T. Thomée and W.L. Wendland, *Mathematical Theory of Finite and Boundary Element Methods*, Birkhäuser-Verlag, Basel, 1990.

[110] J. Shirron, *Numerical Solution of Exterior Helmholtz Problems Using Finite and Infinite Elements*, Ph.D. thesis, College Park, MD, 1995.

[111] G. Strang and G.J. Fix, *An Analysis of the Finite Element Method*, Prentice Hall, Englewood Cliffs, NJ, 1973.

[112] B.S. Szabó and I. Babuška, *Finite Element Analysis*, J. Wiley, 1991.

[113] L.L. Thompson and P.M. Pinsky, Complex wavenumber Fourier analysis of the p-version finite element method, *Comp. Mech.*, 13 (1994), 255 –275.

[114] L.L. Thompson and P.M. Pinsky, A Galerkin Least Squares Finite Element Method for the Two-Dimensional Helmholtz Equation, *Int. J. Numer. Meth. Eng.*, 38 (1995), 371–397.

[115] V.V. Varadan, A. Lakhtakia, and V.K. Varadan, *Field Representations and Introduction to Scattering*, Vol. 1, North Holland 1991.

[116] R. Verfürth, *A Review of* aposteriori *Error Estimation and Adaptive Mesh Refinement Techniques*, J. Wiley and Teubner-Verlag, 1996.

[117] L.B. Wahlbin, Local behavior in finite element methods. In: P.G. Ciarlet and J.L. Lions, eds., *Handbook of Numerical Analysis, Vol. II*, Elsevier, 353–522.

[118] W.L. Wendland, On asymptotic error estimates for the combined BEM and FEM. In: Stein, E., Wendland, W. (eds.) *Finite element and boundary element techniques from mathematical and engineering point of view*, CISM Lecture Notes 301, Springer-Verlag 1988, 273–333.

[119] C. H. Wilcox, An expansion theorem for electromagnetic fields, *Comm. Pure Appl. Math.* 9, 115–134 (1956).

[120] K. Yosida, *Functional Analysis*, Springer-Verlag, Berlin, 1980.

[121] E. Zeidler (ed.) *Teubner Taschenbuch der Mathematik*, Teubner Verlag, Leipzig, 1996.

[122] O.C. Zienkiewicz and J.Z. Zhu, The superconvergent patch recovery and a posteriori error estimates. Part 2: Error estimates and adaptivity, *Int. J. Numer. Meth. Eng.*, 33, 1365–1382 (1992).

Index