Scientific Computation

Editorial Board

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

Zhangxin Chen

Finite Element Methods and Their Applications

With 117 Figures

Springer

Prof. Zhangxin Chen
Department of Mathematics
Box 750156
Southern Methodist University
Dallas, TX 75275-0156, USA
zchen@mail.smu.edu

Library of Congress Control Number: 2005926238

ISBN-10 3-540-24078-0 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-24078-5 Springer Berlin Heidelberg New York

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

Springer is a part of Springer Science+Business Media
springeronline.com

© Springer-Verlag Berlin Heidelberg 2005
Printed in Germany

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

Typesetting: Data prepared by the authors using a Springer TEX macro package
Production: LE-TEX Jelonek, Schmidt & Vöckler GbR, Leipzig
Cover design: *design & production* GmbH, Heidelberg

Printed on acid-free paper 55/3141/JVG 5 4 3 2 1 0

Preface

The finite element method is one of the major tools used in the numerical solution of partial differential equations. This book offers a fundamental and practical introduction to the method, its variants, and their applications. In presenting the material, I have attempted to introduce every concept in the simplest possible setting and to maintain a level of treatment that is as rigorous as possible without being unnecessarily abstract.

The book is based on the material that I have used in a graduate course at Southern Methodist University for several years. Part of the material was also used for my seminar notes at Purdue University, University of Minnesota, and Texas A&M University. Furthermore, this book was the basis for summer schools on the finite element method and its applications held in China, Iran, Mexico, and Venezuela.

This book covers six major topics and four applications. In Chap. 1, the standard (H^1- and H^2-conforming) finite element method is introduced. In Chaps. 2 and 3, two closely related finite element methods, the nonconforming and the mixed finite element methods, are discussed. The discontinuous and characteristic finite element methods are studied in Chaps. 4 and 5; these two methods have been recently developed. The adaptive finite element method is considered in Chap. 6. The last four chapters are devoted to applications of these methods to solid mechanics (Chap. 7), fluid mechanics (Chap. 8), fluid flow in porous media (Chap. 9), and semiconductor modeling (Chap. 10). In each chapter, a brief introduction, the notation, a basic terminology, and necessary concepts are given. Theoretical considerations and bibliographical information are also presented at the end of each chapter. The reader who is not interested in the theory may skip them. Each of the three main types of partial differential equations, i.e., elliptic, parabolic, and hyperbolic equations, is treated in this book. Nonlinear problems are studied as well.

In Chap. 1, we describe the finite element method. We first introduce this method for two simple model problems in Sect. 1.1. Then, in Sect. 1.2, we discuss the small fraction of Sobolev space theory that is sufficient for the foundation of the finite element method as studied in this book. In Sect. 1.3, we develop an abstract variational formulation for this method and give some examples. Section 1.4 is devoted to the construction of general finite element spaces. In Sects. 1.1 and 1.4, we concentrate on polygonal domains; curved

domains are treated in Sect. 1.5. In Sect. 1.6, we very briefly touch on the topic of numerical integration. The finite element method is extended to transient and nonlinear problems in Sects. 1.7 and 1.8, respectively. Section 1.9 is devoted to theoretical consideration of the finite element method; in particular, an approximation theory for the finite element method is established. For self-containedness, in Sect. 1.10, we briefly discuss solution techniques for solving the linear systems arising from the finite element method; these techniques are needed to complete some of the exercises given in Sect. 1.12. For those who have had a course in numerical linear algebra, this section can be skipped. Bibliographical information is given in Sect. 1.11.

In Chap. 2, we discuss the application of the nonconforming finite element method to second- and fourth-order partial differential equation problems (cf. Sects. 2.1 and 2.2). In particular, the nonconforming P_1 (i.e, Crouzeix-Raviart), rotated Q_1, Wilson, Morley, Fraeijs de Veubeke, Zienkiewicz, and Adini elements are described. In Sect. 2.3, we briefly present an application of this method to a nonlinear transient problem.

In Chap. 3, we study the mixed finite element method. As an introduction, in Sect. 3.1, we first describe this method for a one-dimensional model problem. Then we generalize it to a two-dimensional model problem in Sect. 3.2. In Sect. 3.3, we consider the method for general boundary conditions. In Sect. 3.4, we present various mixed finite element spaces, and, in Sect. 3.5, we state the approximation properties of these spaces. In Sect. 3.6, we briefly present an application of the mixed method to a nonlinear transient problem. We also discuss solution techniques for solving the linear algebraic systems arising from this method in Sect. 3.7. These techniques include the classical Uzawa, minimum residual iterative, alternating direction iterative, mixed-hybrid, and equivalence-to-nonconforming algorithms. Here the mixed method is developed in a simple setting. The book by Brezzi-Fortin (1991) should be consulted for a thorough treatment of the subject.

In Chap. 4, we first study the discontinuous Galerkin (DG) finite element method and its stabilized versions for advection problems (cf. Sect. 4.1). Then, in Sect. 4.2, we show how to extend these methods to diffusion problems. In Sect. 4.3, we discuss the recently developed mixed discontinuous finite element method.

In Chap. 5, we discuss the modified method of characteristics (cf. Sect. 5.2), the Eulerian-Lagrangian method (cf. Sect. 5.3), the characteristic mixed method (cf. Sect. 5.4), and the Eulerian-Lagrangian mixed discontinuous method (cf. Sect. 5.5). In Sect. 5.6, we consider the application of these methods to nonlinear problems. In Sect. 5.7, we comment on the characteristic finite element method.

In Chap. 6, we present a brief introduction of some of basic topics on the two components for the adaptive finite element method: the adaptive strategy and a-posteriori error estimation. In Sect. 6.1, we introduce the concept of local grid refinement in space. In Sect. 6.2, we briefly discuss a data

structure which efficiently supports adaptive refinement and unrefinement. In Sect. 6.3, we discuss a-posteriori error estimates for stationary problems, and, in Sect. 6.4, extend them to transient problems. In Sect. 6.5, we briefly consider their application to nonlinear problems.

In Chap. 7, we introduce linear elasticity (cf. Sect. 7.1). In Sect. 7.2, we state variational formulations of the governing equations. Then, in Sect. 7.3, we describe the H^1-conforming, mixed, and nonconforming finite element methods for the discretization of these equations.

In Chap. 8, we describe the derivation of the Stokes and Navier-Stokes equations from the fundamental principles of classical mechanics governing the motion of a continuous medium (cf. Sect. 8.1). In Sect. 8.2, we introduce variational formulations of the Stokes equation. Then, in Sect. 8.3, we apply the H^1-conforming, mixed, and nonconforming finite element methods for the numerical solution of this equation. In Sect. 8.4, we remark on an extension to the Navier-Stokes equation.

In Chap. 9, we study two-phase flow in a porous medium. In Sect. 9.1, we state the governing equations for two-phase flow and their variants defined in terms of pressure and saturation. In Sect. 9.2, we use the mixed finite element method for the pressure solution. Then, in Sect. 9.3, we employ the characteristic finite element method for the saturation solution. In Sect. 9.4, we present a numerical example.

In Chap. 10, we introduce the drift-diffusion, (classical) hydrodynamic, and quantum hydrodynamic models in semiconductor modeling (cf. Sect. 10.1) and finite element methods for solving these models (cf. Sect. 10.2). In Sect. 10.3, we present a numerical example using the hydrodynamic model.

This book can serve as a course that provides an introduction to numerical methods for partial differential equations for graduate students. Some elementary chapters, such as the first three chapters, can be even taught at undergraduate level. It can be also used as a reference book for mathematicians, engineers, and scientists interested in numerical solutions. The necessary prerequisites are relatively moderate: a basic course in advanced calculus and some acquaintance with partial differential equations. For the theoretical considerations in this book, some acquaintance with functional analysis is needed.

Chapters 1 through 6 form the essential material for a course. Because each of Chaps. 2 through 6 is essentially self-contained and independent, different course paths can be chosen. The problem section in each chapter plays a role in the presentation, and the reader should spend the time to solve the problems.

I take this opportunity to thank many people who have helped, in different ways, in the preparation of this book. During my graduate study and post-doctoral research, I had incredibly supportive supervision by Professors Bernardo Cockburn, Jim Douglas, Jr., Richard E. Ewing, and Kaitai Li. Many of my students made invaluable comments at the early stages of this

book. I would also like to thank Professor Ian Gladwell for reading the whole manuscript and making invaluable suggestions.

Dallas, Texas, USA *Zhangxin Chen*
March 2005

Contents

1 Elementary Finite Elements 1
1.1 Introduction 2
1.1.1 A One-Dimensional Model Problem 2
1.1.2 A Two-Dimensional Model Problem 9
1.1.3 An Extension to General Boundary Conditions 14
1.1.4 Programming Considerations 16
1.2 Sobolev Spaces 19
1.2.1 Lebesgue Spaces 20
1.2.2 Weak Derivatives 21
1.2.3 Sobolev Spaces 22
1.2.4 Poincaré's Inequality 23
1.2.5 Duality and Negative Norms 25
1.3 Abstract Variational Formulation 26
1.3.1 An Abstract Formulation 26
1.3.2 The Finite Element Method 28
1.3.3 Examples 30
1.4 Finite Element Spaces 35
1.4.1 Triangles 35
1.4.2 Rectangles 40
1.4.3 Three Dimensions 42
1.4.4 A C^1 Element 44
1.5 General Domains 46
1.6 Quadrature Rules 49
1.7 Finite Elements for Transient Problems 50
1.7.1 A One-Dimensional Model Problem 51
1.7.2 A Semi-Discrete Scheme in Space 52
1.7.3 Fully Discrete Schemes 55
1.8 Finite Elements for Nonlinear Problems 58
1.8.1 Linearization Approaches 59
1.8.2 Implicit Time Approximations 60
1.8.3 Explicit Time Approximations 61
1.9 Approximation Theory 62
1.9.1 Interpolation Errors 62
1.9.2 Error Estimates for Elliptic Problems 67

1.9.3 L^2-Error Estimates 68
1.10 Linear System Solution Techniques 70
1.10.1 Gaussian Elimination 70
1.10.2 The Conjugate Gradient Algorithm 76
1.11 Bibliographical Remarks 81
1.12 Exercises 81

2 Nonconforming Finite Elements 87
2.1 Second-Order Problems 87
2.1.1 Nonconforming Finite Elements on Triangles 89
2.1.2 Nonconforming Finite Elements on Rectangles 92
2.1.3 Nonconforming Finite Elements on Tetrahedra 95
2.1.4 Nonconforming Finite Elements on Parallelepipeds . . . 95
2.1.5 Nonconforming Finite Elements on Prisms 97
2.2 Fourth-Order Problems 98
2.2.1 The Morley Element 100
2.2.2 The Fraeijs de Veubeke Element 102
2.2.3 The Zienkiewicz Element 103
2.2.4 The Adini Element 104
2.3 Nonlinear Problems 105
2.4 Theoretical Considerations 106
2.4.1 An Abstract Formulation 106
2.4.2 Applications 109
2.5 Bibliographical Remarks 113
2.6 Exercises 113

3 Mixed Finite Elements 117
3.1 A One-Dimensional Model Problem 118
3.2 A Two-Dimensional Model Problem 123
3.3 Extension to Boundary Conditions of Other Types 126
3.3.1 A Neumann Boundary Condition 126
3.3.2 A Boundary Condition of Third Type 128
3.4 Mixed Finite Element Spaces 128
3.4.1 Mixed Finite Element Spaces on Triangles 130
3.4.2 Mixed Finite Element Spaces on Rectangles 133
3.4.3 Mixed Finite Element Spaces on Tetrahedra 136
3.4.4 Mixed Finite Element Spaces on Parallelepipeds 137
3.4.5 Mixed Finite Element Spaces on Prisms 140
3.5 Approximation Properties 143
3.6 Mixed Methods for Nonlinear Problems 143
3.7 Linear System Solution Techniques 145
3.7.1 Introduction 145
3.7.2 The Uzawa Algorithm 146
3.7.3 The Minimum Residual Iterative Algorithm 147
3.7.4 Alternating Direction Iterative Algorithms 148

3.7.5 Mixed-Hybrid Algorithms . . . 150
3.7.6 An Equivalence Relationship . . . 152
3.8 Theoretical Considerations . . . 154
3.8.1 An Abstract Formulation . . . 154
3.8.2 The Mixed Finite Element Method . . . 158
3.8.3 Examples . . . 161
3.8.4 Construction of Projection Operators . . . 162
3.8.5 Error Estimates . . . 164
3.9 Bibliographical Remarks . . . 166
3.10 Exercises . . . 167

4 Discontinuous Finite Elements . . . 173
4.1 Advection Problems . . . 173
4.1.1 DG Methods . . . 174
4.1.2 Stabilized DG Methods . . . 178
4.2 Diffusion Problems . . . 183
4.2.1 Symmetric DG Method . . . 186
4.2.2 Symmetric Interior Penalty DG Method . . . 187
4.2.3 Non-Symmetric DG Method . . . 188
4.2.4 Non-Symmetric Interior Penalty DG Method . . . 189
4.2.5 Remarks . . . 192
4.3 Mixed Discontinuous Finite Elements . . . 194
4.3.1 A One-Dimensional Problem . . . 194
4.3.2 Multi-Dimensional Problems . . . 203
4.3.3 Nonlinear Problems . . . 206
4.4 Theoretical Considerations . . . 208
4.4.1 DG Methods . . . 208
4.4.2 Stabilized DG Methods . . . 210
4.5 Bibliographical Remarks . . . 212
4.6 Exercises . . . 212

5 Characteristic Finite Elements . . . 215
5.1 An Example . . . 216
5.2 The Modified Method of Characteristics . . . 218
5.2.1 A One-Dimensional Model Problem . . . 218
5.2.2 Periodic Boundary Conditions . . . 222
5.2.3 Extension to Multi-Dimensional Problems . . . 222
5.2.4 Discussion of a Conservation Relation . . . 224
5.3 The Eulerian-Lagrangian Localized Adjoint Method . . . 226
5.3.1 A One-Dimensional Model Problem . . . 226
5.3.2 Extension to Multi-Dimensional Problems . . . 236
5.4 The Characteristic Mixed Method . . . 242
5.5 The Eulerian-Lagrangian Mixed Discontinuous Method . . . 245
5.6 Nonlinear Problems . . . 248
5.7 Remarks on Characteristic Finite Elements . . . 250

5.8 Theoretical Considerations . . . 250
5.9 Bibliographical Remarks . . . 258
5.10 Exercises . . . 258

6 Adaptive Finite Elements . . . 261
6.1 Local Grid Refinement in Space . . . 262
6.1.1 Regular H-Schemes . . . 263
6.1.2 Irregular H-Schemes . . . 265
6.1.3 Unrefinements . . . 266
6.2 Data Structures . . . 267
6.3 A-Posteriori Error Estimates for Stationary Problems . . . 270
6.3.1 Residual Estimators . . . 271
6.3.2 Local Problem-Based Estimators . . . 277
6.3.3 Averaging-Based Estimators . . . 281
6.3.4 Hierarchical Basis Estimators . . . 283
6.3.5 Efficiency of Error Estimators . . . 287
6.4 A-Posteriori Error Estimates for Transient Problems . . . 289
6.5 A-Posteriori Error Estimates for Nonlinear Problems . . . 292
6.6 Theoretical Considerations . . . 293
6.6.1 An Abstract Theory . . . 294
6.6.2 Applications . . . 297
6.7 Bibliographical Remarks . . . 302
6.8 Exercises . . . 302

7 Solid Mechanics . . . 305
7.1 Introduction . . . 305
7.1.1 Kinematics . . . 305
7.1.2 Equilibrium . . . 306
7.1.3 Material Laws . . . 306
7.2 Variational Formulations . . . 308
7.2.1 The Displacement Form . . . 308
7.2.2 The Mixed Form . . . 309
7.3 Finite Element Methods . . . 310
7.3.1 Finite Elements and Locking Effects . . . 310
7.3.2 Mixed Finite Elements . . . 311
7.3.3 Nonconforming Finite Elements . . . 313
7.4 Theoretical Considerations . . . 314
7.5 Bibliographical Remarks . . . 319
7.6 Exercises . . . 319

8 Fluid Mechanics . . . 321
8.1 Introduction . . . 321
8.2 Variational Formulations . . . 323
8.2.1 The Galerkin Approach . . . 323
8.2.2 The Mixed Formulation . . . 324

8.3 Finite Element Methods . . . 324
8.3.1 Galerkin Finite Elements . . . 324
8.3.2 Mixed Finite Elements . . . 325
8.3.3 Nonconforming Finite Elements . . . 326
8.4 The Navier-Stokes Equation . . . 329
8.5 Theoretical Considerations . . . 330
8.6 Bibliographical Remarks . . . 333
8.7 Exercises . . . 333

9 Fluid Flow in Porous Media . . . 337
9.1 Two-Phase Immiscible Flow . . . 338
9.1.1 The Phase Formulation . . . 340
9.1.2 The Weighted Formulation . . . 342
9.1.3 The Global Formulation . . . 342
9.2 Mixed Finite Elements for Pressure . . . 343
9.3 Characteristic Methods for Saturation . . . 345
9.4 A Numerical Example . . . 346
9.5 Theoretical Considerations . . . 349
9.5.1 Analysis for the Pressure Equation . . . 349
9.5.2 Analysis for the Saturation Equation . . . 351
9.6 Bibliographical Remarks . . . 361
9.7 Exercises . . . 362

10 Semiconductor Modeling . . . 363
10.1 Three Semiconductor Models . . . 364
10.1.1 The Drift-Diffusion Model . . . 364
10.1.2 The Hydrodynamic Model . . . 366
10.1.3 The Quantum Hydrodynamic Model . . . 367
10.2 Numerical Methods . . . 368
10.2.1 The Drift-Diffusion Model . . . 368
10.2.2 The Hydrodynamic Model . . . 371
10.2.3 The Quantum Hydrodynamic Model . . . 378
10.3 A Numerical Example . . . 379
10.4 Bibliographical Remarks . . . 384
10.5 Exercises . . . 384

A Nomenclature . . . 385

References . . . 391

Index . . . 405

1 Elementary Finite Elements

A numerical approach for solving a differential equation problem is to discretize this problem, which has infinitely many degrees of freedom, to produce a discrete problem, which has finitely many degrees of freedom and can be solved using a computer. Compared with the classical finite difference method, the introduction of the finite element method is relatively recent. The advantages of the finite element method over the finite difference method are that general boundary conditions, complex geometry, and variable material properties can be relatively easily handled. Also, the clear structure and versatility of the finite element method makes it possible to develop general purpose software for applications. Furthermore, it has a solid theoretical foundation that gives added reliability, and in many situations it is possible to obtain concrete error estimates in finite element solutions. The finite element method was first introduced by Courant in 1943 (Courant, 1943). From the 1950's to the 1970's, it was developed by engineers and mathematicians into a general method for the numerical solution of partial differential equations.

In this chapter, we describe the finite element method. We first introduce this method for two simple model problems in Sect. 1.1. Then, in Sect. 1.2, we discuss the small fraction of Sobolev space theory that is sufficient for the foundation of the finite element method as studied in this book. In Sect. 1.3, we develop an abstract variational formulation for this method and give some examples. Section 1.4 is devoted to the construction of general finite element spaces. In Sects. 1.1 and 1.4, we concentrate on polygonal domains; curved domains are treated in Sect. 1.5. In Sect. 1.6, we briefly touch on the topic of numerical integration. The finite element method is extended to transient and nonlinear problems in Sects. 1.7 and 1.8, respectively. Section 1.9 is devoted to theoretical considerations of the finite element method; in particular, an approximation theory for the finite element method is established. The reader who is not interested in the theory may simply skip this section. For self-containedness, in Sect. 1.10, we briefly discuss solution techniques for solving the linear systems arising from the finite element method; these techniques are needed to complete some of the exercises given in Sect. 1.12. For those who have had a course in numerical linear algebra, this section can be skipped. Finally, bibliographical information is given in Sect. 1.11. Students

are encouraged to do some of the exercises given in Sect. 1.12; these exercises are closely related to the material presented in this chapter.

1.1 Introduction

The exposition in this section has two purposes: (1) to introduce the terminology and (2) to summarize the basic ingredients that are required for the development of the finite element method.

1.1.1 A One-Dimensional Model Problem

As an introduction, we consider a stationary problem in one dimension

$$\begin{aligned} -\frac{d^2p}{dx^2} &= f(x), \qquad 0 < x < 1 \ , \\ p(0) &= p(1) = 0 \ , \end{aligned} \tag{1.1}$$

where f is a given real-valued piecewise continuous bounded function. Note that (1.1) is a two-point boundary value problem. A number of problems in physics and mechanics arise in form (1.1). For example, consider an elastic bar with tension one, fixed at both ends ($x = 0,\ 1$) and subject to a transversal load of intensity f (cf. Fig. 1.1). Under the assumption of a small displacement, the transversal displacement p satisfies problem (1.1) (cf. Exercise 1.1).

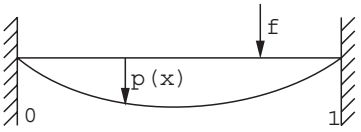

Fig. 1.1. An elastic bar

The finite difference method for (1.1) is to replace the second derivative by a difference quotient that involves the values of p at certain points. The discretization of (1.1) using the finite element method is different. This method starts by rewriting (1.1) in an equivalent variational formulation. For this, we introduce the *scalar-product* notation

$$(v, w) = \int_0^1 v(x)w(x)\ dx \ ,$$

for real-valued piecewise continuous bounded functions v and w, and we define the *linear space*

$$V = \left\{ v : v \text{ is a continuous function on } [0,1], \frac{dv}{dx} \text{ is piecewise continuous and bounded on } (0,1), \text{ and } v(0) = v(1) = 0 \right\}.$$

This space is a subspace of a *Sobolev space* introduced in the next section. We also define the *functional* $F : V \to \mathbb{R}$ by

$$F(v) = \frac{1}{2}\left(\frac{dv}{dx}, \frac{dv}{dx}\right) - (f, v), \qquad v \in V ,$$

where $\mathbb{R}$ is the set of real numbers. It will be shown at the end of this subsection that (1.1) is equivalent to the *minimization problem*

$$\text{Find } p \in V \text{ such that } F(p) \le F(v) \qquad \forall v \in V. \tag{1.2}$$

Problem (1.2) is called a *Ritz variational form* of (1.1).

In mechanics, for example, the quantity $\frac{1}{2}(\frac{dv}{dx}, \frac{dv}{dx})$ is the internal elastic energy, (f, v) is the load potential, and the functional value $F(v)$ represents the total potential energy associated with the displacement $v \in V$. Therefore, problem (1.2) corresponds to the *fundamental principle of minimum potential energy* in mechanics.

In terms of computations, (1.1) can be expressed in a more useful, direct formulation. Multiplying the first equation of (1.1) by any $v \in V$, called a *test function*, and integrating over $(0,1)$, we see that

$$-\left(\frac{d^2p}{dx^2}, v\right) = (f, v) .$$

Application of integration by parts to this equation yields

$$\left(\frac{dp}{dx}, \frac{dv}{dx}\right) = (f, v) , \tag{1.3}$$

where we use the fact that $v(0) = v(1) = 0$ from the definition of V. Equation (1.3) is called a *Galerkin variational* or *weak form* of (1.1). It corresponds to the *principle of virtual work* in mechanics, for example. If p is a solution to (1.1), then it satisfies (1.3). The converse also holds if d^2p/dx^2 exists and is piecewise continuous and bounded in $(0,1)$, for example; see Exercise 1.2. It can be seen that (1.2) and (1.3) are equivalent (see the end of this subsection).

We now construct the finite element method for solving (1.1). Toward that end, for a positive integer M, let $0 = x_0 < x_1 < \ldots < x_M < x_{M+1} = 1$ be a *partition* of $(0,1)$ into a set of subintervals $I_i = (x_{i-1}, x_i)$, with length $h_i = x_i - x_{i-1}$, $i = 1, 2, \ldots, M+1$. Set $h = \max\{h_i : i = 1, 2, \ldots, M+1\}$. The *step size* h measures how fine the partition is. Define the *finite element space*

$$V_h = \{v : v \text{ is a continuous function on } [0,1], v \text{ is linear}$$
$$\text{on each subinterval } I_i \ , \text{ and } v(0) = v(1) = 0\} \ .$$

See Fig. 1.2 for an illustration of a function $v \in V_h$. Note that $V_h \subset V$ (i.e., V_h is a subspace of V).

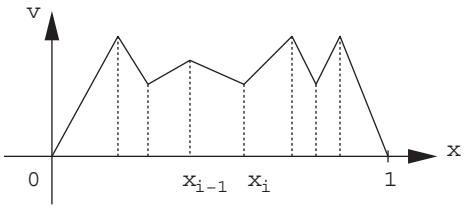

Fig. 1.2. An illustration of a function $v \in V_h$

The discrete version of (1.2) is

$$\text{Find } p_h \in V_h \text{ such that } F(p_h) \le F(v) \qquad \forall v \in V_h \ . \tag{1.4}$$

Method (1.4) is referred to as the *Ritz finite element method*. In the same manner as for (1.3) (see the end of this subsection), (1.4) is equivalent to the problem:

$$\text{Find } p_h \in V_h \text{ such that } \left(\frac{dp_h}{dx}, \frac{dv}{dx}\right) = (f, v) \qquad \forall v \in V_h \ . \tag{1.5}$$

This is usually termed the *Galerkin finite element method*.

It is easy to see that (1.5) has a unique solution. In fact, let $f = 0$, and take $v = p_h$ in (1.5) to give

$$\left(\frac{dp_h}{dx}, \frac{dp_h}{dx}\right) = 0 \ ,$$

so p_h is a constant. It follows from the boundary condition in V_h that $p_h = 0$.

We introduce the *basis functions* $\varphi_i \in V_h$, $i = 1, 2, \ldots, M$,

$$\varphi_i(x_j) = \begin{cases} 1 & \text{if } i = j \ , \\ 0 & \text{if } i \ne j \ . \end{cases}$$

That is, φ_i is a continuous piecewise linear function on $[0,1]$ such that its value is one at *node* x_i and zero at other nodes (cf. Fig. 1.3). It is called a *hat* or *chapeau* function.

Any function $v \in V_h$ has the unique representation

$$v(x) = \sum_{i=1}^{M} v_i \varphi_i(x), \qquad 0 \le x \le 1 \ ,$$

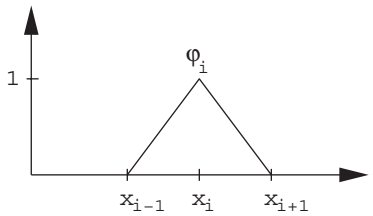

Fig. 1.3. A basis function in one dimension

where $v_i = v(x_i)$. For each j, take $v = \varphi_j$ in (1.5) to see that

$$\left(\frac{dp_h}{dx}, \frac{d\varphi_j}{dx}\right) = (f, \varphi_j)\,, \qquad j = 1, 2, \ldots, M\,. \tag{1.6}$$

Set

$$p_h(x) = \sum_{i=1}^{M} p_i \varphi_i(x), \qquad p_i = p_h(x_i)\,,$$

and substitute it into (1.6) to give

$$\sum_{i=1}^{M} \left(\frac{d\varphi_i}{dx}, \frac{d\varphi_j}{dx}\right) p_i = (f, \varphi_j)\,, \qquad j = 1, 2, \ldots, M\,. \tag{1.7}$$

This is a linear system of M algebraic equations in the M unknowns p_1, $p_2, \ldots, p_M$. It can be written in matrix form

$$\mathbf{A}\mathbf{p} = \mathbf{f}\,, \tag{1.8}$$

where the matrix $\mathbf{A}$ and vectors $\mathbf{p}$ and $\mathbf{f}$ are given by

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1M} \\ a_{21} & a_{22} & \ldots & a_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \ldots & a_{MM} \end{pmatrix}, \quad \mathbf{p} = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_M \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_M \end{pmatrix},$$

with

$$a_{ij} = \left(\frac{d\varphi_i}{dx}, \frac{d\varphi_j}{dx}\right), \quad f_j = (f, \varphi_j)\,, \qquad i, j = 1, 2, \ldots, M\,.$$

The matrix $\mathbf{A}$ is referred to as the *stiffness matrix* and $\mathbf{f}$ is the *load vector*.

By the definition of the basis functions, we observe that

$$\left(\frac{d\varphi_i}{dx}, \frac{d\varphi_j}{dx}\right) = 0 \quad \text{if } |i - j| \geq 2\,,$$

so $\mathbf{A}$ is *tridiagonal*; i.e., only the entries on the main diagonal and the adjacent diagonals may be nonzero. In fact, the entries a_{ij} can be calculated:

$$a_{ii} = \frac{1}{h_i} + \frac{1}{h_{i+1}}, \quad a_{i-1,i} = -\frac{1}{h_i}, \quad a_{i,i+1} = -\frac{1}{h_{i+1}} \ .$$

Also, it can be seen that $\mathbf{A}$ is *symmetric*: $a_{ij} = a_{ji}$, and *positive definite*:

$$\boldsymbol{\eta}^T \mathbf{A} \boldsymbol{\eta} = \sum_{i,j=1}^{M} \eta_i a_{ij} \eta_j > 0 \qquad \text{for all nonzero } \boldsymbol{\eta} \in \mathbb{R}^M \ ,$$

where $\boldsymbol{\eta}^T$ denotes the transpose of $\boldsymbol{\eta}$. Because a positive definite matrix is nonsingular, the linear system (1.8) has a unique solution. Consequently, we have shown that (1.5) has a unique solution $p_h \in V_h$ in a different way.

The symmetry of $\mathbf{A}$ can be seen from the definition of a_{ij}. The positive definiteness can be checked as follows: With

$$\eta = \sum_{i=1}^{M} \eta_i \varphi_i \in V_h, \quad \boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_M) \ ,$$

we see that

$$\begin{aligned} \sum_{i,j=1}^{M} \eta_i a_{ij} \eta_j &= \sum_{i,j=1}^{M} \eta_i \left(\frac{d\varphi_i}{dx}, \frac{d\varphi_j}{dx} \right) \eta_j \\ &= \left(\sum_{i=1}^{M} \eta_i \frac{d\varphi_i}{dx}, \sum_{j=1}^{M} \eta_j \frac{d\varphi_j}{dx} \right) = \left(\frac{d\eta}{dx}, \frac{d\eta}{dx} \right) \geq 0 \ , \end{aligned}$$

so, as for (1.5), the equality holds only for $\eta \equiv 0$ since a constant function η must be zero because of the boundary condition.

We remark that $\mathbf{A}$ is *sparse*; that is, only a few entries in each row of $\mathbf{A}$ are nonzero. In the present one-dimensional case, it is tridiagonal. The sparsity of $\mathbf{A}$ depends upon the fact that a basis function in V_h is different from zero only on a few intervals; that is, it has compact *support*. Thus it interferes only with a few other basis functions. That basis functions can be chosen in this manner is an important distinctive property of the finite element method.

In the case where the partition is uniform, i.e., $h = h_i$, $i = 1, 2, \ldots, M+1$, the stiffness matrix $\mathbf{A}$ takes the form

$$\mathbf{A} = \frac{1}{h} \begin{pmatrix} 2 & -1 & 0 & \ldots & 0 & 0 \\ -1 & 2 & -1 & \ldots & 0 & 0 \\ 0 & -1 & 2 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 2 & -1 \\ 0 & 0 & 0 & \ldots & -1 & 2 \end{pmatrix} .$$

With division by h in **A**, (1.5) can be thought of as a variant of the *central difference scheme* where the right-hand side consists of mean values of $f\varphi_j$ over the interval (x_{j-1}, x_{j+1}).

We end this subsection with two remarks. The first one is on an error estimate for (1.5). In general, the derivation of an *error estimate* for the finite element method is very technical. Here we briefly indicate how to obtain an estimate in one dimension. Subtract (1.5) from (1.3) to get

$$\left(\frac{dp}{dx} - \frac{dp_h}{dx}, \frac{dv}{dx}\right) = 0 \qquad \forall v \in V_h \ . \tag{1.9}$$

We introduce the notation

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

This is a *norm* associated with the *scalar product* $(\cdot, \cdot)$ (cf. Sect. 1.2). We use *Cauchy's inequality* (cf. Exercise 1.4)

$$|(v, w)| \leq \|v\| \, \|w\| \ . \tag{1.10}$$

Note that, using (1.9), with $v \in V_h$ we see that

$$\begin{aligned} \left\| \frac{dp}{dx} - \frac{dp_h}{dx} \right\|^2 &= \left(\frac{dp}{dx} - \frac{dp_h}{dx}, \frac{dp}{dx} - \frac{dp_h}{dx}\right) \\ &= \left(\frac{dp}{dx} - \frac{dp_h}{dx}, \left[\frac{dp}{dx} - \frac{dv}{dx}\right] + \left[\frac{dv}{dx} - \frac{dp_h}{dx}\right]\right) \\ &= \left(\frac{dp}{dx} - \frac{dp_h}{dx}, \frac{dp}{dx} - \frac{dv}{dx}\right) \ , \end{aligned}$$

so, by (1.10),

$$\left\| \frac{dp}{dx} - \frac{dp_h}{dx} \right\| \leq \left\| \frac{dp}{dx} - \frac{dv}{dx} \right\| \qquad \forall v \in V_h \ . \tag{1.11}$$

This equation implies that p_h is the best possible approximation of p in V_h in terms of the norm in (1.11).

To obtain an error bound, we take v in (1.11) to be the *interpolant* $\tilde{p}_h \in V_h$ of p; i.e., $\tilde{p}_h$ is defined by

$$\tilde{p}_h(x_i) = p(x_i), \qquad i = 0, 1, \ldots, M+1 \ . \tag{1.12}$$

It is an easy exercise (cf. Exercise 1.5) to see that, for $x \in [0, 1]$,

$$\begin{aligned} &|(p - \tilde{p}_h)(x)| \leq \frac{h^2}{8} \max_{y \in [0,1]} \left| \frac{d^2 p(y)}{dx^2} \right| , \\ &\left| \left(\frac{dp}{dx} - \frac{d\tilde{p}_h}{dx} \right)(x) \right| \leq h \max_{y \in [0,1]} \left| \frac{d^2 p(y)}{dx^2} \right| \ . \end{aligned} \tag{1.13}$$

With $v = \tilde{p}_h$ in (1.11) and the second equation of (1.13), we obtain

$$\left\| \frac{dp}{dx} - \frac{dp_h}{dx} \right\| \leq h \max_{y\in[0,1]} \left| \frac{d^2p(y)}{dx^2} \right| . \tag{1.14}$$

Using the fact that $p(0) - p_h(0) = 0$, we have

$$p(x) - p_h(x) = \int_0^x \left(\frac{dp}{dx} - \frac{dp_h}{dx} \right)(y)\, dy, \qquad x \in [0,1] ,$$

which, together with (1.14), implies

$$|p(x) - p_h(x)| \leq h \max_{y\in[0,1]} \left| \frac{d^2p(y)}{dx^2} \right|, \qquad x \in [0,1] . \tag{1.15}$$

Note that (1.15) is less sharp in h than the first estimate in (1.13) for the *interpolation error*. With a more delicate analysis, we can show that the first error estimate in (1.13) holds for p_h as well as $\tilde{p}_h$. In fact, it can be shown that $p_h = \tilde{p}_h$ (cf. Exercise 1.6), which is true only for one dimension.

In summary, we have obtained the quantitative estimates in (1.14) and (1.15), which show that the approximate solution of (1.5) approaches the exact solution of (1.1) as h goes to zero. This implies *convergence* of the finite element method (1.5).

The second remark is on the equivalence between (1.2) and (1.3). Let p be a solution of (1.2). Then, for any $v \in V$ and any $\epsilon \in \mathbb{R}$, we have

$$F(p) \leq F(p + \epsilon v) .$$

With the definition

$$\begin{aligned} G(\epsilon) &= F(p + \epsilon v) \\ &= \frac{1}{2} \left(\frac{dp}{dx}, \frac{dp}{dx} \right) + \epsilon \left(\frac{dp}{dx}, \frac{dv}{dx} \right) + \frac{\epsilon^2}{2} \left(\frac{dv}{dx}, \frac{dv}{dx} \right) - \epsilon(f, v) - (f, p) , \end{aligned}$$

we see that G has a minimum at $\epsilon = 0$, so $\frac{dG}{d\epsilon}(0) = 0$. Since

$$\frac{dG}{d\epsilon}(0) = \left(\frac{dp}{dx}, \frac{dv}{dx} \right) - (f, v) ,$$

p is a solution of (1.3). Conversely, suppose that p is a solution of (1.3). With any $v \in V$, set $w = v - p \in V$; we find that

$$\begin{aligned} F(v) = F(p + w) &= \frac{1}{2} \left(\frac{d(p+w)}{dx}, \frac{d(p+w)}{dx} \right) - (f, p + w) \\ &= \frac{1}{2} \left(\frac{dp}{dx}, \frac{dp}{dx} \right) - (f, p) + \left(\frac{dp}{dx}, \frac{dw}{dx} \right) - (f, w) + \frac{1}{2} \left(\frac{dw}{dx}, \frac{dw}{dx} \right) \\ &= \frac{1}{2} \left(\frac{dp}{dx}, \frac{dp}{dx} \right) - (f, p) + \frac{1}{2} \left(\frac{dw}{dx}, \frac{dw}{dx} \right) \geq F(p) , \end{aligned}$$

which implies that p is a solution of (1.2). Because of the equivalence between (1.1) and (1.3), (1.2) is equivalent to (1.1), too.

1.1.2 A Two-Dimensional Model Problem

In this subsection, we consider a stationary problem in two dimensions

$$\begin{aligned} -\Delta p &= f && \text{in } \Omega\,, \\ p &= 0 && \text{on } \Gamma\,, \end{aligned} \tag{1.16}$$

where Ω is a bounded domain in the plane with boundary Γ, f is a given real-valued piecewise continuous bounded function in Ω, and the *Laplacian operator* Δ is defined by

$$\Delta p = \frac{\partial^2 p}{\partial x_1^2} + \frac{\partial^2 p}{\partial x_2^2}\,.$$

Corresponding to the one-dimensional problem (1.1) for an elastic bar, consider an elastic membrane fixed at its boundary and subject to a transversal load of intensity f (cf. Fig. 1.4). Under the assumption of small displacements, we can check that the transversal displacement p satisfies (1.16).

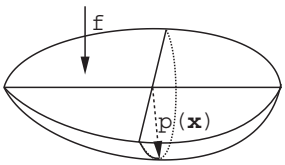

Fig. 1.4. An elastic membrane

We introduce the linear space

$$V = \left\{ v : \ v \text{ is a continuous function on } \Omega, \frac{\partial v}{\partial x_1} \text{ and } \frac{\partial v}{\partial x_2} \text{ are piecewise continuous and bounded on } \Omega, \text{ and } v = 0 \text{ on } \Gamma \right\}.$$

This space is a subspace of a Sobolev space introduced in the next section. Let us recall *Green's formula*. For a vector-valued function $\mathbf{b} = (b_1, b_2)$, the *divergence theorem* reads:

$$\int_\Omega \nabla \cdot \mathbf{b}\ d\mathbf{x} = \int_\Gamma \mathbf{b} \cdot \boldsymbol{\nu}\ d\ell\,, \tag{1.17}$$

where we recall that the divergence operator is given by

$$\nabla \cdot \mathbf{b} = \frac{\partial b_1}{\partial x_1} + \frac{\partial b_2}{\partial x_2}\,,$$

$\boldsymbol{\nu}$ is the outward unit normal to Γ, and the dot product $\mathbf{b} \cdot \boldsymbol{\nu}$ is defined by

$$\mathbf{b} \cdot \boldsymbol{\nu} = b_1\nu_1 + b_2\nu_2 \ .$$

With $v,\ w \in V$, we take $\mathbf{b} = (\frac{\partial v}{\partial x_1} w, 0)$ and $\mathbf{b} = (0, \frac{\partial v}{\partial x_2} w)$ in (1.17), respectively, to see that

$$\int_\Omega \frac{\partial^2 v}{\partial x_i^2} w \ d\mathbf{x} + \int_\Omega \frac{\partial v}{\partial x_i} \frac{\partial w}{\partial x_i} \ d\mathbf{x} = \int_\Gamma \frac{\partial v}{\partial x_i} w \nu_i \ d\ell, \quad i = 1, 2 \ . \tag{1.18}$$

Using the definition of the *gradient operator*, i.e.,

$$\nabla v = \left(\frac{\partial v}{\partial x_1}, \frac{\partial v}{\partial x_2} \right),$$

we sum over $i = 1, 2$ in (1.18) to obtain

$$\int_\Omega \Delta v \ w \ d\mathbf{x} = \int_\Gamma \frac{\partial v}{\partial \boldsymbol{\nu}} w \ d\ell - \int_\Omega \nabla v \cdot \nabla w \ d\mathbf{x} \ , \tag{1.19}$$

where the *normal derivative* is expressed by

$$\frac{\partial v}{\partial \boldsymbol{\nu}} = \frac{\partial v}{\partial x_1}\nu_1 + \frac{\partial v}{\partial x_2}\nu_2 \ .$$

Relation (1.19) is *Green's formula*, and it also holds in three dimensions (cf. Exercise 1.7).

Introduce the notation

$$a(p, v) = \int_\Omega \nabla p \cdot \nabla v \ d\mathbf{x}, \quad (f, v) = \int_\Omega f v \ d\mathbf{x} \ .$$

The form $a(\cdot, \cdot)$ is called a *bilinear form* on $V \times V$ (cf. Sect. 1.3). Also, we define the functional $F : V \to \mathbb{R}$ by

$$F(v) = \frac{1}{2} a(v, v) - (f, v), \qquad v \in V \ .$$

As in one dimension, (1.16) can be formulated as the minimization problem

$$\text{Find } p \in V \text{ such that } F(p) \leq F(v) \quad \forall v \in V \ .$$

This problem is equivalent to the variational problem (1.20) below, exactly using the same proof as for (1.2) and (1.3).

Multiplying the first equation of (1.16) by $v \in V$ and integrating over Ω, we see that

$$-\int_\Omega \Delta p \ v \ d\mathbf{x} = \int_\Omega f v \ d\mathbf{x} \ .$$

Applying (1.19) to this equation and using the homogeneous boundary condition lead to

$$\int_\Omega \nabla p \cdot \nabla v \; d\mathbf{x} = \int_\Omega f v \; d\mathbf{x} \qquad \forall v \in V \; .$$

Thus we derive the variational form

$$\text{Find } p \in V \text{ such that } a(p, v) = (f, v) \qquad \forall v \in V \; . \tag{1.20}$$

We now construct the finite element method for (1.16). For simplicity, in this section, we assume that Ω is a polygonal domain. A curved domain Ω will be handled in Sect. 1.5. Let K_h be a partition, called a *triangulation*, of Ω into non-overlapping (open) triangles K_i (cf. Fig. 1.5):

$$\bar{\Omega} = \bar{K}_1 \cup \bar{K}_2 \cup \ldots \cup \bar{K}_{\mathcal{M}} \; ,$$

such that no vertex of one triangle lies in the interior of an edge of another triangle, where $\bar{\Omega}$ represents the closure of Ω (i.e., $\bar{\Omega} = \Omega \cup \Gamma$) and a similar meaning holds for each K_i.

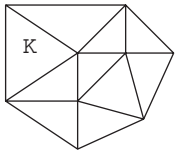

Fig. 1.5. A finite element partition in two dimensions

For (open) triangles $K \in K_h$, we define the *mesh parameters*

$$\text{diam}(K) = \text{the longest edge of } \bar{K} \text{ and } h = \max_{K \in K_h} \text{diam}(K) \; .$$

Now, we introduce the finite element space

$$V_h = \{v : v \text{ is a continuous function on } \Omega, v \text{ is linear} \\ \text{on each triangle } K \in K_h, \text{ and } v = 0 \text{ on } \Gamma\} \; .$$

Notice that $V_h \subset V$. The finite element method for (1.16) is formulated as

$$\text{Find } p_h \in V_h \text{ such that } a(p_h, v) = (f, v) \qquad \forall v \in V_h \; . \tag{1.21}$$

Existence and uniqueness of a solution to (1.21) can be checked as for (1.5). Also, in the same fashion as in the proof of the equivalence between (1.2) and (1.3), one can check that (1.21) is equivalent to a discrete minimization problem:

$$\text{Find } p_h \in V_h \text{ such that } F(p_h) \le F(v) \qquad \forall v \in V_h \; .$$

Denote the vertices (*nodes*) of the triangles in K_h by $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{\tilde{M}}$. The basis functions φ_i in V_h, $i = 1, 2, \ldots, \tilde{M}$, are defined by

$$\varphi_i(\mathbf{x}_j) = \begin{cases} 1 & \text{if } i = j \ , \\ 0 & \text{if } i \neq j \ . \end{cases}$$

The *support* of φ_i, i.e., the set of $\mathbf{x}$ where $\varphi_i(\mathbf{x}) \neq 0$, consists of the triangles with the common node $\mathbf{x}_i$; see Fig. 1.6. The function φ_i is also called a hat or chapeau function.

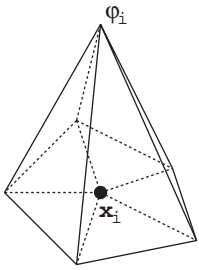

Fig. 1.6. A basis function in two dimensions

Let M be the number of interior vertices in K_h; for convenience, let the first M vertices be the interior ones. As in the previous subsection, any function $v \in V_h$ has the unique representation

$$v(\mathbf{x}) = \sum_{i=1}^{M} v_i \varphi_i(\mathbf{x}), \qquad \mathbf{x} \in \Omega \ ,$$

where $v_i = v(\mathbf{x}_i)$. Due to the boundary condition, we exclude the vertices on the boundary of Ω.

In the same way as for (1.5), (1.21) can be written in matrix form (cf. Exercise 1.8)

$$\mathbf{A}\mathbf{p} = \mathbf{f}, \tag{1.22}$$

where, as before, the matrix $\mathbf{A}$ and vectors $\mathbf{p}$ and $\mathbf{f}$ are given by

$$\mathbf{A} = (a_{ij}) \, , \quad \mathbf{p} = (p_j) \, , \quad \mathbf{f} = (f_j) \ ,$$

with

$$a_{ij} = a\,(\varphi_i, \varphi_j) \, , \quad f_j = (f, \varphi_j) \, , \quad i, j = 1, 2, \ldots, M \ .$$

As in one dimension, it can be checked that the stiffness matrix $\mathbf{A}$ is symmetric positive definite. In particular, it is nonsingular. Consequently, (1.22) and thus (1.21) has a unique solution. Also, notice that $\mathbf{A}$ is sparse from the construction of the basis functions.

As an example, we consider the case where the domain is the unit square $\Omega = (0,1) \times (0,1)$ and K_h is the uniform triangulation of Ω as illustrated in

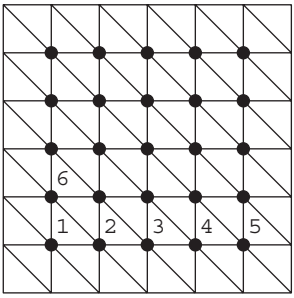

Fig. 1.7. An example of a triangulation

Fig. 1.7 with the indicated *enumeration* of nodes. In this case, the matrix $\mathbf{A}$ has the form (cf. Exercise 1.9)

$$\mathbf{A} = \begin{pmatrix} 4 & -1 & 0 & 0 & \dots & 0 & -1 & 0 & \dots & 0 & 0 \\ -1 & 4 & -1 & 0 & \dots & 0 & 0 & -1 & \dots & 0 & 0 \\ 0 & -1 & 4 & -1 & \dots & 0 & 0 & 0 & \dots & -1 & 0 \\ 0 & 0 & -1 & 4 & \dots & 0 & 0 & 0 & \dots & 0 & -1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 4 & -1 & 0 & \dots & 0 & 0 \\ -1 & 0 & 0 & 0 & \dots & -1 & 4 & -1 & \dots & 0 & 0 \\ 0 & -1 & 0 & 0 & \dots & 0 & -1 & 4 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & -1 & 0 & \dots & 0 & 0 & 0 & \dots & 4 & -1 \\ 0 & 0 & 0 & -1 & \dots & 0 & 0 & 0 & \dots & -1 & 4 \end{pmatrix} .$$

Note that associated with the four corner nodes (e.g., node 1), there are only three nonzeros per row; the adjacent diagonal entry for such a node (e.g., node 5) may be zero. For other nodes adjacent to the boundary (e.g., node 2), there are solely four nonzeros per row. From this form of $\mathbf{A}$, the left-hand side of the ith equation in (1.22) is a linear combination of the values of p_h at most at the five nodes illustrated in Fig. 1.8. After division by h^2, system (1.22) can be treated as a linear system generated by a *five-point difference stencil scheme* for (1.16).

In practical computations (cf. Sect. 1.1.4), the entries a_{ij} in $\mathbf{A}$ are obtained by summing the contributions from different triangles $K \in K_h$:

$$a_{ij} = a(\varphi_i, \varphi_j) = \sum_{K \in K_h} a^K(\varphi_i, \varphi_j) \ ,$$

where

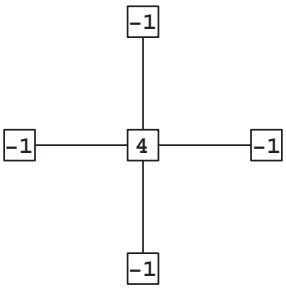

Fig. 1.8. A five-point stencil scheme

$$a_{ij}^K \equiv a^K(\varphi_i, \varphi_j) = \int_K \nabla\varphi_i \cdot \nabla\varphi_j \, d\mathbf{x} \,. \tag{1.23}$$

Using the definition of the basis functions, we see that $a^K(\varphi_i, \varphi_j) = 0$ unless nodes $\mathbf{x}_i$ and $\mathbf{x}_j$ are both vertices of K.

We end with a remark on an error estimate. As noted earlier, the derivation of an estimate is very delicate. An approximation theory will be presented in Sect. 1.9. Until then, all error estimates for multi-dimensional problems will be just stated without proof. By the same argument as for (1.11), we have

$$\|\nabla p - \nabla p_h\| \leq \|\nabla p - \nabla v\| \qquad \forall v \in V_h \,,$$

where p and p_h are the respective solutions of (1.20) and (1.21), and we recall that $\|\cdot\|$ is the norm

$$\|\nabla p\| = \left(\int_\Omega \left(\left(\frac{\partial p}{\partial x_1} \right)^2 + \left(\frac{\partial p}{\partial x_2} \right)^2 \right) d\mathbf{x} \right)^{1/2} .$$

This means that p_h is the best possible approximation of p in V_h in terms of the norm deduced from the bilinear form $a(\cdot,\cdot)$. Applying the approximation theory developed in Sect. 1.9, it holds that

$$\|p - p_h\| + h\,\|\nabla p - \nabla p_h\| \leq Ch^2 \,, \tag{1.24}$$

where the constant C depends on the second partial derivatives of p and the smallest angle of the triangles $K \in K_h$, but does not depend on h. error estimate (1.24) indicates that if the solution is sufficiently smooth, p_h tends to p in the norm $\|\cdot\|$ as h approaches zero.

1.1.3 An Extension to General Boundary Conditions

We now extend the finite element method to the stationary problem with another type of boundary condition

$$\begin{aligned} -\Delta p &= f && \text{in } \Omega\ , \\ \gamma p + \frac{\partial p}{\partial \boldsymbol{\nu}} &= g && \text{on } \Gamma\ , \end{aligned} \tag{1.25}$$

where γ and g are given functions and we recall that $\partial p/\partial \boldsymbol{\nu}$ is the normal derivative. This type of boundary condition is called a *third*, *mixed*, *Robin*, or *Dankwerts* boundary condition. When $\gamma = 0$, the boundary condition is a *second* or *Neumann* condition. When γ is infinite, the boundary condition reduces to a *first* or *Dirichlet condition*, which was considered in the previous subsection. A *fourth type* of boundary condition (i.e., a *periodic* boundary condition) will be considered in Chap. 5. In this subsection, we treat the case where γ is bounded.

Note that if $\gamma = 0$ on Γ, Green's formula (1.19) and (1.25) imply that (cf. Exercise 1.11)

$$\int_\Omega f \, d\mathbf{x} + \int_\Gamma g \, d\ell = 0\ . \tag{1.26}$$

For (1.25) to have a solution, the *compatibility condition* (1.26) must be satisfied. In this case, p is unique only up to an additive constant.

Introduce the linear space

$$V = \left\{ v : v \text{ is a continuous function on } \Omega, \text{ and } \frac{\partial v}{\partial x_1} \text{ and } \frac{\partial v}{\partial x_2} \right.$$
$$\left. \text{are piecewise continuous and bounded on } \Omega \right\}\ ,$$

and the notation

$$\begin{aligned} a(v,w) &= \int_\Omega \nabla v \cdot \nabla w \, d\mathbf{x} + \int_\Gamma \gamma v w \, d\ell, && v,\, w \in V\ , \\ (f,v) &= \int_\Omega f v \, d\mathbf{x}, \quad (g,v)_\Gamma = \int_\Gamma g v \, d\ell, && v \in V\ . \end{aligned}$$

Then, as in the previous subsection, (1.25) can be written (cf. Exercise 1.12):

$$\text{Find } p \in V \text{ such that } a(p,v) = (f,v) + (g,v)_\Gamma \qquad \forall v \in V\ . \tag{1.27}$$

Note that the boundary condition in (1.25) is not imposed in the definition of V. It appears implicitly in (1.27). A boundary condition that need not be imposed is called a *natural condition*. The pure Neumann boundary condition is natural. The Dirichlet boundary condition has been imposed explicitly in V in Sect. 1.1.2, and is termed an *essential condition*.

If $\gamma \equiv 0$, the definition of V needs to be modified to take into account the up-to a constant uniqueness of solution to (1.25). That is, the space V can be modified to, say,

$$V = \Big\{ v : v \text{ is a continuous function on } \Omega, \frac{\partial v}{\partial x_1} \text{ and } \frac{\partial v}{\partial x_2}$$
$$\text{are piecewise continuous and bounded on } \Omega \,,$$
$$\text{and } \int_\Omega v \; d\mathbf{x} = 0 \Big\} \,.$$

To construct the finite element method for (1.25), let K_h be a triangulation of Ω as in the previous subsection. The finite element space V_h is defined by

$$V_h = \{ v : v \text{ is a continuous function on } \Omega \text{ and}$$
$$\text{is linear on each triangle } K \in K_h \} \,.$$

Note that the functions in V_h are not required to satisfy any boundary condition. Now, the finite element solution satisfies

$$\text{Find } p_h \in V_h \text{ such that } a(p_h, v) = (f, v) + (g, v)_\Gamma \qquad \forall v \in V_h \,. \tag{1.28}$$

Again, for the pure Neumann boundary condition, V_h needs to be modified to

$$V_h = \Big\{ v : v \text{ is a continuous function on } \Omega \text{ and is linear}$$
$$\text{on each triangle } K \in K_h, \text{ and } \int_\Omega v \; d\mathbf{x} = 0 \Big\} \,.$$

As in the last two subsections, (1.28) can be formulated in matrix form, and an error estimate can be similarly stated under an appropriate smoothness assumption on the solution p that involves its second partial derivatives.

The *Poisson equation* has been considered in (1.16) and (1.25). More general partial differential equations will be treated in subsequent sections and chapters.

1.1.4 Programming Considerations

The essential features of a typical computer program implementing the finite element method are included in the following parts:

- Input of data such as the domain Ω, the right-hand side function f, the boundary data γ and g (cf. (1.25)), and the coefficients that may appear in a differential problem;
- Construction of the triangulation K_h;
- Computation and assembly of the stiffness matrix $\mathbf{A}$ and the right-hand side vector $\mathbf{f}$;
- Solution of the linear system of algebraic equations $\mathbf{Ap} = \mathbf{f}$;
- Output of the computational results.

The data input can be easily implemented in a small subroutine, and the result output depends on the computer system and software the user has. Here we briefly discuss the other three parts. As an illustration, we focus on two dimensions.

1.1.4.1 Construction of the Triangulation K_h

The triangulation K_h can be constructed from a successive refinement of an initial coarse partition of Ω; fine triangles can be obtained by connecting the midpoints of edges of coarse triangles, for example. A sequence of uniform refinements will lead to *quasi-uniform* grids where the triangles in K_h essentially have the same size in all regions of Ω (cf. Fig. 1.9). If the boundary Γ of Ω is a curve, special care needs to be taken of near Γ (cf. Sect. 1.5).

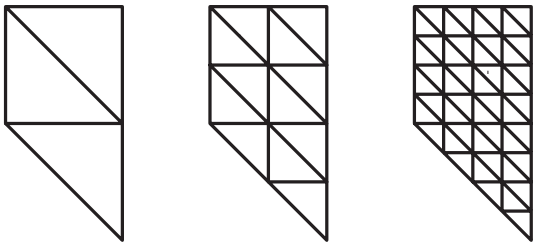

Fig. 1.9. Uniform refinement

In practical applications, it is often necessary to use triangles in K_h that vary considerably in size in different regions of Ω. For example, one utilizes smaller triangles in regions where the exact solution has a fast variation or where certain derivatives are large; see Fig. 1.10, where a *local refinement* strategy is carried out. In this strategy, proper care is taken of in the transition zone between regions with triangles of different sizes so that a so-called *regular* local refinement results (i.e, no vertex of one triangle lies in the interior of an edge of another triangle; see Chap. 6). Methods that automatically refine grids where needed are called *adaptive methods*, and will be studied in detail in Chap. 6.

Let a triangulation K_h have M nodes and $\mathcal{M}$ triangles. The triangulation can be represented by two arrays $\mathbf{Z}(2, M)$ and $\mathcal{Z}(3, \mathcal{M})$, where $\mathbf{Z}(i, j)$ $(i = 1, 2)$ indicates the coordinates of the jth node, $j = 1, 2, \ldots, M$, and $\mathcal{Z}(i, k)$ $(i = 1, 2, 3)$ enumerates the nodes of the kth triangle, $k = 1, 2, \ldots, \mathcal{M}$. An example is demonstrated in Fig. 1.11, where the triangle numbers are denoted in circles. For this example, the array $\mathcal{Z}(3, \mathcal{M})$ is of the form, where $M = \mathcal{M} = 11$:

$$\mathcal{Z} = \begin{pmatrix} 1 & 1 & 2 & 3 & 4 & 4 & 5 & 6 & 7 & 7 & 8 \\ 2 & 4 & 5 & 4 & 5 & 7 & 9 & 7 & 9 & 10 & 10 \\ 4 & 3 & 4 & 6 & 7 & 6 & 7 & 8 & 10 & 8 & 11 \end{pmatrix} .$$

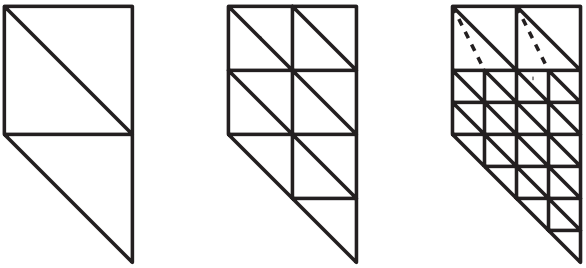

Fig. 1.10. Nonuniform refinement

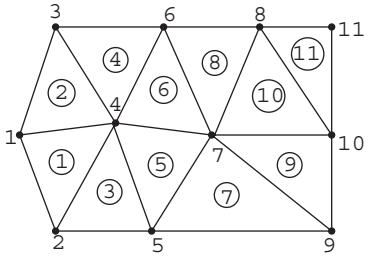

Fig. 1.11. Node and triangle enumeration

If a direct method (Gaussian elimination) is employed to solve the linear system $\mathbf{Ap} = \mathbf{f}$, the nodes should be enumerated in such a way that the *band width* of each row in $\mathbf{A}$ is as small as possible. This matter will be studied in Sect. 1.10, in connection with the discussion of solution methods for linear systems.

In general, when local refinement is involved in a triangulation K_h, it is very difficult to enumerate the nodes and triangles efficiently; some strategies will be given in Chap. 6. For a simple domain Ω (e.g., a convex polygonal Ω), it is rather easy to construct and represent a triangulation that utilizes uniform refinement in the whole domain.

1.1.4.2 Assembly of the Stiffness Matrix

After the triangulation K_h is constructed, one computes the *element stiffness matrices* with entries a_{ij}^K given by (1.23). We recall that $a_{ij}^K = 0$ unless nodes $\mathbf{x}_i$ and $\mathbf{x}_j$ are both vertices of $K \in K_h$.

For a kth triangle K_k, $\mathcal{Z}(m,k)$ $(m = 1,2,3)$ are the numbers of the vertices of K_k, and the element stiffness matrix $\mathbf{A}^{(k)} = \left(a_{mn}^k\right)_{m,n=1}^3$ is now calculated as follows:

$$a_{mn}^k = \int_{K_k} \nabla\varphi_m \cdot \nabla\varphi_n \, d\mathbf{x}, \quad m,\ n = 1,2,3\ ,$$

where the (linear) basis function φ_m on K_k satisfies

$$\varphi_m(\mathbf{x}_{\mathcal{Z}(n,k)}) = \begin{cases} 1 & \text{if } m = n \ , \\ 0 & \text{if } m \neq n \ . \end{cases}$$

The right-hand side on K_k is computed by

$$f_m^k = \int_{K_k} f\varphi_m \, d\mathbf{x}, \quad m = 1, 2, 3 \ .$$

Note that m and n are the local numbers of the three vertices of K_k, while i and j used in (1.23) are the global numbers of vertices in K_h.

To assemble the global matrix $\mathbf{A} = (a_{ij})$ and the right-hand side $\mathbf{f} = (f_j)$, one loops over all triangles K_k and successively adds the contributions from different $K_k's$:

$$\begin{aligned} &\text{For } k = 1, 2, \ldots, \mathcal{M}, \text{ compute} \\ &a_{\mathcal{Z}(m,k),\mathcal{Z}(n,k)} = a_{\mathcal{Z}(m,k),\mathcal{Z}(n,k)} + a_{mn}^k \ , \\ &f_{\mathcal{Z}(m,k)} = f_{\mathcal{Z}(m,k)} + f_m^k, \quad m, \ n = 1, 2, 3 \ . \end{aligned}$$

The approach used is *element-oriented*; that is, we loop over elements (i.e., triangles). Experience shows that this approach is more efficient than the *node-oriented* approach (i.e., looping over all nodes); the latter approach wastes too much time in repeated computations of $\mathbf{A}$ and $\mathbf{f}$.

1.1.4.3 Solution of a Linear System

The solution of the linear system $\mathbf{Ap} = \mathbf{f}$ can be performed via a direct method (Gaussian elimination) or an iterative method (e.g., the conjugate gradient method), which will be discussed in Sect. 1.10. Here we just mention that in use of these two methods, it is not necessary to exploit an array $\mathbf{A}(M, M)$ to store the stiffness matrix $\mathbf{A}$. Instead, since $\mathbf{A}$ is sparse and usually a band matrix, only the nonzero entries of $\mathbf{A}$ need to be stored, say, in an one-dimensional array.

1.2 Sobolev Spaces

In the previous section, an introductory finite element method was developed for two simple model problems. To present the finite element method in a general formulation, we need to use function spaces. This section is devoted to the development of the function spaces that are slightly more general than the spaces of continuous functions with piecewise continuous derivatives utilized in the previous section. We establish the small fraction of these spaces that is sufficient to develop the foundation of the finite element method as studied in this book.

1.2.1 Lebesgue Spaces

In this section, we assume that Ω is an open subset of $\mathbb{R}^d$, $1 \le d \le 3$, with piecewise smooth boundary. For a real-valued function v on Ω, we use the notation

$$\int_\Omega v(\mathbf{x})\ d\mathbf{x}$$

to denote the integral of f in the sense of Lebesgue (Rudin, 1987). For $1 \le q < \infty$, define

$$\|v\|_{L^q(\Omega)} = \left(\int_\Omega |v(\mathbf{x})|^q\ d\mathbf{x} \right)^{1/q} .$$

For $q = \infty$, set

$$\|v\|_{L^\infty(\Omega)} = \text{ess sup}\{|v(\mathbf{x})| : \mathbf{x} \in \Omega\} ,$$

where ess sup denotes the *essential supremum*. Now, for $1 \le q \le \infty$, we define the *Lebesgue spaces*

$$L^q(\Omega) = \{v :\ v \text{ is defined on } \Omega \text{ and } \|v\|_{L^q(\Omega)} < \infty\} .$$

For $q = 2$, for example, $L^2(\Omega)$ consists of all *square integrable functions* on Ω (in the sense of Lebesgue). To avoid trivial differences, we identify two functions u and v whenever $\|u - v\|_{L^q(\Omega)} = 0$; i.e., $u(\mathbf{x}) = v(\mathbf{x})$ for $\mathbf{x} \in \Omega$, except on a set of measure zero.

Given a linear (vector) space V, a *norm* in V, $\|\cdot\|$, is a function from V to $\mathbb{R}$ such that

- $\|v\| \ge 0 \quad \forall v \in V$; $\|v\| = 0$ if and only if $v = 0$.
- $\|cv\| = |c|\ \|v\| \quad \forall c \in \mathbb{R},\ v \in V$.
- $\|u + v\| \le \|u\| + \|v\| \quad \forall u, v \in V$ (the triangle inequality).

A linear space V endowed with a norm $\|\cdot\|$ is called a *normed linear space*. V is termed *complete* if every *Cauchy sequence* $\{v_i\}$ in V has a limit v that is an element of V. The Cauchy sequence $\{v_i\}$ means that $\|v_i - v_j\| \to 0$ as $i, j \to \infty$, and completeness says that $\|v_i - v\| \to 0$ as $i \to \infty$. A normed linear space $(V, \|\cdot\|)$ is called a *Banach space* if it is complete with respect to the norm $\|\cdot\|$. For $1 \le q \le \infty$, the space $L^q(\Omega)$ is a Banach space (Adams, 1975).

There are several useful inequalities that hold for functions in $L^q(\Omega)$. We state them without proof (Adams, 1975).

Hölder's inequality: For $1 \le q, q' \le \infty$ such that $1/q + 1/q' = 1$, it holds that

$$\|uv\|_{L^1(\Omega)} \le \|u\|_{L^q(\Omega)} \|v\|_{L^{q'}(\Omega)} \qquad \forall u \in L^q(\Omega),\ v \in L^{q'}(\Omega) . \tag{1.29}$$

When $q = q' = 2$, this inequality is also called *Cauchy's* or *Schwarz's inequality*:

$$\|uv\|_{L^1(\Omega)} \leq \|u\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \qquad \forall u,\, v \in L^2(\Omega)\ . \tag{1.30}$$

The triangle inequality applied to $L^q(\Omega)$ is referred to as *Minkowski's inequality*:

$$\|u+v\|_{L^q(\Omega)} \leq \|u\|_{L^q(\Omega)} + \|v\|_{L^q(\Omega)} \qquad \forall u,\, v \in L^q(\Omega)\ . \tag{1.31}$$

1.2.2 Weak Derivatives

We introduce the notation

$$D^\alpha v = \frac{\partial^{|\alpha|} v}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_d^{\alpha_d}}\ ,$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ is a multi-index (called a d-tuple), with $\alpha_1, \alpha_2, \ldots, \alpha_d$ nonnegative integers, and $|\alpha| = \alpha_1 + \alpha_2 + \ldots + \alpha_d$ is the length of α. This notation indicates a partial derivative of v. For example, as $d = 2$, a second partial derivative can be written as $D^\alpha v$ with $\alpha = (2,0)$, $\alpha = (1,1)$, or $\alpha = (0,2)$.

In calculus, derivatives of a function are defined pointwise. The variational formulation in the finite element method is given globally, i.e., in terms of integrals on Ω. Pointwise values of derivatives are not needed; only derivatives that can be interpreted as functions in Lebesgue spaces are used. Hence it is natural to introduce a global definition of *derivative* more suitable to the Lebesgue spaces.

For a continuous function v defined on Ω, the *support* of v is the closure of the (open) set $\{v : v(\mathbf{x}) \neq 0, \mathbf{x} \in \Omega\}$. If this set is *compact* (i.e., bounded), then v is called to have *compact support* in Ω. When Ω is bounded, it is equivalent to saying that v vanishes in a neighborhood of the boundary Γ of Ω.

For $\Omega \subset \mathbb{R}^d$, indicate by $\mathcal{D}(\Omega)$ or $C_0^\infty(\Omega)$ the subset of $C^\infty(\Omega)$ (the linear space of functions infinitely differentiable) functions that have compact support in Ω. We use the space $\mathcal{D}(\Omega)$ to introduce the concept of weak (generalized) derivatives. For this, we need the following function space:

$$L^1_{loc}(\Omega) = \{v : v \in L^1(K) \text{ for any compact } K \text{ inside } \Omega\}\ .$$

Note that $L^1_{loc}(\Omega)$ contains all of $C^0(\Omega)$ (continuous functions in Ω). Functions in $L^1_{loc}(\Omega)$ can behave arbitrarily badly near the boundary. With $\text{dist}(\mathbf{x}, \Gamma)$ denoting the distance from $\mathbf{x}$ to Γ, the function $e^{e^{1/\text{dist}(\mathbf{x},\Gamma)}} \in L^1_{loc}(\Omega)$, for example.

A function $v \in L^1_{loc}(\Omega)$ is said to have a *weak derivative*, $D^\alpha_w v$, if there is a function $u \in L^1_{loc}(\Omega)$ such that

$$\int_\Omega u(\mathbf{x})\varphi(\mathbf{x})\ d\mathbf{x} = (-1)^{|\alpha|}\int_\Omega v(\mathbf{x})D^\alpha\varphi(\mathbf{x})\ d\mathbf{x} \qquad \forall\varphi\in\mathcal{D}(\Omega)\ .$$

If such a function u does exist, we write $D^\alpha_w v = u$.

For any multi-index α, if $v \in C^{|\alpha|}(\Omega)$, then the weak derivative $D^\alpha_w v$ exists and equals $D^\alpha v$ (cf. Exercise 1.13). Consequently, we will ignore the difference in the definition of D^α_w and D^α. Namely, if classical derivatives do not exist, the differentiation symbol D^α will refer to weak derivatives.

Example. We consider a simple example with $d = 1$ and $\Omega = (-1, 1)$. Let $v(x) = 1 - |x|$. Then $D^1 v$ exists and is given by

$$u(x) = \begin{cases} -1 & \text{if } x > 0\ , \\ 1 & \text{if } x < 0\ . \end{cases}$$

In fact, for $\varphi \in \mathcal{D}(\Omega)$, an application of integration by parts yields

$$\begin{aligned}
&\int_{-1}^{1} v(x)\frac{d\varphi}{dx}(x)\ dx \\
&\quad = \int_{-1}^{0} v(x)\frac{d\varphi}{dx}(x)\ dx + \int_{0}^{1} v(x)\frac{d\varphi}{dx}(x)\ dx \\
&\quad = [v\varphi]_{-1}^{0} - \int_{-1}^{0}(+1)\varphi(x)\ dx + [v\varphi]_{0}^{1} - \int_{0}^{1}(-1)\varphi(x)\ dx \\
&\quad = -\int_{-1}^{1} u(x)\varphi(x)\ dx\ ,
\end{aligned}$$

since v is continuous at 0.

Note that v is not differentiable at 0 in the classical sense. However, its first weak derivative exists. One can show that its higher order derivative $D^i v$ does not exist for $i > 2$ (cf. Exercise 1.14).

1.2.3 Sobolev Spaces

We now use weak derivatives to generalize the Lebesgue spaces introduced in Sect. 1.2.1.

For $r = 1, 2, \ldots$ and $v \in L^1_{loc}(\Omega)$, assume that the weak derivatives $D^\alpha v$ exist for all $|\alpha| \le r$. We define the *Sobolev norm*

$$\|v\|_{W^{r,q}(\Omega)} = \left(\sum_{|\alpha|\le r}\|D^\alpha v\|^q_{L^q(\Omega)}\right)^{1/q}\ ,$$

if $1 \le q < \infty$. For $q = \infty$, define

$$\|v\|_{W^{r,\infty}(\Omega)} = \max_{|\alpha|\le r}\|D^\alpha v\|_{L^\infty(\Omega)}\ .$$

The *Sobolev spaces* are defined by

$$W^{r,q}(\Omega) = \left\{ v \in L^1_{loc}(\Omega) : \|v\|_{W^{r,q}(\Omega)} < \infty \right\}, \qquad 1 \le q \le \infty .$$

One can check that $\|\cdot\|_{W^{r,q}(\Omega)}$ is a norm; moreover, the Sobolev space $W^{r,q}(\Omega)$ is a Banach space (Adams, 1975).

We denote by $W_0^{r,q}(\Omega)$ the completion of $\mathcal{D}(\Omega)$ with respect to the norm $\|\cdot\|_{W^{r,q}(\Omega)}$.

For $\Omega \subset \mathbb{R}^d$ with smooth boundary and $v \in W^{1,q}(\Omega)$, the restriction to the boundary Γ, $v|_\Gamma$, can be interpreted as a function in $L^q(\Gamma)$ (Adams, 1975), $1 \le q \le \infty$. This does not assert that pointwise values of v on Γ make sense. For $q = 2$, for example, $v|_\Gamma$ is only square integrable on Γ. Using this property, the space $W_0^{r,q}(\Omega)$ can be characterized as

$$W_0^{r,q}(\Omega) = \left\{ v \in W^{r,q}(\Omega) : D^\alpha v|_\Gamma = 0 \text{ in } L^2(\Gamma),\ |\alpha| < r \right\} .$$

For later applications, the *seminorms* will be used:

$$|v|_{W^{r,q}(\Omega)} = \left(\sum_{|\alpha|=r} \|D^\alpha v\|^q_{L^q(\Omega)} \right)^{1/q}, \quad 1 \le q < \infty ,$$
$$|v|_{W^{r,\infty}(\Omega)} = \max_{|\alpha|=r} \|D^\alpha v\|_{L^\infty(\Omega)} .$$

Furthermore, for $q = 2$, we will utilize the symbols

$$H^r(\Omega) = W^{r,2}(\Omega),\ H_0^r(\Omega) = W_0^{r,2}(\Omega), \quad r = 1, 2, \ldots .$$

That is, the functions in $H^r(\Omega)$, together with their derivatives $D^\alpha v$ of order $|\alpha| \le r$, are square integrable in Ω. Note that $H^0(\Omega) = L^2(\Omega)$.

The Sobolev spaces $W^{r,q}(\Omega)$ have a number of important properties. Given the indices defining these spaces, it is natural that there are inclusion relations to provide some type of ordering among them. We list a couple of inclusion relations; see Exercise 1.15.

For nonnegative integers r and k such that $r \le k$, it holds that

$$W^{k,q}(\Omega) \subset W^{r,q}(\Omega), \qquad 1 \le q \le \infty . \tag{1.32}$$

In addition, when Ω is bounded,

$$W^{r,q'}(\Omega) \subset W^{r,q}(\Omega), \qquad 1 \le q \le q' \le \infty , \tag{1.33}$$

for $r = 1, 2, \ldots$.

1.2.4 Poincaré's Inequality

We show an important inequality which will be heavily used in this book, *Poincaré's inequality*. It is sometimes called Poincaré-Friedrichs' inequality or simply Friedrichs' inequality.

Before introducing this inequality in its general form, we first consider one dimension. For any $v \in C_0^\infty(I)$ ($I = (0,1)$, the unit interval), because $v(0) = 0$, we see that

$$v(x) = \int_0^x \frac{dv(y)}{dy}\, dy\ .$$

Consequently, by Cauchy's inequality (1.10), we have

$$\begin{aligned} |v(x)| &\leq \int_0^1 \left|\frac{dv(y)}{dy}\right|\, dy \leq \left(\int_0^1 dy\right)^{1/2} \left(\int_0^1 \left(\frac{dv}{dy}\right)^2\, dy\right)^{1/2} \\ &= \left(\int_0^1 \left(\frac{dv}{dy}\right)^2\, dy\right)^{1/2}\ , \end{aligned}$$

which, by squaring and integrating over I, yields

$$\|v\|_{L^2(I)} \leq |v|_{H^1(I)}\ .$$

Because $C_0^\infty(I)$ is dense in $H_0^1(I)$, we see that

$$\|v\|_{L^2(I)} \leq |v|_{H^1(I)} \qquad \forall v \in V = H_0^1(I)\ . \tag{1.34}$$

This is Poincaré's inequality in one dimension.

We can extend this argument to the case where Ω is a d-dimensional cube: $\Omega = \{(x_1, x_2, \ldots, x_d) : 0 < x_i < l,\ i = 1, 2, \ldots, d\}$, where $l > 0$ is a real number. Again, since $C_0^\infty(\Omega)$ is dense in $H_0^1(\Omega)$, it is sufficient to prove Poincaré's inequality for $v \in C_0^\infty(\Omega)$. Then we see that

$$v(x_1, x_2, \ldots, x_d) = v(0, x_2, \ldots, x_d) + \int_0^{x_1} \frac{\partial v}{\partial x_1}(y, x_2, \ldots, x_d)\, dy\ .$$

Because the boundary term vanishes, it follows from Cauchy's inequality (1.10) that

$$\begin{aligned} |v(\mathbf{x})|^2 &\leq \int_0^{x_1} dy \int_0^{x_1} \left|\frac{\partial v}{\partial x_1}(y, x_2, \ldots, x_d)\right|^2 dy \\ &\leq l \int_0^l \left|\frac{\partial v}{\partial x_1}(y, x_2, \ldots, x_d)\right|^2 dy\ . \end{aligned}$$

Integrating over Ω implies

$$\|v\|_{L^2(\Omega)} \leq l\, |v|_{H^1(\Omega)} \qquad \forall v \in H_0^1(\Omega)\ . \tag{1.35}$$

For a general open set $\Omega \subset \mathbb{R}^d$ with piecewise smooth boundary, if $v \in H^1(\Omega)$ vanishes on a part of the boundary Γ with this part having positive $(d-1)$-dimensional measure, then there is a positive constant C, depending only on Ω, such that (Adams, 1975)

$$\|v\|_{L^2(\Omega)} \leq C\,|v|_{H^1(\Omega)} \ . \tag{1.36}$$

If Ω is bounded, this inequality implies that the seminorm $|\cdot|_{H^1(\Omega)}$ is equivalent to the norm $\|\cdot\|_{H^1(\Omega)}$ in $H_0^1(\Omega)$. In general, an induction argument can be used to show that $|\cdot|_{H^r(\Omega)}$ is equivalent to $\|\cdot\|_{H^r(\Omega)}$ in $H_0^r(\Omega)$, $r = 1, 2, \ldots$.

1.2.5 Duality and Negative Norms

Let V be a Banach space. A mapping $L : V \to \mathbb{R}$ is called a *linear functional* if

$$L(\alpha u + \beta v) = \alpha L(u) + \beta L(v), \quad \alpha, \beta \in \mathbb{R},\ u,\ v \in V \ .$$

We say that L is *bounded* in the norm $\|\cdot\|_V$ if there is a constant $\tilde{L} > 0$ such that

$$|L(v)| \leq \tilde{L}\|v\|_V \qquad \forall v \in V \ .$$

The set of bounded linear functionals on V is termed the *dual space* of V, and is denoted by V'.

A bounded linear functional L is actually *Lipschitz continuous* (and thus continuous); i.e.,

$$|L(v) - L(w)| = |L(v - w)| \leq \tilde{L}\|v - w\|_V \qquad \forall v,\ w \in V \ .$$

Conversely, a continuous linear functional is also bounded. In fact, if it is not bounded, there is a sequence $\{v_i\}$ in V such that $|L(v_i)|/\|v_i\|_V \geq i$. Setting $w_i = v_i/\left(i\|v_i\|_V\right)$, we see that $|L(w_i)| \geq 1$ and $\|w_i\|_V = 1/i$. Then $w_i \to 0$ as $i \to \infty$, which, together with continuity of L, implies $L(w_i) \to 0$ as $i \to \infty$. This contradicts with $|L(w_i)| \geq 1$.

For $L \in V'$, define

$$\|L\|_{V'} = \sup_{0 \neq v \in V} \frac{L(v)}{\|v\|_V} \ .$$

Since L is bounded, this quantity is always finite. In fact, it induces a norm on V', called the *dual norm* (cf. Exercise 1.16), and V' is a Banach space with respect to it (Adams, 1975).

Let us consider the dual space of $L^q(\Omega)$, $1 \leq q < \infty$. For $f \in L^{q'}(\Omega)$, where $1/q + 1/q' = 1$, set

$$L(v) = \int_\Omega f(\mathbf{x}) v(\mathbf{x})\ d\mathbf{x}, \qquad v \in L^q(\Omega) \ .$$

It follows from Hölder's inequality (1.29) that L is bounded in the L^q-norm:

$$|L(v)| \leq \|f\|_{L^{q'}(\Omega)} \|v\|_{L^q(\Omega)}, \qquad v \in L^q(\Omega) \ .$$

Thus every function $f \in L^{q'}(\Omega)$ can be viewed as a bounded linear functional on $L^q(\Omega)$. Due to the Riesz Representation Theorem (cf. Sect. 1.3.1), all

bounded linear functionals on $L^q(\Omega)$ arise in this form, so $L^{q'}(\Omega)$ can be viewed as the dual space of $L^q(\Omega)$. The number q' is often termed the *dual index* of q.

For $1 \le q \le \infty$ and a positive integer r, the dual space of the Sobolev space $W^{r,q}(\Omega)$ is indicated by $W^{-r,q'}(\Omega)$, where q' is the dual index of q. The norm on $W^{-r,q'}(\Omega)$ is defined via *duality*:

$$\|L\|_{W^{-r,q'}(\Omega)} = \sup_{0 \neq v \in W^{r,q}(\Omega)} \frac{L(v)}{\|v\|_{W^{r,q}(\Omega)}}, \qquad L \in W^{-r,q'}(\Omega) \ .$$

1.3 Abstract Variational Formulation

The introductory finite element method discussed in Sect. 1.1 will be written in an abstract formulation in this section. We first provide this formulation and its theoretical analysis, and then give several concrete examples. These examples will utilize the Sobolev spaces introduced in Sect. 1.2, particularly, the spaces $H^r(\Omega)$, $r = 0, 1, 2, \ldots$.

1.3.1 An Abstract Formulation

A linear space V, together with an inner product $(\cdot, \cdot)$ defined on it, is called an *inner product* space and is represented by $\big(V, (\cdot, \cdot)\big)$. With the inner product $(\cdot, \cdot)$, there is an associated norm defined on V:

$$\|v\| = \sqrt{(v, v)}, \qquad v \in V \ .$$

Hence an inner product space can be always made to be a normed linear space. If the corresponding normed linear space $\big(V, \|\cdot\|\big)$ is complete, then $\big(V, (\cdot, \cdot)\big)$ is termed a *Hilbert space*.

The space $H^r(\Omega)$ $(r = 1, 2, \ldots)$, with the inner product

$$(u, v)_{H^r(\Omega)} = \sum_{|\alpha| \le r} \int_\Omega D^\alpha u(\mathbf{x}) D^\alpha v(\mathbf{x}) \, d\mathbf{x}, \quad u, \ v \in H^r(\Omega)$$

and the corresponding norm $\|\cdot\|_{H^r(\Omega)}$, is a Hilbert space (Adams, 1975).

Suppose that V is a Hilbert space with the scalar product $(\cdot, \cdot)$ and the corresponding norm $\|\cdot\|_V$. Let $a(\cdot, \cdot) : V \times V \to \mathbb{R}$ be a *bilinear form* in the sense that

$$a(u, \alpha v + \beta w) = \alpha a(u, v) + \beta a(u, w) \ ,$$
$$a(\alpha u + \beta v, w) = \alpha a(u, w) + \beta a(v, w) \ ,$$

for $\alpha, \beta \in \mathbb{R}$, $u, v, w \in V$. Also, assume that $L : V \to \mathbb{R}$ is a linear functional. We define the functional $F : V \to \mathbb{R}$ by

$$F(v) = \frac{1}{2} a(v, v) - L(v), \qquad v \in V \ .$$

We now consider the abstract *minimization problem*

$$\text{Find } p \in V \text{ such that } F(p) \leq F(v) \qquad \forall v \in V \ , \tag{1.37}$$

and the abstract *variational problem*

$$\text{Find } p \in V \text{ such that } a(p,v) = L(v) \qquad \forall v \in V \ . \tag{1.38}$$

To analyze (1.37) and (1.38), we need some properties of a and L:

- $a(\cdot,\cdot)$ is *symmetric* if

$$a(u,v) = a(v,u) \qquad \forall u,\ v \in V \ . \tag{1.39}$$

- $a(\cdot,\cdot)$ is *continuous* or *bounded* in the norm $\|\cdot\|_V$ if there is a constant $a^* > 0$ such that

$$|a(u,v)| \leq a^* \|u\|_V \|v\|_V \qquad \forall u,\ v \in V \ . \tag{1.40}$$

- $a(\cdot,\cdot)$ is V-*elliptic* or *coercive* if there exists a constant $a_* > 0$ such that

$$|a(v,v)| \geq a_* \|v\|_V^2 \qquad \forall v \in V \ . \tag{1.41}$$

- L is bounded in the norm $\|\cdot\|_V$:

$$|L(v)| \leq \tilde{L} \|v\|_V \qquad \forall v \in V \ . \tag{1.42}$$

The following theorem is needed in the proof of Theorem 1.1 below (Conway, 1985).

Theorem (Riesz Representation Theorem). *Let H be a Hilbert space with the scalar product $(\cdot,\cdot)_H$. Then, for any continuous linear functional $\mathcal{L}$ on H there is a unique $u \in H$ such that*

$$\mathcal{L}(v) = (u,v)_H \ .$$

We now prove the next theorem.

Theorem 1.1 (Lax-Milgram). *Under assumptions* (1.39)–(1.42), *problem* (1.38) *has a unique solution $p \in V$, which satisfies the bound*

$$\|p\|_V \leq \frac{\tilde{L}}{a_*} \ . \tag{1.43}$$

Proof. Since the bilinear form a is symmetric and V-elliptic, it induces a scalar product in V:

$$[u,v] = a(u,v), \qquad u,\ v \in V \ .$$

Moreover, by (1.40) and (1.41), we see that

$$a_*\|v\|_V^2 \le [v,v] \le a^*\|v\|_V^2 \qquad \forall v \in V \ .$$

That is, the norm induced by $[\cdot,\cdot]$ is equivalent to $\|\cdot\|_V$. To this new norm, L is still a continuous linear functional. Thus, according to the Riesz Representation Theorem, there is a unique $p \in V$ such that

$$[p,v] = L(v) \qquad \forall v \in V;$$

i.e.,

$$a(p,v) = L(v) \qquad \forall v \in V \ ,$$

which is (1.38). To show stability, we take $v = p$ in (1.38) and use (1.41) and (1.42) to see that

$$a_*\|p\|_V^2 \le a(p,p) = L(p) \le \tilde{L}\|p\|_V \ ,$$

which yields (1.43). □

Under assumptions (1.39)–(1.42), it can be shown in the same way as in Sect. 1.1.1 that problems (1.37) and (1.38) are equivalent. One can check that (1.38) still possesses a unique solution even without the symmetry assumption (1.39) (Ciarlet, 1978). In this case, however, there is no corresponding minimization problem.

1.3.2 The Finite Element Method

Suppose that V_h is a finite element (finite dimensional) subspace of V. The respective discrete counterparts of (1.37) and (1.38) are

$$\text{Find } p_h \in V_h \text{ such that } F(p_h) \le F(v) \qquad \forall v \in V_h \ , \tag{1.44}$$

and

$$\text{Find } p_h \in V_h \text{ such that } a(p_h,v) = L(v) \qquad \forall v \in V_h \ . \tag{1.45}$$

Theorem 1.1 remains valid for problems (1.44) and (1.45) under assumptions (1.39)–(1.42). Moreover, the solution $p_h \in V_h$ satisfies

$$\|p_h\|_V \le \frac{\tilde{L}}{a_*} \ . \tag{1.46}$$

Let $\{\varphi_i\}_{i=1}^M$ be a basis of V_h, where M is the dimension of V_h. We choose $v = \varphi_j$ in (1.45) to give

$$a(p_h,\varphi_j) = L(\varphi_j), \qquad j = 1,2,\ldots,M \ . \tag{1.47}$$

Set

$$p_h = \sum_{i=1}^{M} p_i \varphi_i, \quad p_i \in \mathbb{R} ,$$

and substitute it into (1.47) to give

$$\sum_{i=1}^{M} a(\varphi_i, \varphi_j) p_i = L(\varphi_j), \quad j = 1, 2, \ldots, M .$$

In matrix form, it is

$$\mathbf{A}\mathbf{p} = \mathbf{L}, \tag{1.48}$$

where

$$\begin{aligned} &\mathbf{A} = (a_{ij}), \qquad \mathbf{p} = (p_i), \qquad \mathbf{L} = (L_i) , \\ &a_{ij} = a(\varphi_i, \varphi_j), \quad L_i = L(\varphi_i), \quad i,\ j = 1, 2, \ldots, M . \end{aligned}$$

Theorem 1.2. *Under assumptions* (1.39) *and* (1.41), *the stiffness matrix* $\mathbf{A}$ *is symmetric and positive definite.*

Proof. Since $a(\varphi_i, \varphi_j) = a(\varphi_j, \varphi_i)$ by (1.39), $\mathbf{A}$ is obviously symmetric. For $\mathbf{v} = (v_i) \in \mathbb{R}^M$, define

$$v = \sum_{i=1}^{M} v_i \varphi_i \in V_h .$$

Then we find that

$$a(v, v) = \sum_{i,j=1}^{M} v_i a(\varphi_i, \varphi_j) v_j = \mathbf{v}^T \mathbf{A} \mathbf{v} .$$

If $\mathbf{v} \neq \mathbf{0}$ and thus $v \neq 0$, it follows from (1.41) that

$$\mathbf{v}^T \mathbf{A} \mathbf{v} \geq a_* \|v\|_V^2 > 0 ,$$

so $\mathbf{A}$ is positive definite. □

We now prove an error estimate (*Céa's Lemma*).

Theorem 1.3. *Under assumptions* (1.39)–(1.42), *if* p *and* p_h *are the respective solutions to* (1.38) *and* (1.45), *then*

$$\|p - p_h\|_V \leq \frac{a^*}{a_*} \|p - v\|_V \qquad \forall v \in V_h . \tag{1.49}$$

Proof. Because $V_h \subset V$, we subtract (1.45) from (1.38) to see that

$$a(p - p_h, w) = 0 \qquad \forall w \in V_h . \tag{1.50}$$

Using (1.40), (1.41), and (1.50), for any $v \in V_h$ it follows that

$$\begin{aligned} a_* \|p - p_h\|_V^2 &\leq a(p - p_h, p - p_h) = a(p - p_h, p - v) \\ &\leq a^* \|p - p_h\|_V \|p - v\|_V , \end{aligned}$$

which implies (1.49). □

1.3.3 Examples

We apply the theory developed in Sects. 1.3.1 and 1.3.2 to some concrete examples. More applications will be provided in subsequent chapters. These examples will employ finite element spaces V_h that consist of piecewise polynomials on partitions or triangulations $K_h = \{K\}$ of $\Omega \subset \mathbb{R}^d$ $(d = 1, 2, 3)$ into *elements* K. For $d = 1$, the elements K will be intervals; for $d = 2$, they will be triangles or quadrilaterals; for $d = 3$, they will be tetrahedra, rectangular parallelepipeds, or prisms (cf. Sect. 1.4). We will need certain *regularity* on the functions in V_h, depending on second- or fourth-order differential problems studied. For the spaces introduced in Sect. 1.1, for example, the functions in V_h are required to be continuous on $\bar{\Omega}$. The continuity requirement for V_h is related to the space $H^1(\Omega)$. As a matter of fact, if V_h is composed of piecewise polynomials, one can show that $V_h \subset H^1(\Omega)$ if and only if $V_h \subset C^0(\bar{\Omega})$. That is, $V_h \subset H^1(\Omega)$ if and only if the functions in V_h are continuous on $\bar{\Omega}$. Similarly, $V_h \subset H^2(\Omega)$ if and only if $V_h \subset C^1(\bar{\Omega})$; i.e., $V_h \subset H^2(\Omega)$ if and only if the functions in V_h and their first partial derivatives are continuous on $\bar{\Omega}$. These equivalences give the *regularity requirement* on the functions in V_h (cf. Exercise 1.17).

Example 1.1. Let us return to the one-dimensional problem (1.1). Define $I = (0, 1)$, $V = H_0^1(I)$, $\|\cdot\|_V = \|\cdot\|_{H^1(I)}$, and

$$a(v, w) = \left(\frac{dv}{dx}, \frac{dw}{dx}\right), \quad L(v) = (f, v), \qquad v,\ w \in V\ ,$$

where $f \in L^2(I)$ is given. We now check conditions (1.39)–(1.42). First, it is obvious that $a(\cdot, \cdot)$ is symmetric. Second, by Cauchy's inequality (1.10), note that

$$|a(v, w)| \le \left\|\frac{dv}{dx}\right\|_{L^2(I)} \left\|\frac{dw}{dx}\right\|_{L^2(I)} \le \|v\|_{H^1(I)} \|w\|_{H^1(I)}\ ,$$

for $v,\ w \in V$, so (1.40) holds with $a^* = 1$. Third, using Poincaré's inequality (1.34) in one dimension, we observe that

$$a(v, v) = \left\|\frac{dv}{dx}\right\|^2_{L^2(I)} \ge \frac{1}{2}\left(\|v\|^2_{L^2(I)} + \left\|\frac{dv}{dx}\right\|^2_{L^2(I)}\right)\ ,$$

so (1.41) is true with $a_* = 1/2$. Finally, by (1.10),

$$|L(v)| \le \|f\|_{L^2(I)} \|v\|_{L^2(I)} \le \|f\|_{L^2(I)} \|v\|_{H^1(I)}\,;$$

i.e., (1.42) is satisfied with $\tilde{L} = \|f\|_{L^2(I)}$. Thus Theorem 1.1 applies to problem (1.1).

For $h > 0$, let K_h be a partition of $(0, 1)$ into subintervals as in Sect. 1.1.1. Associated with K_h, let $V_h \subset V$ be the space of piecewise linear polynomials introduced in Sect. 1.1.1. Applying Theorem 1.3 and the approximation

analysis given in Sect. 1.1.1, we see that if the solution p is in the space $H^2(I)$, then

$$\|p - p_h\|_{L^2(I)} + h|p - p_h|_{H^1(I)} \leq Ch^2|p|_{H^2(I)} \ , \tag{1.51}$$

where p and p_h are the respective solutions of (1.1) and (1.5).

Example 1.2. We now consider the Poisson equation (1.16) in two dimensions. For a polygon $\Omega \subset \mathbb{R}^2$, set $V = H_0^1(\Omega)$, $\|\cdot\|_V = \|\cdot\|_{H^1(\Omega)}$, and

$$a(v,w) = (\nabla v, \nabla w) \, , \quad L(v) = (f,v), \qquad v, \ w \in V \ ,$$

where $f \in L^2(\Omega)$ is given. Conditions (1.39), (1.40), and (1.42) can be verified in the same fashion as in the previous example: $a^* = 1$ and $\tilde{L} = \|f\|_{L^2(\Omega)}$. We use the two-dimensional version of Poincaré's inequality (1.36) to prove (1.41) with

$$a_* = \frac{1}{2} \min \left\{ \frac{1}{C^2}, 1 \right\} \ ,$$

where C is given by (1.36). Therefore, Theorem 1.1 applies to problem (1.16).

For $h > 0$, let K_h be a triangulation of Ω into triangles, as defined in Sect. 1.1.2. For $K \in K_h$, as previously, we define the mesh parameters

$$h_K = \text{diam}(K) = \text{the longest edge of } \bar{K}, \ h = \max_{K \in K_h} \text{diam}(K) \ .$$

We also need the quantity

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

We say that a triangulation is *regular* if there is a constant β_1, independent of h, such that

$$\frac{h_K}{\rho_K} \leq \beta_1 \qquad \forall K \in K_h \ . \tag{1.52}$$

This condition says that the triangles in K_h are not arbitrarily thin, or equivalently, the angles of the triangles are not arbitrarily small. The constant β_1 is a measure of the smallest angle over all $K \in K_h$.

The finite element space V_h is defined as in Sect. 1.1.2; i.e.,

$$V_h = \left\{ v \in H^1(\Omega) : \ v\big|_K \in P_1(K), \, K \in K_h, \text{ and } v\big|_\Gamma = 0 \right\} \ ,$$

where $P_1(K)$ is the set of linear polynomials on K. If the solution p to the Poisson equation (1.16) is in $H^2(\Omega)$, the approximation theory in Sect. 1.9 will yield an error estimate of the form

$$\|p - p_h\|_{H^1(\Omega)} \leq Ch|p|_{H^2(\Omega)}, \tag{1.53}$$

where the constant C depends only on β_1 in (1.52), but not on h or p. To state an estimate in the L^2-norm, we require that the polygonal domain Ω be convex. In the convex case, it holds that (cf. Sect. 1.9)

$$\|p - p_h\|_{L^2(\Omega)} \le Ch^2 |p|_{H^2(\Omega)} \ . \tag{1.54}$$

Estimates (1.53) and (1.54) are *optimal* (i.e., the estimates with the largest power of h one can obtain between the exact solution and its approximate solution). Instead of a polygonal domain, if the boundary Γ of Ω is smooth, convexity is not required for (1.54).

Example 1.3. The analysis in the previous example can be extended to a more general second-order problem:

$$\begin{aligned} -\nabla \cdot (\mathbf{a}\nabla p) + \boldsymbol{\beta} \cdot \nabla p + cp = f \qquad & \text{in } \Omega \ , \\ p = 0 \qquad & \text{on } \Gamma \ , \end{aligned} \tag{1.55}$$

where $\mathbf{a}$ is a 2×2 matrix, $\boldsymbol{\beta}$ is a constant vector, and c is a bounded, nonnegative function. They, together with $f \in L^2(\Omega)$, are given functions. Assume that $\mathbf{a}$ satisfies

$$0 < a_* \le |\boldsymbol{\eta}|^2 \sum_{i,j=1}^{2} a_{ij}(\mathbf{x}) \eta_i \eta_j \le a^* < \infty, \quad \mathbf{x} \in \Omega, \ \boldsymbol{\eta} \ne \mathbf{0} \in \mathbb{R}^2 \ .$$

This problem is an example of a *convection-diffusion-reaction problem*; the first term corresponds to diffusion with the diffusion coefficient $\mathbf{a}$, the second term to convection in the direction $\boldsymbol{\beta}$, and the third term to reaction with the coefficient c. We here consider the case where the size of $|\boldsymbol{\beta}|$ is moderate. For convection- or advection-dominated problems, the reader should refer to Chaps. 4 and 5. Many problems arise in form (1.55), e.g., the problems from multiphase flows in porous media and semiconductor modeling (cf. Chaps. 9 and 10).

Define $V = H_0^1(\Omega)$, $\|\cdot\|_V = \|\cdot\|_{H^1(\Omega)}$, and

$$\begin{aligned} a(v, w) &= (\mathbf{a}\nabla v, \nabla w) + (\boldsymbol{\beta} \cdot \nabla v, w) + (cv, w) \ , \\ L(v) &= (f, v), \qquad v, \ w \in V \ . \end{aligned}$$

Note that $a(\cdot, \cdot)$ is not symmetric due to the presence of $\boldsymbol{\beta}$, so (1.39) does not hold. However, conditions (1.40)–(1.42) do hold. The proof of (1.40) and (1.42) is the same as in the previous example. To verify (1.41), by (1.19) we see that

$$(\boldsymbol{\beta} \cdot \nabla v, v) = (\boldsymbol{\beta} \cdot \boldsymbol{\nu} v, v)_\Gamma - (v, \boldsymbol{\beta} \cdot \nabla v) \ ,$$

so, by the fact that $v|_\Gamma = 0$,

$$(\boldsymbol{\beta} \cdot \nabla v, v) = 0 \ .$$

Hence we obtain

$$a(v, v) = (\mathbf{a}\nabla v, \nabla v) + (cv, v), \qquad v \in V \ .$$

Consequently, in the same manner as in Example 1.2, (1.41) follows from the assumptions on **a** and c.

Following a remark after Theorem 1.1, the variational problem associated with (1.55) has a unique solution. Also, the corresponding discrete problem and its error estimates can be stated as in the previous example.

While we have considered only the Dirichlet boundary condition in these three examples, boundary conditions of other types can be analyzed as in Sect. 1.1.3 (cf. Exercises 1.20 and 1.21).

Example 1.4. In this example, we consider a *fourth-order* problem in one dimension:

$$\begin{aligned} &\frac{d^4p}{dx^4} = f(x), \quad 0 < x < 1 \,, \\ &p(0) = p(1) = \frac{dp}{dx}(0) = \frac{dp}{dx}(1) = 0 \,, \end{aligned} \tag{1.56}$$

where $f \in L^2(I)$ is given. With $I = (0,1)$, define

$$V = H_0^2(I) = \left\{ v \in H^2(I) : v(0) = v(1) = \frac{dv}{dx}(0) = \frac{dv}{dx}(1) = 0 \right\} \,,$$

with the norm $\|\cdot\|_V = \|\cdot\|_{H^2(I)}$. Also, set

$$a(v,w) = \left(\frac{d^2v}{dx^2}, \frac{d^2w}{dx^2} \right), \quad L(v) = (f,v), \qquad v,\ w \in V \,.$$

From Poincaré's inequality (1.34), we can deduce that

$$\|v\|_{L^2(I)} \le \left\| \frac{dv}{dx} \right\|_{L^2(I)} \le \left\| \frac{d^2v}{dx^2} \right\|_{L^2(I)} \qquad \forall v \in V \,.$$

As a result, this example can be analyzed in the same way as for Example 1.1.

As in Sect. 1.1.1, let $K_h : 0 = x_0 < x_1 < \ldots < x_M < x_{M+1} = 1$ be a partition of I into subintervals $I_i = (x_{i-1}, x_i)$, with length $h_i = x_i - x_{i-1}$, $i = 1, 2, \ldots, M+1$. Set $h = \max\{h_i : i = 1, 2, \ldots, M+1\}$. Introduce the finite element space

$$\begin{aligned} V_h = \Big\{ &v : v \text{ and } \frac{dv}{dx} \text{ are continuous on } I, v \text{ is a polynomial} \\ &\text{of degree 3 on each subinterval } I_i \,, \\ &\text{and } v(0) = v(1) = \frac{dv}{dx}(0) = \frac{dv}{dx}(1) = 0 \Big\} \,. \end{aligned}$$

As parameters, or *degrees of freedom*, to describe the functions $v \in V_h$, we can use the values and first derivatives of v at the nodes $\{x_i\}_{i=1}^{M+1}$ of K_h. It can be shown that this is a legitimate choice; that is, a function in V_h is

uniquely determined by these degrees of freedom (see Example 1.13 in the next section). Now, the finite element method for (1.56) can be defined as in (1.45). Moreover, if the solution p to (1.56) is in $H^3(I)$, the following error estimate holds (cf. Sect. 1.9):

$$\|p - p_h\|_V \leq Ch|p|_{H^3(I)} \ .$$

Note that because the first derivative of $v \in V_h$ is required to be continuous on I, it has at least four degrees of freedom on each subinterval in K_h. Thus the degree of v must be greater than or equal to three.

Example 1.5. In this example, we extend the one-dimensional fourth-order problem to two dimensions, i.e., to the *biharmonic problem*:

$$\begin{aligned} \Delta^2 p &= f && \text{in } \Omega \ , \\ p = \frac{\partial p}{\partial \boldsymbol{\nu}} &= 0 && \text{on } \Gamma \ , \end{aligned} \tag{1.57}$$

where $\Delta^2 = \Delta\Delta$. This problem is a formulation of the Stokes equation in fluid mechanics (cf. Exercise 8.2). It also models the displacement of a thin elastic plate, *clamped* at its boundary and under a transversal load of intensity f; see Fig. 1.12, where the thin elastic plate has a surface given by $\Omega \subset \mathbb{R}^2$ (refer to Chap. 7). The first boundary condition $p|_\Gamma = 0$ says that the displacement p is held fixed (at the zero height) at the boundary Γ, while the second condition $\partial p/\partial \boldsymbol{\nu}|_\Gamma = 0$ means that the rotation of the plate is also prescribed at Γ. These boundary conditions thus imply that the plate is *clamped*.

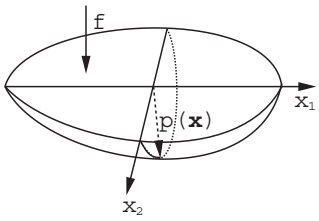

Fig. 1.12. An elastic plate

We introduce the space

$$V = H_0^2(\Omega) = \left\{ v \in H^2(\Omega) : v = \frac{\partial v}{\partial \boldsymbol{\nu}} = 0 \text{ on } \Gamma \right\} \ ,$$

with the norm $\|\cdot\|_V = \|\cdot\|_{H^2(\Omega)}$. With this definition and Green's formula (1.19), we see that

$$
\begin{aligned}
(\Delta^2 p, v) &= \left(\frac{\partial \Delta p}{\partial \boldsymbol{\nu}}, v\right)_\Gamma - (\nabla \Delta p, \nabla v) \\
&= -\left(\Delta p, \frac{\partial v}{\partial \boldsymbol{\nu}}\right)_\Gamma + (\Delta p, \Delta v) \\
&= (\Delta p, \Delta v)\,, \qquad v \in V\,.
\end{aligned}
$$

We thus define

$$
a(v, w) = (\Delta v, \Delta w)\,, \qquad L(v) = (f, v), \quad v,\ w \in V\,.
$$

Conditions (1.39), (1.40), and (1.42) can be easily seen, while (1.41) follows from the inequality (Girault-Raviart, 1981)

$$
\|v\|_{H^2(\Omega)} \le C \|\Delta v\|_{L^2(\Omega)} \qquad \forall v \in V\,.
$$

Let K_h be a triangulation of Ω into triangles as in Example 1.2. Associated with K_h, we define the finite element space

$$
V_h = \left\{ v : v \text{ and } \nabla v \text{ are continuous on } \Omega, v \text{ is a polynomial of} \right.
$$
$$
\left. \text{degree 5 on each triangle, and } v = \frac{\partial v}{\partial \boldsymbol{\nu}} = 0 \text{ on } \Gamma \right\}\,.
$$

This space is often known as the *Argyris triangle*. Since the first partial derivatives of the functions in V_h are required to be continuous on $\Omega \subset \mathbb{R}^2$, there are at least six degrees of freedom on each interior edge in K_h. Thus the polynomial degree must be greater than or equal to five. The degrees of freedom for describing the functions in V_h will be discussed in Example 1.13 in the next section. With this V_h, the finite element method (1.45) applies to (1.57). If the solution p to (1.57) is in $H^3(\Omega)$, there is the error estimate

$$
\|p - p_h\|_V \le Ch|p|_{H^3(\Omega)}\,.
$$

Problem (1.57) will be further considered in Chaps. 2 and 7; other spaces such as the *reduced Argyris triangle* and *Morley element* for the approximation of problem (1.57) will be studied.

1.4 Finite Element Spaces

1.4.1 Triangles

In Sects. 1.1.2 and 1.3.3, we have considered the finite element space of piecewise linear functions for the approximation of second-order partial differential equations. In this section, we consider more general finite element spaces. First, we treat the case where $\Omega \subset \mathbb{R}^2$ is a polygonal domain in the plane. Let

K_h be a triangulation of Ω into triangles K as in Sect. 1.1.2. We introduce the notation

$$P_r(K) = \{v : \ v \text{ is a polynomial of degree at most } r \text{ on } K\} \ ,$$

where $r = 0, 1, 2, \ldots$. For $r = 1$, $P_1(K)$ is the space of *linear* functions, used previously, of the form

$$v(\mathbf{x}) = v_{00} + v_{10}x_1 + v_{01}x_2, \quad \mathbf{x} = (x_1, x_2) \in K, \, v \in P_1(K) \ ,$$

where $v_{ij} \in \mathbb{R}$, $i, j = 0, 1$. Note that $\dim(P_1(K)) = 3$; i.e., its *dimension* is three.

For $r = 2$, $P_2(K)$ is the space of *quadratic* functions on K:

$$v(\mathbf{x}) = v_{00} + v_{10}x_1 + v_{01}x_2 + v_{20}x_1^2 + v_{11}x_1x_2 + v_{02}x_2^2, \quad v \in P_2(K) \ ,$$

where $v_{ij} \in \mathbb{R}$, $i, j = 0, 1, 2$. We see that $\dim(P_2(K)) = 6$.

In general, we have

$$P_r(K) = \left\{ v : \ v(\mathbf{x}) = \sum_{0 \le i+j \le r} v_{ij} x_1^i x_2^j, \, \mathbf{x} \in K, \, v_{ij} \in \mathbb{R} \right\}, \ r \ge 0 \ ,$$

so

$$\dim(P_r(K)) = \frac{(r+1)(r+2)}{2} \ .$$

Example 1.6. Define

$$V_h = \left\{ v : \ v \text{ is continuous on } \Omega \text{ and } v\big|_K \in P_1(K), \, K \in K_h \right\} \ ,$$

where $v\big|_K$ represents the restriction of v to K. As parameters, or *global degrees of freedom*, to describe the functions in V_h, we use the values at the vertices (nodes) of K_h. It can be shown that this is a legitimate choice; that is, a function in V_h is uniquely determined by these global degrees of freedom. To see this, for each triangle $K \in K_h$, let its vertices be indicated by $\mathbf{m}_1$, $\mathbf{m}_2$, and $\mathbf{m}_3$; see Fig. 1.13. Also, let the (local) basis functions of $P_1(K)$ be λ_i, $i = 1, 2, 3$, which are defined by

$$\lambda_i(\mathbf{m}_j) = \begin{cases} 1 & \text{if } i = j \ , \\ 0 & \text{if } i \ne j, \end{cases} \qquad i, j = 1, 2, 3 \ .$$

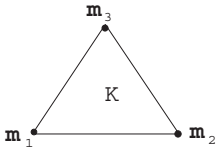

Fig. 1.13. The element degrees of freedom for $P_1(K)$

These basis functions can be determined in the following approach: Let an equation of the straight line through the vertices $\mathbf{m}_2$ and $\mathbf{m}_3$ be given by

$$c_0 + c_1 x_1 + c_2 x_2 = 0 \ ,$$

and then define

$$\lambda_1(\mathbf{x}) = \gamma(c_0 + c_1 x_1 + c_2 x_2), \quad \mathbf{x} = (x_1, x_2) \ ,$$

where the constant γ is chosen such that $\lambda_1(\mathbf{m}_1) = 1$. The functions λ_2 and λ_3 can be determined in the same approach. These functions λ_1, λ_2, and λ_3 are sometimes called the *barycentric coordinates* of a triangle. If K is the reference triangle with vertices $(1,0)$, $(0,1)$, and $(0,0)$, λ_1, λ_2, and λ_3 are, respectively, x_1, x_2, and $1 - x_1 - x_2$. Now, any function $v \in P_1(K)$ has the unique representation

$$v(\mathbf{x}) = \sum_{i=1}^{3} v(\mathbf{m}_i)\lambda_i(\mathbf{x}), \qquad \mathbf{x} \in K \ .$$

Thus $v \in P_1(K)$ is uniquely determined by its values at the three vertices. Therefore, on each triangle $K \in K_h$, the degrees of freedom, *element degrees of freedom*, can be these (nodal) values. These degrees of freedom are the same as the global degrees of freedom and are used to construct the basis functions in V_h (cf. Sect. 1.1.2).

We claim that for v such that $v|_K \in P_1(K)$, $K \in K_h$, if it is continuous at internal vertices, then $v \in C^0(\bar{\Omega})$. Obviously, it suffices to show that v is continuous across all interelement edges. Let triangles K_1, $K_2 \in K_h$ share a common edge e with the end points $\mathbf{m}_1$ and $\mathbf{m}_2$, and set $v_i = v|_{K_i} \in P_1(K_i)$, $i = 1, 2$. Then the difference $v_1 - v_2$ defined on e vanishes at $\mathbf{m}_1$ and $\mathbf{m}_2$. Because $v_1 - v_2$ is linear on e (in either x_1 or x_2), it vanishes on the entire edge e. Thus v is continuous across e, and we proved the claim that $v \in C^0(\bar{\Omega})$.

For a problem with an essential boundary condition, this condition needs to be incorporated into the definition of V_h as in Sect. 1.1.2. This remark also applies to the examples below.

Example 1.7. Let

$$V_h = \left\{ v : \ v \text{ is continuous on } \Omega \text{ and } v\big|_K \in P_2(K),\ K \in K_h \right\} \ .$$

Namely, V_h is the space of continuous piecewise quadratic functions. The global degrees of freedom of a function $v \in V_h$ are chosen by the values of v at the vertices and the midpoints of edges in K_h. It can be seen that v is uniquely defined by these degrees of freedom. For each $K \in K_h$, the element degrees of freedom are shown in Fig. 1.14, where the midpoints of edges of K are denoted by $\mathbf{m}_{ij}$, $i < j$, $i, j = 1, 2, 3$.

In fact, because $\dim(P_2(K))$ equals the number of degrees of freedom (6), it suffices to show that if $v \in P_2(K)$ satisfies

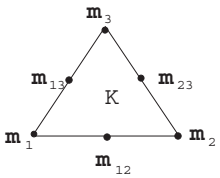

Fig. 1.14. The element degrees of freedom for $P_2(K)$

$$v(\mathbf{m}_i) = 0, \quad v(\mathbf{m}_{ij}) = 0, \qquad i < j,\ i,\ j = 1, 2, 3\ ,$$

then $v \equiv 0$. For this, consider edge $\mathbf{m}_2\mathbf{m}_3$. Since $v \in P_2(K)$ is quadratic (in a single variable) on this edge and vanishes at three distinct points $\mathbf{m}_2$, $\mathbf{m}_{23}$, and $\mathbf{m}_3$, then $v \equiv 0$ on $\mathbf{m}_2\mathbf{m}_3$ (cf. Exercise 1.24) and it can be written (cf. Exercise 1.25) as

$$v(\mathbf{x}) = \lambda_1(\mathbf{x}) w(\mathbf{x}), \qquad \mathbf{x} \in K\ ,$$

where $w \in P_1(K)$. Similarly, we can show that $v \equiv 0$ on edge $\mathbf{m}_1\mathbf{m}_3$ and

$$v(\mathbf{x}) = \lambda_1(\mathbf{x})\lambda_2(\mathbf{x}) w_0, \qquad \mathbf{x} \in K\ ,$$

where w_0 is a constant. Note that

$$0 = v(\mathbf{m}_{12}) = \frac{1}{2} \cdot \frac{1}{2}\ w_0\ ;$$

i.e., $w_0 = 0$. Therefore, $v \equiv 0$ on K.

It can be seen (cf. Exercise 1.26) that a function $v \in P_2(K)$ has the representation

$$\begin{aligned} v(\mathbf{x}) = &\sum_{i=1}^{3} v(\mathbf{m}_i)\lambda_i(\mathbf{x})\big(2\lambda_i(\mathbf{x}) - 1\big) \\ &+ \sum_{i,j=1;\ i<j}^{3} 4v(\mathbf{m}_{ij})\lambda_i(\mathbf{x})\lambda_j(\mathbf{x}), \qquad \mathbf{x} \in K\ . \end{aligned} \tag{1.58}$$

Also, as in Example 1.6, we can prove that if v is continuous at the internal vertices and midpoints of edges and $v \in P_2(K)$, $K \in K_h$, then $v \in C^0(\bar{\Omega})$.

Example 1.8. Set

$$V_h = \big\{v :\ v \text{ is continuous on } \Omega \text{ and } v\big|_K \in P_3(K),\ K \in K_h\big\}\ .$$

That is, V_h is the space of continuous piecewise cubic functions. Let $K \in K_h$ have vertices $\mathbf{m}_i$, $i = 1, 2, 3$. Define, for $i,\ j = 1, 2, 3,\ i \neq j$,

$$\mathbf{m}_0 = \frac{1}{3}\left(\mathbf{m}_1 + \mathbf{m}_2 + \mathbf{m}_3\right), \quad \mathbf{m}_{i,j} = \frac{1}{3}\left(2\mathbf{m}_i + \mathbf{m}_j\right)\ ,$$

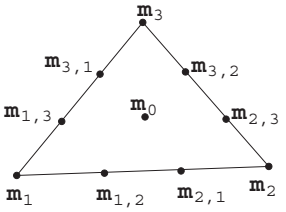

Fig. 1.15. The element degrees of freedom for $P_3(K)$

where $\mathbf{m}_0$ is the *center of gravity* of K (centroid); see Fig. 1.15. It can be proven that a function $v \in P_3(K)$ is uniquely determined by the following values:

$$v(\mathbf{m}_i), \quad v(\mathbf{m}_0), \quad v(\mathbf{m}_{i,j}), \quad i, j = 1, 2, 3,\ i \neq j\ .$$

These values can be used as the degrees of freedom.

In fact, because $\dim(P_3(K))$ equals the number of degrees of freedom (10), it is sufficient to prove that if $v \in P_3(K)$ satisfies

$$v(\mathbf{m}_0) = v(\mathbf{m}_i) = v(\mathbf{m}_{i,j}) = 0, \quad i, j = 1, 2, 3,\ i \neq j\ ,$$

then $v \equiv 0$ on K. Indeed, for such a v, it can be seen as in Example 1.7 that v has the representation

$$v(\mathbf{x}) = \lambda_1(\mathbf{x})\lambda_2(\mathbf{x})\lambda_3(\mathbf{x})w_0, \qquad \mathbf{x} \in K\ ,$$

where w_0 is a constant. Since

$$0 = v(\mathbf{m}_0) = \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{3}\ w_0\ ,$$

we see that $w_0 = 0$, and thus $v \equiv 0$ on K.

Example 1.9. The degrees of freedom for $P_3(K)$ (and thus for V_h) can be chosen in a different way. A function $v \in P_3(K)$ is also uniquely defined by (cf. Fig. 1.16)

$$v(\mathbf{m}_i), \quad v(\mathbf{m}_0), \quad \frac{\partial v}{\partial x_j}(\mathbf{m}_i), \quad i = 1, 2, 3,\ j = 1, 2\ .$$

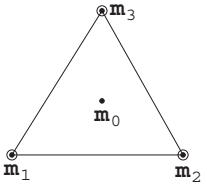

Fig. 1.16. The second set of degrees of freedom for $P_3(K)$

In fact, it suffices to show that if $v \in P_3(K)$ satisfies

$$v(\mathbf{m}_0) = v(\mathbf{m}_i) = \frac{\partial v}{\partial x_j}(\mathbf{m}_i) = 0, \quad i = 1,2,3,\ j = 1,2\ , \tag{1.59}$$

then $v \equiv 0$ on K. Using (1.59), we see that

$$\frac{\partial v}{\partial \mathbf{t}}(\mathbf{m}_i) = \frac{\partial v}{\partial x_1}(\mathbf{m}_i)t_1 + \frac{\partial v}{\partial x_2}(\mathbf{m}_i)t_2 = 0, \quad i = 1,2,3\ ,$$

where $\mathbf{t} = (t_1, t_2)$ is a tangential direction. Particularly, we see that

$$\frac{\partial v}{\partial \mathbf{t}}(\mathbf{m}_2) = \frac{\partial v}{\partial \mathbf{t}}(\mathbf{m}_3) = 0\ ,$$

where $\mathbf{t}$ is the direction from $\mathbf{m}_2$ to $\mathbf{m}_3$, which, together with $v(\mathbf{m}_2) = v(\mathbf{m}_3) = 0$ and the fact that v is cubic (in a single variable) on edge $\mathbf{m}_2\mathbf{m}_3$, implies that $v \equiv 0$ on this edge. The same argument shows that $v \equiv 0$ on edges $\mathbf{m}_1\mathbf{m}_3$ and $\mathbf{m}_1\mathbf{m}_2$. Then, in the same way as in Example 1.8, we see that $v \equiv 0$ on K.

The corresponding finite element space $V_h \subset C^0(\bar{\Omega})$ is defined by

$$V_h = \left\{ v : v \text{ and } \frac{\partial v}{\partial x_i}\ (i = 1,2) \text{ are continuous at vertices of } K_h;\ v\big|_K \in P_3(K),\ K \in K_h \right\}.$$

We have considered the cases $r \leq 3$. In general, for any $r \geq 1$, we define

$$V_h = \left\{ v :\ v \text{ is continuous on } \Omega \text{ and } v\big|_K \in P_r(K),\ K \in K_h \right\}\ .$$

A function $v \in P_r(K)$ can be uniquely determined by its values at the three vertices, $3(r-1)$ distinct points on the edges, and $(r-1)(r-2)/2$ interior points in K. The values at these points can be employed as the degrees of freedom in V_h.

1.4.2 Rectangles

We now consider the case where Ω is a rectangular domain and K_h is a partition of Ω into non-overlapping rectangles such that the horizontal and vertical edges of rectangles are parallel to the x_1- and x_2-coordinate axes, respectively. We also require that no vertex of any rectangle lie in the interior of an edge of another rectangle. We introduce the notation

$$Q_r(K) = \left\{ v :\ v(\mathbf{x}) = \sum_{i,j=0}^{r} v_{ij} x_1^i x_2^j,\ \mathbf{x} \in K,\ v_{ij} \in \mathbb{R} \right\}, \quad r \geq 0\ .$$

Note that $\dim(Q_r(K)) = (r+1)^2$.

For $r = 1$, we define

$$V_h = \left\{ v : \ v \text{ is continuous on } \Omega \text{ and } v\big|_K \in Q_1(K),\ K \in K_h \right\} .$$

A function $v \in Q_1(K)$ is *bilinear* and of the form

$$v(\mathbf{x}) = v_{00} + v_{10}x_1 + v_{01}x_2 + v_{11}x_1x_2, \quad \mathbf{x} = (x_1, x_2) \in K,\ v_{ij} \in \mathbb{R} .$$

As in the triangular case, it can be checked that v is uniquely determined by its values at the four vertices of K, which can be chosen as the degrees of freedom for V_h (cf. Fig. 1.17).

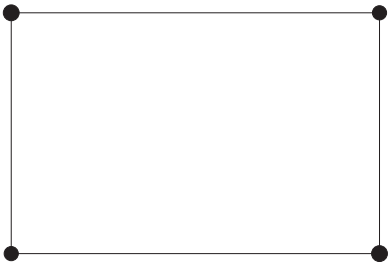

Fig. 1.17. The element degrees of freedom for $Q_1(K)$

For $r = 2$, define

$$V_h = \left\{ v : \ v \text{ is continuous on } \Omega \text{ and } v\big|_K \in Q_2(K),\ K \in K_h \right\} ,$$

where $Q_2(K)$ is the set of *biquadratic* functions on K. The degrees of freedom can be chosen by the values of functions at the vertices, midpoints of edges, and center of each rectangle (cf. Fig. 1.18). Other cases $r \geq 3$ can be analogously discussed.

The use of rectangles requires that the geometry of Ω be special. Thus it is of interest to utilize more general quadrilaterals, which will be considered in the next section, in connection with *isoparametric finite elements* (cf. Exercise 1.29).

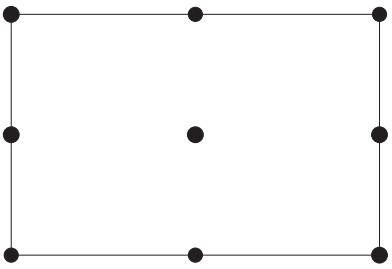

Fig. 1.18. The element degrees of freedom for $Q_2(K)$

1.4.3 Three Dimensions

Example 1.10. In three dimensions, for a polygonal domain $\Omega \subset \mathbb{R}^3$, let K_h be a partition of Ω into non-overlapping *tetrahedra* such that no vertex of any tetrahedron lies in the interior of an edge or face of another tetrahedron. For each $K \in K_h$ and $r \geq 0$, set

$$P_r(K) = \left\{ v : \ v(\mathbf{x}) = \sum_{0 \leq i+j+k \leq r} v_{ijk} x_1^i x_2^j x_3^k, \, \mathbf{x} \in K, \, v_{ijk} \in \mathbb{R} \right\} ,$$

where $\mathbf{x} = (x_1, x_2, x_3)$ and

$$\dim(P_r(K)) = \frac{(r+1)(r+2)(r+3)}{6} .$$

For $r = 1$, the function values of $v \in P_1(K)$ at the four vertices of K can be utilized as the degrees of freedom (cf. Fig. 1.19). Other cases $r \geq 2$ can be also handled.

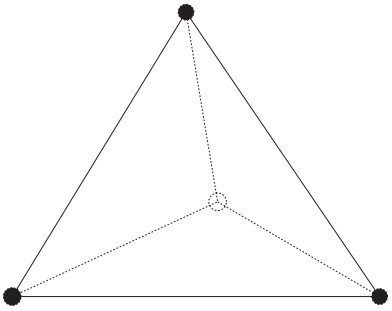

Fig. 1.19. The element degrees of freedom for $P_1(K)$ on a tetrahedron

Example 1.11. Let Ω be a rectangular domain in $\mathbb{R}^3$, and K_h be a partition of Ω into non-overlapping *rectangular parallelepipeds* such that the faces are parallel to the x_1-, x_2-, and x_3-coordinate axes, respectively. For each $K \in K_h$, the polynomials we use for V_h are of the type

$$Q_r(K) = \left\{ v : \ v(\mathbf{x}) = \sum_{i,j,k=0}^{r} v_{ijk} x_1^i x_2^j x_3^k, \, \mathbf{x} \in K, \, v_{ijk} \in \mathbb{R} \right\} , \quad r \geq 0 .$$

Note that $\dim(Q_r(K)) = (r+1)^3$. For $r = 1$, the function values of $v \in Q_1(K)$ at the eight vertices of K can be utilized as the degrees of freedom (cf. Fig. 1.20).

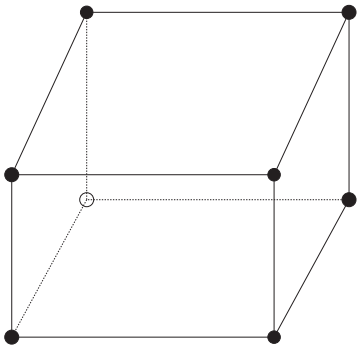

Fig. 1.20. The element degrees of freedom for $Q_1(K)$ on a cube

Example 1.12. Let $\Omega \subset \mathbb{R}^3$ be a domain of the form $\Omega = G \times [l_1, l_2]$, where $G \subset \mathbb{R}^2$ and l_1 and l_2 are real numbers. Let K_h be a partition of Ω into *prisms* such that their bases are triangles in the (x_1, x_2)-plane with three vertical edges parallel to the x_3-axis. Define $P_{l,r}$ to be the space of polynomials of degree l in the two variables x_1 and x_2 and of degree r in the variable x_3. That is, for each $K \in K_h$ and $l,\ r \geq 0$,

$$P_{l,r}(K) = \left\{ v :\ v(\mathbf{x}) = \sum_{0 \leq i+j \leq l} \sum_{k=0}^{r} v_{ijk} x_1^i x_2^j x_3^k,\ \mathbf{x} \in K,\ v_{ijk} \in \mathbb{R} \right\} .$$

Note that $\dim(P_{l,r}(K)) = (l+1)(l+2)(r+1)/2$. For $l = 1$ and $r = 1$, the function values of $v \in P_{1,1}(K)$ at the six vertices of K can be utilized as the degrees of freedom (cf. Fig. 1.21).

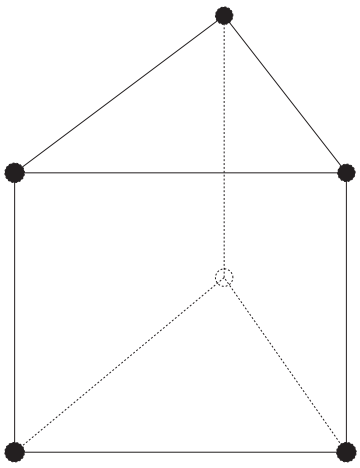

Fig. 1.21. The element degrees of freedom for $P_{1,1}(K)$ on a prism

1.4.4 A C^1 Element

In the final example, we consider a finite element space V_h that is a subspace of $C^1(\bar{\Omega})$. This space has been briefly studied in Example 1.5. As noted there, to satisfy $V_h \subset C^1(\bar{\Omega})$, we require that polynomials be degree at least five on each triangle. Special constructions of K_h must be employed to satisfy the C^1-condition if polynomials of lower degree are used (e.g., the Hsieh-Clough-Tocher element; see Ciarlet, 1978).

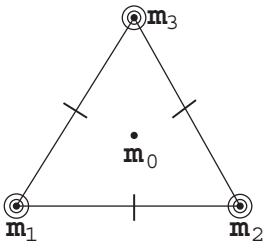

Fig. 1.22. Argyris' triangle

Example 1.13. Let K_h be a triangulation of Ω into triangles K as in Sect. 1.1.2, and define

$$V_h = \{v : \ v \text{ and } \nabla v \text{ are continuous on } \Omega; v|_K \in P_5(K),\ K \in K_h\} \ .$$

For each $K \in K_h$, its vertices and midpoints of the edges are denoted by $\mathbf{m}_i$ and $\mathbf{m}_{ij}$, $i < j$, $i, j = 1, 2, 3$; see Fig. 1.14. A function $v \in P_5(K)$ is uniquely determined by the degrees of freedom

$$\begin{aligned} &D^\alpha v(\mathbf{m}_i), \quad i = 1, 2, 3,\ |\alpha| \le 2 \ , \\ &\frac{\partial v}{\partial \boldsymbol{\nu}}(\mathbf{m}_{ij}), \quad i < j,\ i, j = 1, 2, 3 \ , \end{aligned} \tag{1.60}$$

where we recall that $\partial v/\partial \boldsymbol{\nu}$ is the normal derivative (cf. Fig. 1.22). In fact, because both $\dim(P_5(K))$ and the number of degrees of freedom are equal to 21, it suffices to show that if these degrees of freedom vanish, then $v \equiv 0$ on K. Toward that end, note that if $\mathbf{t}$ is the direction from $\mathbf{m}_2$ to $\mathbf{m}_3$, then

$$v(\mathbf{m}_i) = \frac{\partial v}{\partial \mathbf{t}}(\mathbf{m}_i) = \frac{\partial^2 v}{\partial \mathbf{t}^2}(\mathbf{m}_i) = 0, \quad i = 2, 3 \ . \tag{1.61}$$

Because v is a polynomial of degree five (in either x_1 or x_2) on edge $\mathbf{m}_2\mathbf{m}_3$, $v \equiv 0$ on this edge. Also, since

$$\frac{\partial v}{\partial \boldsymbol{\nu}}(\mathbf{m}_{23}) = \frac{\partial v}{\partial \boldsymbol{\nu}}(\mathbf{m}_i) = \frac{\partial}{\partial \mathbf{t}} \left(\frac{\partial v}{\partial \boldsymbol{\nu}} \right) (\mathbf{m}_i) = 0, \quad i = 2, 3 \ , \tag{1.62}$$

and $\partial v/\partial \boldsymbol{\nu}$ is a polynomial of degree at most 4 on edge $\mathbf{m}_2\mathbf{m}_3$, we see that $\partial v/\partial \boldsymbol{\nu} = 0$ on this edge. The fact that both v and $\partial v/\partial \boldsymbol{\nu}$ vanish on $\mathbf{m}_2\mathbf{m}_3$ implies

$$v(\mathbf{x}) = \big(\lambda_1(\mathbf{x})\big)^2 w(\mathbf{x}), \qquad \mathbf{x} \in K \,,$$

where $w \in P_3(K)$. The same argument yields

$$v(\mathbf{x}) = \big(\lambda_1(\mathbf{x})\big)^2 \big(\lambda_2(\mathbf{x})\big)^2 \big(\lambda_3(\mathbf{x})\big)^2 w_0, \qquad \mathbf{x} \in K \,,$$

where w_0 is a constant. Because $v \in P_5(K)$, the only possibility is that $w_0 = 0$, and thus $v \equiv 0$ on K.

We claim that for v such that $v|_K \in P_5(K)$, $K \in K_h$, if its degrees of freedom given by (1.60) are continuous, then $v \in C^1(\bar{\Omega})$. Let triangles K_1, $K_2 \in K_h$ share a common edge e with the end points $\mathbf{m}_2$ and $\mathbf{m}_3$ and midpoint $\mathbf{m}_{23}$, and set $v_i = v|_{K_i} \in P_1(K_i)$, $i = 1, 2$. Suppose that

$$\begin{aligned} &D^\alpha v_1(\mathbf{m}_i) = D^\alpha v_2(\mathbf{m}_i), \quad i = 2, 3, \ |\alpha| \le 2 \,, \\ &\frac{\partial v_1}{\partial \boldsymbol{\nu}}(\mathbf{m}_{23}) = \frac{\partial v_2}{\partial \boldsymbol{\nu}}(\mathbf{m}_{23}) \,. \end{aligned}$$

Then the difference $v_1 - v_2$ satisfies (1.61) and (1.62), so

$$v_1 - v_2 = \frac{\partial (v_1 - v_2)}{\partial \boldsymbol{\nu}} = 0 \quad \text{on } e \,.$$

Furthermore, if $v_1 - v_2 = 0$ on e, we see that

$$\frac{\partial (v_1 - v_2)}{\partial \mathbf{t}} = 0 \quad \text{on } e \,.$$

Therefore, v and its first partial derivatives are continuous across e, and the claim is proven.

The $C^1(\bar{\Omega})$ element introduced in this example is often known as the *Argyris triangle*. It has 21 degrees of freedom on each triangle $K \in K_h$. One can reduce this number of degrees of freedom by restricting to the class of polynomials of degree five whose normal derivatives on each edge of K are polynomials of degree three rather than four. For this class of polynomials, the normal derivative along an edge is uniquely determined by the derivatives at its endpoints (vertices). The number of degrees of freedom on each $K \in K_h$ for this *reduced Argyris triangle* (preferably, *Bell's triangle*; cf. Fig. 1.23) is 18:

$$D^\alpha v(\mathbf{m}_i), \qquad i = 1, 2, 3, \ |\alpha| \le 2 \,.$$

In summary, a *finite element* is a triple $(K, P(K), \Sigma_K)$, where K is a geometric object (i.e., element), $P(K)$ is a finite dimensional linear space of functions on K, and Σ_K is a set of degrees of freedom, such that a function $v \in P(K)$ is uniquely defined by Σ_K. For instance, in Example 1.6, K is a triangle, $P(K) = P_1(K)$, and Σ_K is the set of the values at the vertices of

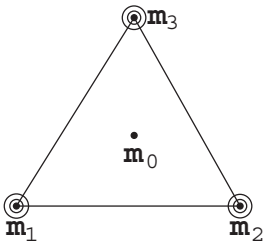

Fig. 1.23. Bell's triangle

K. When Σ_K includes the values of partial derivatives of functions, the finite element is said to be of *Hermite type*, as in Examples 1.9 and 1.13. When all degrees of freedom are given by function values, the finite element is called a *Lagrange* element.

1.5 General Domains

In the construction of finite element spaces so far, we have assumed that the domain Ω is polygonal. In this section, we consider the case where Ω is curved. For simplicity, we focus on two space dimensions.

For a two-dimensional domain Ω, the simplest approximation for its curved boundary Γ is a polygonal line; see Fig. 1.24. The resulting error due to this approximation is of order $\mathcal{O}(h^2)$, where h is the mesh size as usual; see Exercise 1.28. To obtain a more accurate approximation, we can approximate Γ with piecewise polynomials of degree $r \geq 2$. The error in this approximation becomes $\mathcal{O}(h^{r+1})$. In the partition of such an approximated domain, the elements closest to Γ then have one curved edge.

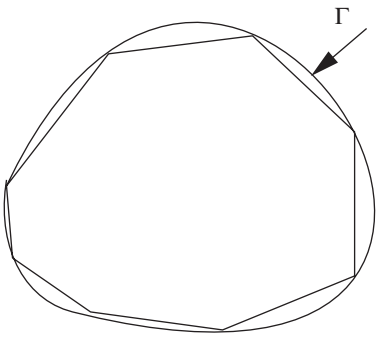

Fig. 1.24. A polygonal line approximation of Γ

As an example, let $(\hat{K}, P(\hat{K}), \Sigma_{\hat{K}})$ be a finite element, where $\hat{K}$ is the *reference triangle* with vertices $\hat{\mathbf{m}}_1 = (0,0)$, $\hat{\mathbf{m}}_2 = (1,0)$, and $\hat{\mathbf{m}}_3 = (0,1)$ in the $\hat{\mathbf{x}}$-plane. Furthermore, assume that this element is of the Lagrange type; that is, all degrees of freedom are defined by the function values at certain points $\hat{\mathbf{m}}_i$, $i = 1, 2, \ldots, l$ (cf. Sect. 1.4). Suppose that $\mathbf{F}$ is a one-to-one mapping of $\hat{K}$ onto a curved triangle K in the $\mathbf{x}$-plane with inverse $\mathbf{F}^{-1}$; i.e., $K = \mathbf{F}(\hat{K})$ (refer to Fig. 1.25). We then define

$$P(K) = \{v : v(\mathbf{x}) = \hat{v}(\mathbf{F}^{-1}(\mathbf{x})),\ \mathbf{x} \in K,\ \hat{v} \in P(\hat{K})\}\ ,$$
$$\Sigma_K \text{ consists of function values at } \mathbf{m}_i = \mathbf{F}(\hat{\mathbf{m}}_i),\ i = 1, 2, \ldots, l\ .$$

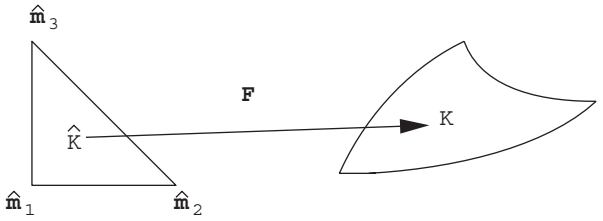

Fig. 1.25. The mapping $\mathbf{F}$

If $\mathbf{F} = (F_1, F_2)$ is of the same type as the functions in $P(K)$, i.e., $F_1,\ F_2 \in P(K)$, then we say that the element $(K, P(K), \Sigma_K)$ is an *isoparametric element*. In general, $\mathbf{F}^{-1}$ is not a polynomial, and thus the functions $v \in P(K)$ for a curved element are not polynomials either.

Let $K_h = \{K\}$ be a triangulation of Ω into "triangles" where some of them may have one or more curved edges, and let Ω_h be the union of these triangles in K_h. Note that Ω_h is an approximation of Ω with a piecewise smooth boundary. Now, the finite element space V_h is

$$V_h = \left\{v \in H^1(\Omega_h) : v|_K \in P(K),\ K \in K_h\right\}\ .$$

With this space, the finite element method can be defined as in (1.21) for the Poisson equation (1.16), for example. Moreover, error estimates analogous to (1.53) and (1.54) hold.

We now consider the computation of a stiffness matrix. Let $\{\hat{\varphi}_i\}_{i=1}^{l}$ be a basis of $P(\hat{K})$. We define

$$\varphi_i(\mathbf{x}) = \hat{\varphi}_i\big(\mathbf{F}^{-1}(\mathbf{x})\big), \quad \mathbf{x} \in K,\ i = 1, 2, \ldots, l\ .$$

For (1.16), we need to compute (cf. Sect. 1.1.2)

$$a^K(\varphi_i, \varphi_j) = \int_K \nabla\varphi_i \cdot \nabla\varphi_j\ d\mathbf{x}, \quad i,\ j = 1, 2, \ldots, l\ . \tag{1.63}$$

It follows from the chain rule that

$$\frac{\partial \varphi_i}{\partial x_k} = \frac{\partial}{\partial x_k}\left(\hat{\varphi}_i\big(\mathbf{F}^{-1}(\mathbf{x})\big)\right) = \frac{\partial \hat{\varphi}_i}{\partial \hat{x}_1}\frac{\partial \hat{x}_1}{\partial x_k} + \frac{\partial \hat{\varphi}_i}{\partial \hat{x}_2}\frac{\partial \hat{x}_2}{\partial x_k}\,,$$

for $k = 1, 2$. Consequently, we see that

$$\nabla \varphi_i = \mathbf{G}^{-T} \nabla \hat{\varphi}_i\,,$$

where $\mathbf{G}^{-T}$ is the transpose of the Jacobian of $\mathbf{F}^{-1}$:

$$\mathbf{G}^{-T} = \begin{pmatrix} \dfrac{\partial \hat{x}_1}{\partial x_1} & \dfrac{\partial \hat{x}_2}{\partial x_1} \\ \dfrac{\partial \hat{x}_1}{\partial x_2} & \dfrac{\partial \hat{x}_2}{\partial x_2} \end{pmatrix}.$$

When we apply the change of variable $\mathbf{F} : \hat{K} \to K$ to (1.63), we have

$$a^K(\varphi_i, \varphi_j) = \int_{\hat{K}} \left(\mathbf{G}^{-T}\nabla \hat{\varphi}_i\right) \cdot \left(\mathbf{G}^{-T}\nabla \hat{\varphi}_j\right) |\det\, \mathbf{G}|\; d\hat{\mathbf{x}}\,, \tag{1.64}$$

for $i, j = 1, 2, \ldots, l$, where $|\det\, \mathbf{G}|$ is the absolute value of the determinant of the Jacobian $\mathbf{G}$:

$$\mathbf{G} = \begin{pmatrix} \dfrac{\partial x_1}{\partial \hat{x}_1} & \dfrac{\partial x_1}{\partial \hat{x}_2} \\ \dfrac{\partial x_2}{\partial \hat{x}_1} & \dfrac{\partial x_2}{\partial \hat{x}_2} \end{pmatrix}.$$

Applying an algebraic computation, we see that

$$\mathbf{G}^{-T} = \left(\mathbf{G}^{-1}\right)^T = \frac{1}{\det\, \mathbf{G}}\mathbf{G}'\,,$$

where

$$\mathbf{G}' = \begin{pmatrix} \dfrac{\partial x_2}{\partial \hat{x}_2} & -\dfrac{\partial x_2}{\partial \hat{x}_1} \\ -\dfrac{\partial x_1}{\partial \hat{x}_2} & \dfrac{\partial x_1}{\partial \hat{x}_1} \end{pmatrix}.$$

Hence (1.64) becomes

$$a^K(\varphi_i, \varphi_j) = \int_{\hat{K}} \left(\mathbf{G}'\nabla \hat{\varphi}_i\right) \cdot \left(\mathbf{G}'\nabla \hat{\varphi}_j\right) \frac{1}{|\det\, \mathbf{G}|}\; d\hat{\mathbf{x}}\,, \tag{1.65}$$

for $i, j = 1, 2, \ldots, l$. Therefore, the matrix entry a_{ij} on K can be calculated by either (1.64) or (1.65). In general, it is difficult to evaluate these two integrals analytically. However, they can be relatively easily evaluated using a numerical integration formula (or a *quadrature rule*); refer to the next section for more details.

We now describe an example of constructing the mapping $\mathbf{F} : \hat{K} \to K$. Let the reference triangle $\hat{K}$ have vertices $\hat{\mathbf{m}}_i$, $i = 1, 2, 3$, and midpoints $\hat{\mathbf{m}}_i$ of the edges, $i = 4, 5, 6$. Furthermore, let $P(\hat{K}) = P_2(\hat{K})$ and let $\Sigma_{\hat{K}}$ be composed of the function values at $\hat{\mathbf{m}}_i$, $i = 1, 2, \ldots, 6$. Define the basis functions $\hat{\varphi}_i \in P_2(\hat{K})$ by

$$\hat{\varphi}_i(\hat{\mathbf{m}}_j) = \delta_{ij}, \quad i, j = 1, 2, \ldots, 6 \, .$$

Also, let the points $\mathbf{m}_i$, $i = 1, 2, \ldots, 6$ in the $\mathbf{x}$-plane satisfy that $\mathbf{m}_4$ and $\mathbf{m}_6$ are the midpoints of the line segments $\mathbf{m}_1\mathbf{m}_2$ and $\mathbf{m}_1\mathbf{m}_3$, respectively, and $\mathbf{m}_5$ is slightly displaced from the line segment $\mathbf{m}_2\mathbf{m}_3$; see Fig. 1.26. We now define $\mathbf{F}$ by

$$\mathbf{F}(\hat{\mathbf{x}}) = \sum_{i=1}^{6} \mathbf{m}_i \hat{\varphi}_i(\hat{\mathbf{x}}), \quad \hat{\mathbf{x}} \in \hat{K} \, .$$

Clearly, $\mathbf{m}_i = \mathbf{F}(\hat{\mathbf{m}}_i)$, $i = 1, 2, \ldots, 6$. Moreover, it can be shown that $\mathbf{F}$ is one-to-one for sufficiently small h_K (Johnson, 1994), i.e., for sufficiently fine triangulations near Γ.

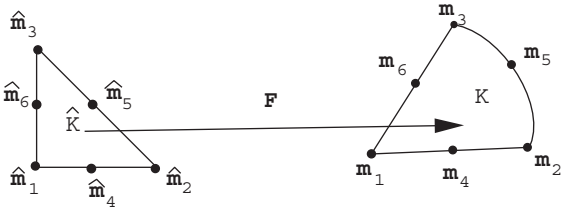

Fig. 1.26. An example of the mapping $\mathbf{F}$

1.6 Quadrature Rules

As mentioned in the previous section, some integrals such as (1.64) and (1.65) can be evaluated only approximately. We can use a *quadrature rule* of the type

$$\int_K g(\mathbf{x}) \, d\mathbf{x} \approx \sum_{i=1}^{m} w_i g(\mathbf{x}_i) \, , \tag{1.66}$$

where $w_i > 0$ and $\mathbf{x}_i$ are certain weights and points in the element K, respectively. If the quadrature rule (1.66) is exact for polynomials of degree r, i.e.,

$$\int_K g(\mathbf{x}) \, d\mathbf{x} = \sum_{i=1}^{m} w_i g(\mathbf{x}_i), \quad g \in P_r(K) \, , \tag{1.67}$$

then the error in using (1.66) can be bounded by (Ciarlet-Raviart, 1972)

$$\left| \int_K g(\mathbf{x})\ d\mathbf{x} - \sum_{i=1}^{m} w_i g(\mathbf{x}_i) \right| \leq C h_K^{r+1} \sum_{|\alpha|=r+1} \int_K |D^\alpha g(\mathbf{x})|\ d\mathbf{x}\ ,$$

where $r > 0$; refer to Sect. 1.2 for the definition of $D^\alpha g$. Several examples are presented below, where r indicates the maximum degree of polynomials for which (1.67) holds.

Example 1.14. Let K be a triangle with vertices $\mathbf{m}_i$, midpoints $\mathbf{m}_{ij}$, $i, j = 1, 2, 3$, $i < j$, and the center of gravity $\mathbf{m}_0$. Also, let $|K|$ indicate the area of K. Then

$$\begin{aligned}
&\int_K g(\mathbf{x})\ d\mathbf{x} \approx |K| g(\mathbf{m}_0) \qquad \text{where } r = 1\ , \\
&\int_K g(\mathbf{x})\ d\mathbf{x} \approx \frac{|K|}{3}(g(\mathbf{m}_{12}) + g(\mathbf{m}_{23}) + g(\mathbf{m}_{13})) \quad \text{where } r = 2\ , \\
&\int_K g(\mathbf{x})\ d\mathbf{x} \approx |K| \left\{ \sum_{i=1}^{3} \frac{g(\mathbf{m}_i)}{20} + \frac{9g(\mathbf{m}_0)}{20} + \frac{2}{15}(g(\mathbf{m}_{12}) + g(\mathbf{m}_{23}) \right. \\
&\qquad\qquad \left. + g(\mathbf{m}_{13})) \right\} \quad \text{where } r = 3\ .
\end{aligned}$$

Example 1.15. Let K be a rectangle centered at the origin and with edges parallel to the x_1- and x_2-coordinate axes of lengths $2h_1$ and $2h_2$, respectively. Then

$$\begin{aligned}
&\int_K g(\mathbf{x})\ d\mathbf{x} \approx |K| g(0) \qquad \text{where } r = 1\ , \\
&\int_K g(\mathbf{x})\ d\mathbf{x} \approx \frac{|K|}{4} \left\{ g\left(\frac{h_1}{\sqrt{3}}, \frac{h_2}{\sqrt{3}}\right) + g\left(\frac{h_1}{\sqrt{3}}, -\frac{h_2}{\sqrt{3}}\right) \right. \\
&\qquad\qquad \left. + g\left(-\frac{h_1}{\sqrt{3}}, \frac{h_2}{\sqrt{3}}\right) + g\left(-\frac{h_1}{\sqrt{3}}, -\frac{h_2}{\sqrt{3}}\right) \right\} \\
&\qquad\qquad \text{where } r = 3\ .
\end{aligned}$$

1.7 Finite Elements for Transient Problems

In this section, we briefly study the finite element method for a *transient* (parabolic) problem in a bounded domain $\Omega \subset \mathbb{R}^d$, $d \geq 1$:

$$\begin{aligned}
&\phi \frac{\partial p}{\partial t} - \nabla \cdot (\mathbf{a} \nabla p) = f \qquad \text{in } \Omega \times J\ , \\
&p = 0 \qquad \text{on } \Gamma \times J\ , \\
&p(\cdot, 0) = p_0 \qquad \text{in } \Omega\ ,
\end{aligned} \tag{1.68}$$

where $J = (0, T]$ $(T > 0)$ is the time interval of interest and ϕ, f, $\mathbf{a}$, and p_0 are given functions. A typical such problem is heat conduction in an inhomogeneous body Ω with heat capacity ϕ and conductivity tensor $\mathbf{a}$. We will first present a *semi-discrete* approximation scheme where (1.68) is discretized only in space using the finite element method. Then we consider *fully discrete* approximation schemes where the time discretization is based on the *backward Euler method* or the *Crank-Nicholson method*. For more details on the finite element method for transient problems, refer to the book by Thomée (1984).

1.7.1 A One-Dimensional Model Problem

To understand some of the major properties of the solution to problem (1.68), we consider the following one-dimensional version that models heat conduction in a bar:

$$\begin{aligned} &\frac{\partial p}{\partial t} - \frac{\partial^2 p}{\partial x^2} = 0, && 0 < x < \pi,\ t \in J\ , \\ &p(0,t) = p(\pi,t) = 0, && t \in J\ , \\ &p(x,0) = p_0(x), && 0 < x < \pi\ . \end{aligned} \tag{1.69}$$

Application of separation of variables yields

$$p(x,t) = \sum_{j=1}^{\infty} p_0^j e^{-j^2 t} \sin(jx)\ , \tag{1.70}$$

where the *Fourier coefficients* p_0^j of the initial datum p_0 are given by

$$p_0^j = \sqrt{\frac{2}{\pi}} \int_0^{\pi} p_0(x) \sin(jx)\ dx, \qquad j = 1, 2, \ldots\ .$$

Note that $\{\sqrt{\frac{2}{\pi}} \sin(jx)\}_{j=1}^{\infty}$ forms an *orthonormal system* in the sense that

$$\frac{2}{\pi} \int_0^{\pi} \sin(jx) \sin(kx)\ dx = \begin{cases} 1 & \text{if } j = k\ , \\ 0 & \text{if } j \neq k\ . \end{cases} \tag{1.71}$$

It follows from (1.70) that the solution p is a linear combination of sine waves $\sin(jx)$ with amplitudes $p_0^j e^{-j^2 t}$ and frequencies j. Because $e^{-j^2 t}$ is very small for $j^2 t$ moderately large, each component $\sin(jx)$ lives on a time scale of order $\mathcal{O}(j^{-2})$. Consequently, high frequency components are quickly damped, and the solution p becomes smoother as t increases. This property can be also understood from the following stability estimates:

$$\begin{aligned} &\|p(t)\|_{L^2(\Omega)} \le \|p_0\|_{L^2(\Omega)}, \qquad t \in J\ , \\ &\left\| \frac{\partial p}{\partial t}(t) \right\|_{L^2(\Omega)} \le \frac{C}{t} \|p_0\|_{L^2(\Omega)}, \quad t \in J\ . \end{aligned} \tag{1.72}$$

We prove these two estimates formally (a proof that is not concerned with any of the convergence questions). From (1.70) and (1.71) it follows that

$$\begin{aligned} \|p(t)\|^2_{L^2(\Omega)} &= \int_0^\pi p^2(x,t)\ dx = \frac{\pi}{2} \sum_{j=1}^{\infty} \left(p_0^j\right)^2 e^{-2j^2 t} \\ &\le \frac{\pi}{2} \sum_{j=1}^{\infty} \left(p_0^j\right)^2 = \|p_0\|^2_{L^2(\Omega)}\ . \end{aligned}$$

Also, note that

$$\frac{\partial p}{\partial t} = \sum_{j=1}^{\infty} p_0^j \big(-j^2\big) e^{-j^2 t} \sin(jx)\ ,$$

so that

$$\left\| \frac{\partial p}{\partial t}(t) \right\|^2_{L^2(\Omega)} = \frac{\pi}{2} \sum_{j=1}^{\infty} \left(p_0^j\right)^2 \big(-j^2\big)^2 e^{-2j^2 t}\ .$$

Using the fact that $0 \le \gamma^2 e^{-\gamma} \le C$ for any $\gamma \ge 0$, we see that

$$\left\| \frac{\partial p}{\partial t}(t) \right\|^2_{L^2(\Omega)} \le \frac{C}{t^2} \|p_0\|^2_{L^2(\Omega)}\ .$$

It follows from the second estimate in (1.72) that if $\|p_0\|_{L^2(\Omega)} < \infty$, then $\|\partial p/\partial t(t)\|_{L^2(\Omega)} = \mathcal{O}(t^{-1})$ as $t \to 0$. An initial phase (for t small) where certain derivatives of p are large is referred to as an *initial transient*. In general, the solution p of a parabolic problem has an initial transient. It will become smoother as t increases. This observation is very important when the parabolic problem is numerically solved. It is desirable to vary the grid size (in space and time) according to the smoothness of p. For a region where p is nonsmooth, a fine grid is used; for a region where p becomes smoother, the grid size is increased. That is, an *adaptive finite element method* should be employed, which will be discussed in Chap. 6. We mention that transients may also occur at times $t > 0$ if the boundary data or the source term (the right-hand side function in (1.68)) changes abruptly in time.

1.7.2 A Semi-Discrete Scheme in Space

We now return to problem (1.68). For simplicity, we study a special case of this problem where $\phi = 1$ and $\mathbf{a} = \mathbf{I}$ (the identity tensor). Set

$$V = H_0^1(\Omega) = \left\{ v \in H^1(\Omega) : \ v\big|_\Gamma = 0 \right\} \ .$$

As in Sect. 1.1.2, we exploit the notation

$$a(p,v) = \int_\Omega \nabla p \cdot \nabla v \ d\mathbf{x}, \quad (f,v) = \int_\Omega f v \ d\mathbf{x} \ .$$

Then (1.68) is written in the variational form: Find $p : J \to V$ such that

$$\begin{aligned} &\left(\frac{\partial p}{\partial t}, v\right) + a(p,v) = (f,v) \qquad \forall v \in V,\ t \in J \ , \\ &p(\mathbf{x},0) = p_0(\mathbf{x}) \qquad\qquad\quad \forall \mathbf{x} \in \Omega \ . \end{aligned} \tag{1.73}$$

Let V_h be a finite element subspace of V. Replacing V in (1.73) by V_h, we have the finite element method: Find $p_h : J \to V_h$ such that

$$\begin{aligned} &\left(\frac{\partial p_h}{\partial t}, v\right) + a(p_h,v) = (f,v) \qquad \forall v \in V_h,\ t \in J \ , \\ &(p_h(\cdot,0), v) = (p_0, v) \qquad\qquad\quad \forall v \in V_h \ . \end{aligned} \tag{1.74}$$

This system is discretized in space, but continuous in time. For this reason, it is called a *semi-discrete scheme*. Let the basis functions in V_h be denoted by φ_i, $i = 1, 2, \ldots, M$, and express p_h as

$$p_h(\mathbf{x},t) = \sum_{i=1}^{M} p_i(t)\varphi_i(\mathbf{x}), \qquad (\mathbf{x},t) \in \Omega \times J \ . \tag{1.75}$$

For $j = 1, 2, \ldots, M$, we take $v = \varphi_j$ in (1.74) and utilize (1.75) to see that, for $t \in J$,

$$\begin{aligned} &\sum_{i=1}^{M} (\varphi_i, \varphi_j) \frac{dp_i}{dt} + \sum_{i=1}^{M} a(\varphi_i, \varphi_j) p_i = (f, \varphi_j), \quad j = 1, 2, \ldots, M \ , \\ &\sum_{i=1}^{M} (\varphi_i, \varphi_j) p_i(0) = (p_0, \varphi_j), \qquad\qquad\qquad\qquad j = 1, 2, \ldots, M \ , \end{aligned}$$

which, in matrix form, is given by

$$\begin{aligned} &\mathbf{B}\frac{d\mathbf{p}(t)}{dt} + \mathbf{A}\mathbf{p}(t) = \mathbf{f}(t), \qquad t \in J \ , \\ &\mathbf{B}\mathbf{p}(0) = \mathbf{p}_0 \ , \end{aligned} \tag{1.76}$$

where the $M \times M$ matrices $\mathbf{A}$ and $\mathbf{B}$ and the vectors $\mathbf{p}$, $\mathbf{f}$, and $\mathbf{p}_0$ are

$$\begin{aligned}
&\mathbf{A} = (a_{ij}), && a_{ij} = a(\varphi_i, \varphi_j) \ , \\
&\mathbf{B} = (b_{ij}), && b_{ij} = (\varphi_i, \varphi_j) \ , \\
&\mathbf{p} = (p_j), && \mathbf{f} = (f_j) \, , \quad f_j = (f, \varphi_j) \ , \\
&\mathbf{p}_0 = ((p_0)_j), && (p_0)_j = (p_0, \varphi_j) \ .
\end{aligned}$$

Both $\mathbf{A}$ and $\mathbf{B}$ are symmetric and positive definite, as was shown in the stationary case. Their *condition numbers* are of the order $\mathcal{O}(h^{-2})$ and $\mathcal{O}(1)$ as $h \to 0$ (cf. Sect. 1.10), respectively, where we recall that for a symmetric matrix, its condition number is defined as the ratio of its largest eigenvalue to its smallest eigenvalue. For this reason, the matrices $\mathbf{A}$ and $\mathbf{B}$ are referred to as the *stiffness* and *mass* matrices, respectively. Thus (1.76) is a *stiff* system of ordinary differential equations (ODEs). The usual way to solve an ODE system is to discretize the time derivative as well. One approach is to exploit the numerical methods developed already for ODEs. Because of the large number of simultaneous equations, however, simple numerical methods for transient partial differential problems have been developed independent of the methods for ODEs. This will be discussed in the next subsection.

We show a stability result for the semi-discrete system (1.74) with $f = 0$. We choose $v = p_h(t)$ in the first equation of (1.74) to obtain

$$\left(\frac{\partial p_h}{\partial t}, p_h \right) + a(p_h, p_h) = 0 \ ,$$

which gives

$$\frac{1}{2} \frac{d}{dt} \|p_h(t)\|^2_{L^2(\Omega)} + a(p_h, p_h) = 0 \ .$$

Also, take $v = p_h(0)$ in the second equation of (1.74) and use Cauchy's inequality (1.10) to see that

$$\|p_h(0)\|_{L^2(\Omega)} \le \|p_0\|_{L^2(\Omega)} \ .$$

Then it follows that

$$\|p_h(t)\|^2_{L^2(\Omega)} + 2 \int_0^t a(p_h(\ell), p_h(\ell)) \, d\ell = \|p_h(0)\|^2_{L^2(\Omega)} \le \|p_0\|^2_{L^2(\Omega)} \ .$$

Consequently, we obtain

$$\|p_h(t)\|_{L^2(\Omega)} \le \|p_0\|_{L^2(\Omega)}, \qquad t \in J \ . \tag{1.77}$$

This inequality is similar to the first inequality in (1.72). In fact, the latter inequality can be shown in the same manner as for (1.77). The derivation of an error estimate for (1.74) is much more elaborate than that for a stationary problem. We just state an estimate for the case where V_h is the space of piecewise linear functions on a *quasi-uniform* triangulation of Ω in the sense that there is a positive constant β_2, independent of h, such that

$$h_K \geq \beta_2 h \ , \forall K \in K_h \ , \tag{1.78}$$

where we recall that $h_K = \text{diam}(K)$, $K \in K_h$, and $h = \max\{h_K : K \in K_h\}$. Condition (1.78) says that all elements $K \in K_h$ are roughly of the same size. The error estimate reads as follows (Thomée, 1984; Johnson, 1994):

$$\max_{t\in J} \|(p - p_h)(t)\|_{L^2(\Omega)} \leq C \left(1 + \left|\ln \frac{T}{h^2}\right|\right) \max_{t\in J} h^2 \|p(t)\|_{H^2(\Omega)} \ . \tag{1.79}$$

Due to the presence of the term $\ln h^{-2}$, this estimate is only *almost optimal*.

1.7.3 Fully Discrete Schemes

We consider three fully discrete schemes: the backward and forward Euler methods and the Crank-Nicholson method.

1.7.3.1 The Backward Euler Method

Let $0 = t^0 < t^1 < \ldots < t^N = T$ be a partition of J into subintervals $J^n = (t^{n-1}, t^n)$, with length $\Delta t^n = t^n - t^{n-1}$. For a generic function v of time, set $v^n = v(t^n)$. The *backward Euler method* for the semi-discrete version (1.74) is: Find $p_h^n \in V_h$, $n = 1, 2, \ldots, N$, such that

$$\begin{aligned} &\left(\frac{p_h^n - p_h^{n-1}}{\Delta t^n}, v\right) + a\,(p_h^n, v) = (f^n, v) \qquad \forall v \in V_h \ , \\ &\left(p_h^0, v\right) = (p_0, v) \qquad\qquad\qquad\qquad\quad \forall v \in V_h \ . \end{aligned} \tag{1.80}$$

Note that (1.80) comes from replacing the time derivative in (1.74) by the difference quotient $(p_h^n - p_h^{n-1})/\Delta t^n$. This replacement results in a discretization error of order $\mathcal{O}\,(\Delta t^n)$. As in (1.76), (1.80) can be expressed in matrix form

$$\begin{aligned} &(\mathbf{B} + \mathbf{A}\Delta t^n)\,\mathbf{p}^n = \mathbf{B}\mathbf{p}^{n-1} + \mathbf{f}^n \Delta t^n \ , \\ &\mathbf{B}\mathbf{p}(0) = \mathbf{p}_0 \ , \end{aligned} \tag{1.81}$$

where

$$p_h^n = \sum_{i=1}^{M} p_i^n \varphi_i, \qquad n = 0, 1, \ldots, N \ ,$$

and

$$\mathbf{p}^n = (p_1^n, p_2^n, \ldots, p_M^n)^T \ .$$

Clearly, (1.81) is an *implicit* scheme; that is, we need to solve a system of linear equations at each time step.

Let us state a basic *stability* estimate for (1.80) in the case $f = 0$. Choosing $v = p_h^n$ in (1.80), we see that

$$\|p_h^n\|^2 - \left(p_h^{n-1}, p_h^n\right) + a\left(p_h^n, p_h^n\right)\Delta t^n = 0 \ .$$

It follows from Cauchy's inequality (1.10) that

$$\left(p_h^{n-1}, p_h^n\right) \le \|p_h^{n-1}\| \ \|p_h^n\| \le \frac{1}{2}\|p_h^{n-1}\|^2 + \frac{1}{2}\|p_h^n\|^2 \ .$$

Consequently, we get

$$\frac{1}{2}\|p_h^n\|^2 - \frac{1}{2}\|p_h^{n-1}\|^2 + a\left(p_h^n, p_h^n\right)\Delta t^n \le 0 \ .$$

We sum over n and use the second equation in (1.80) to give

$$\|p_h^j\|^2 + 2\sum_{n=1}^{j} a\left(p_h^n, p_h^n\right)\Delta t^n \le \|p_h^0\|^2 \le \|p^0\|^2 \ .$$

Because $a\left(p_h^n, p_h^n\right) \ge 0$, we obtain the stability result

$$\|p_h^j\| \le \|p^0\|, \qquad j = 0, 1, \ldots, N \ . \tag{1.82}$$

Note that (1.82) holds regardless of the size of the time steps Δt^j. In other words, the backward Euler method (1.80) is *unconditionally stable*. This is a very desirable feature of a time discretization scheme for a parabolic problem.

We remark that an estimate for the error $p - p_h$ can be derived. The error stems from a combination of the space and time discretizations. When V_h is the finite element space of piecewise linear functions, for example, the error $p^n - p_h^n$ $(0 \le n \le N)$ in the L^2-norm is of order $\mathcal{O}\left(\Delta t + h^2\right)$ (Thomée, 1984) under appropriate smoothness assumptions on p, where $\Delta t = \max\{\Delta t^j, 1 \le j \le N\}$.

1.7.3.2 The Crank-Nicholson Method

The *Crank-Nicholson method* for (1.74) is defined: Find $p_h^n \in V_h$, $n = 1, 2, \ldots, N$, such that

$$\begin{aligned} &\left(\frac{p_h^n - p_h^{n-1}}{\Delta t^n}, v\right) + a\left(\frac{p_h^n + p_h^{n-1}}{2}, v\right) = \left(\frac{f^n + f^{n-1}}{2}, v\right) \\ &\qquad\qquad \forall v \in V_h \ , \\ &\left(p_h^0, v\right) = (p_0, v) \qquad \forall v \in V_h \ . \end{aligned} \tag{1.83}$$

In the present case, the difference quotient $(p_h^n - p_h^{n-1})/\Delta t^n$ now replaces the average $\left(\partial p(t^n)/\partial t + \partial p(t^{n-1})/\partial t\right)/2$. The resulting discretization error in time is $\mathcal{O}\left((\Delta t^n)^2\right)$. Similarly to (1.81), the linear system from (1.83) is

$$\begin{aligned} &\left(\mathbf{B} + \frac{\Delta t^n}{2}\mathbf{A}\right)\mathbf{p}^n = \left(\mathbf{B} - \frac{\Delta t^n}{2}\mathbf{A}\right)\mathbf{p}^{n-1} + \frac{\mathbf{f}^n + \mathbf{f}^{n-1}}{2}\Delta t^n \ , \\ &\mathbf{B}\mathbf{p}(0) = \mathbf{p}_0 \ , \end{aligned} \tag{1.84}$$

for $n = 1, 2, \ldots, N$. Again, this is an implicit method. When $f = 0$, by taking $v = (p_h^n + p_h^{n-1})/2$ in (1.83) one can show that the stability result (1.82) unconditionally holds for the Crank-Nicholson method, too; see Exercise 1.30. For the piecewise linear finite element space V_h, for each n the error $p^n - p_h^n$ in the L^2-norm is $\mathcal{O}\left((\Delta t)^2 + h^2\right)$ this time. Note that the Crank-Nicholson method is more accurate in time than the backward Euler method and is slightly more expensive from the computational point of view.

1.7.3.3 The Forward Euler Method

We conclude with the *forward Euler method*. This method takes the form: Find $p_h^n \in V_h$, $n = 1, 2, \ldots, N$, such that

$$\begin{aligned} &\left(\frac{p_h^n - p_h^{n-1}}{\Delta t^n}, v\right) + a\left(p_h^{n-1}, v\right) = \left(f^{n-1}, v\right) \qquad \forall v \in V_h \ , \\ &\left(p_h^0, v\right) = (p_0, v) \qquad\qquad\qquad\qquad\qquad\quad \forall v \in V_h \ , \end{aligned} \tag{1.85}$$

and the corresponding matrix form is

$$\begin{aligned} &\mathbf{B}\mathbf{p}^n = \left(\mathbf{B} - \mathbf{A}\Delta t^n\right)\mathbf{p}^{n-1} + \mathbf{f}^{n-1}\Delta t^n \ , \\ &\mathbf{B}\mathbf{p}(0) = \mathbf{p}_0 \ . \end{aligned} \tag{1.86}$$

Introducing the Cholesky decomposition $\mathbf{B} = \mathbf{D}\mathbf{D}^T$ (cf. Sect. 1.10) and using the new variable $\mathbf{q} = \mathbf{D}^T\mathbf{p}$, where $\mathbf{D}^T$ is the transpose of $\mathbf{D}$, problem (1.86) is of the simpler form

$$\begin{aligned} &\mathbf{q}^n = \left(\mathbf{I} - \tilde{\mathbf{A}}\Delta t^n\right)\mathbf{q}^{n-1} + \mathbf{D}^{-1}\mathbf{f}^{n-1}\Delta t^n \ , \\ &\mathbf{q}(0) = \mathbf{D}^{-1}\mathbf{p}_0 \ , \end{aligned} \tag{1.87}$$

where $\tilde{\mathbf{A}} = \mathbf{D}^{-1}\mathbf{A}\mathbf{D}^{-T}$. Clearly, (1.87) is an *explicit scheme* in $\mathbf{q}$. A stability result similar to (1.82) can be proven only under the *stability condition*

$$\Delta t^n \leq Ch^2, \qquad n = 1, 2, \ldots, N \ , \tag{1.88}$$

where C is a constant independent of Δt and h. This can be seen as follows: With $f = 0$, the first equation of (1.87) becomes

$$\mathbf{q}^n = (\mathbf{I} - \tilde{\mathbf{A}}\Delta t^n)\mathbf{q}^{n-1} \ . \tag{1.89}$$

Define the matrix norm

$$\|\tilde{\mathbf{A}}\| = \max_{\boldsymbol{\eta} \in \mathbb{R}^M, \boldsymbol{\eta} \neq \mathbf{0}} \frac{\|\tilde{\mathbf{A}}\boldsymbol{\eta}\|}{\|\boldsymbol{\eta}\|} \ ,$$

where $\|\boldsymbol{\eta}\|$ is the Euclidean norm of $\boldsymbol{\eta}$: $\|\boldsymbol{\eta}\|^2 = \eta_1^2 + \eta_2^2 + \ldots + \eta_M^2$, $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_M)$. Assume that the symmetric matrix $\tilde{\mathbf{A}}$ has eigenvalues $\mu_i > 0$, $i = 1, 2, \ldots, M$. Then we see that (Axelsson, 1994)

$$\|\tilde{\mathbf{A}}\| = \max_{i=1,2,\ldots,M} |\mu_i| \ .$$

Thus it follows that

$$\|\mathbf{I} - \tilde{\mathbf{A}}\Delta t^n\| = \max_{i=1,2,\ldots,M} |1 - \mu_i \Delta t^n| \ .$$

Let the maximum occur as $i = M$, for example. Then

$$\|\mathbf{I} - \tilde{\mathbf{A}}\Delta t^n\| \leq 1 \ ,$$

only if $\mu_M \Delta t^n \leq 2$. Since $\mu_M = \mathcal{O}(h^{-2})$ (cf. Sect. 1.10), $\Delta t^n \leq 2/\mu_M = \mathcal{O}(h^2)$, which is (1.88).

The stability condition (1.88) requires that the time step be sufficiently small. In other words, the forward Euler method (1.85) is *conditionally stable*. This condition is very restrictive, particularly for long time integration. In contrast, the backward Euler and Crank-Nicholson methods are unconditionally stable, but require more work per time step. These two methods are more efficient for parabolic problems since the extra cost involved at each step for an implicit method is more than compensated for by the fact that bigger time steps can be utilized.

1.8 Finite Elements for Nonlinear Problems

In this section, we briefly consider an application of the finite element method to the *nonlinear transient problem*

$$\begin{aligned} &c(p)\frac{\partial p}{\partial t} - \nabla \cdot \big(a(p)\nabla p\big) = f(p) && \text{in } \Omega \times J \ , \\ &p = 0 && \text{on } \Gamma \times J \ , \\ &p(\cdot, 0) = p_0 && \text{in } \Omega \ , \end{aligned} \tag{1.90}$$

where $c(p) = c(\mathbf{x}, t, p)$, $a(p) = a(\mathbf{x}, t, p)$, and $f(p) = f(\mathbf{x}, t, p)$ depend on the unknown p. In (1.90) and below, for notational convenience, we drop the dependence of these coefficients on $\mathbf{x}$ and t. We assume that (1.90) admits a unique solution. Furthermore, we assume that the coefficients $c(p)$, $a(p)$, and $f(p)$ are *globally Lipschitz continuous* in p; i.e., for some constants C_ξ, they satisfy

$$|\xi(p_1) - \xi(p_2)| \leq C_\xi |p_1 - p_2|, \qquad p_1, \, p_2 \in \mathbb{R}, \, \xi = c, \, a, \, f \ . \tag{1.91}$$

With $V = H_0^1(\Omega)$, problem (1.90) can be written in the variational form: Find $p : J \to V$ such that

$$\begin{aligned} &\left(c(p)\frac{\partial p}{\partial t}, v\right) + \big(a(p)\nabla p, \nabla v\big) = \big(f(p), v\big) \qquad \forall v \in V,\, t \in J\ , \\ &p(\mathbf{x}, 0) = p_0(\mathbf{x}) \qquad\qquad\qquad\qquad\quad \forall \mathbf{x} \in \Omega\ . \end{aligned} \tag{1.92}$$

Let V_h be a finite element subspace of V. The finite element version of (1.92) is: Find $p_h : J \to V_h$ such that

$$\begin{aligned} &\left(c(p_h)\frac{\partial p_h}{\partial t}, v\right) + \big(a(p_h)\nabla p_h, \nabla v\big) = \big(f(p_h), v\big) \qquad \forall v \in V_h\ , \\ &(p_h(\cdot, 0), v) = (p_0, v) \qquad\qquad\qquad\quad \forall v \in V_h\ . \end{aligned} \tag{1.93}$$

As for (1.76), after the introduction of basis functions in V_h, (1.93) can be stated in matrix form

$$\begin{aligned} &\mathbf{C}(\mathbf{p})\frac{d\mathbf{p}}{dt} + \mathbf{A}(\mathbf{p})\mathbf{p} = \mathbf{f}(\mathbf{p}), \qquad t \in J\ , \\ &\mathbf{B}\mathbf{p}(0) = \mathbf{p}_0\ . \end{aligned} \tag{1.94}$$

Under the assumption that the coefficient $c(p)$ is bounded below by a positive constant, this nonlinear system of ODEs locally has a unique solution. In fact, because of assumption (1.91) on c, a, and f, the solution $\mathbf{p}(t)$ exists for all t. Several approaches for solving (1.94) are discussed in this section.

1.8.1 Linearization Approaches

The nonlinear system (1.94) can be linearized by allowing the nonlinearities to lag one time step behind. Thus the backward Euler method for (1.90) takes the form: Find $p_h^n \in V_h$, $n = 1, 2, \ldots, N$, such that

$$\begin{aligned} &\left(c\left(p_h^{n-1}\right)\frac{p_h^n - p_h^{n-1}}{\Delta t^n}, v\right) + \left(a\left(p_h^{n-1}\right)\nabla p_h^n, \nabla v\right) \\ &\qquad\qquad = \left(f\left(p_h^{n-1}\right), v\right) \qquad \forall v \in V_h\ , \\ &\left(p_h^0, v\right) = (p_0, v) \qquad \forall v \in V_h\ . \end{aligned} \tag{1.95}$$

In matrix form it is given by

$$\begin{aligned} &\mathbf{C}\left(\mathbf{p}^{n-1}\right)\frac{\mathbf{p}^n - \mathbf{p}^{n-1}}{\Delta t^n} + \mathbf{A}\big(\mathbf{p}^{n-1}\big)\mathbf{p}^n = \mathbf{f}\big(\mathbf{p}^{n-1}\big)\ , \\ &\mathbf{B}\mathbf{p}(0) = \mathbf{p}_0\ . \end{aligned} \tag{1.96}$$

Note that (1.96) is a system of linear equations for $\mathbf{p}^n$, which can be solved using *iterative algorithms* as discussed in Sect. 1.10, for example. When V_h is

the finite element space of piecewise linear functions, the error $p^n - p_h^n$ $(0 \leq n \leq N)$ in the L^2-norm is of order $\mathcal{O}\left(\Delta t + h^2\right)$ as for problem (1.68) under appropriate smoothness assumptions on p and for Δt small enough (Thomée, 1984; Chen-Douglas, 1991). We may use the Crank-Nicholson discretization method in (1.95). However, the linearization decreases the order of the time discretization error to $\mathcal{O}(\Delta t)$, giving $\mathcal{O}\left(\Delta t + h^2\right)$ overall. This is true for any higher-order time discretization method with the present linearization technique. This drawback can be overcome by using *extrapolation techniques* in the linearization of the coefficients c, a, and f (cf. Sect. 5.7). Combined with an appropriate extrapolation, the Crank-Nicholson method can be shown to produce an error of order $\mathcal{O}\big((\Delta t)^2\big)$ in time (Douglas, 1961; Thomée, 1984). In general, higher-order extrapolations increase data storage.

1.8.2 Implicit Time Approximations

We now consider a fully *implicit time approximation* scheme for (1.90): Find $p_h^n \in V_h$, $n = 1, 2, \ldots, N$, such that

$$\begin{aligned} &\left(c\left(p_h^n\right) \frac{p_h^n - p_h^{n-1}}{\Delta t^n}, v\right) + \left(a\left(p_h^n\right) \nabla p_h^n, \nabla v\right) \\ &\qquad\qquad = \left(f\left(p_h^n\right), v\right) \qquad \forall v \in V_h \,, \\ &\left(p_h^0, v\right) = (p_0, v) \qquad \forall v \in V_h \,. \end{aligned} \tag{1.97}$$

Its matrix form is

$$\begin{aligned} &\mathbf{C}\left(\mathbf{p}^n\right) \frac{\mathbf{p}^n - \mathbf{p}^{n-1}}{\Delta t^n} + \mathbf{A}\big(\mathbf{p}^n\big)\mathbf{p}^n = \mathbf{f}\big(\mathbf{p}^n\big) \,, \\ &\mathbf{B}\mathbf{p}(0) = \mathbf{p}_0 \,. \end{aligned} \tag{1.98}$$

Now, (1.98) is a system of nonlinear equations in $\mathbf{p}^n$, which must be solved at each time step via an iteration method. Let us consider *Newton's method* (or Newton-Raphson's method). Note that the first equation of (1.98) can be rewritten as

$$\left(\mathbf{A}\big(\mathbf{p}^n\big) + \frac{1}{\Delta t^n}\mathbf{C}\left(\mathbf{p}^n\right)\right)\mathbf{p}^n - \frac{1}{\Delta t^n}\mathbf{C}\left(\mathbf{p}^n\right)\mathbf{p}^{n-1} - \mathbf{f}\big(\mathbf{p}^n\big) = \mathbf{0} \,.$$

We express this equation as

$$\mathbf{F}\big(\mathbf{p}^n\big) = \mathbf{0} \,. \tag{1.99}$$

Newton's method for (1.99) can be defined in the form

$$\begin{aligned}&\text{Set } \mathbf{v}^0 = \mathbf{p}^{n-1};\\ &\text{Iterate } \mathbf{v}^k = \mathbf{v}^{k-1} + \mathbf{d}^k, \qquad k = 1, 2, \dots ,\end{aligned}$$

where $\mathbf{d}^k$ solves the system

$$\mathbf{G}\big(\mathbf{v}^{k-1}\big)\mathbf{d}^k = -\mathbf{F}\big(\mathbf{v}^{k-1}\big) ,$$

with $\mathbf{G}$ being the Jacobian matrix of the vector function $\mathbf{F}$:

$$\mathbf{G} = \left(\frac{\partial F_i}{\partial p_j}\right)_{i,j=1,2,\dots,M} .$$

If the matrix $\mathbf{G}\big(\mathbf{p}^n\big)$ is nonsingular and the second partial derivatives of $\mathbf{F}$ are bounded, Newton's method converges quadratically in a neighborhood of $\mathbf{p}^n$; i.e., there are constants $\epsilon > 0$ and C such that if $\big|\mathbf{v}^{k-1} - \mathbf{p}^n\big| \le \epsilon$, then

$$\big|\mathbf{v}^k - \mathbf{p}^n\big| \le C\big|\mathbf{v}^{k-1} - \mathbf{p}^n\big|^2 .$$

The main difficulty with Newton's method is to get a sufficiently good initial guess. Once it is obtained, Newton's method converges with very few iterations. This method is a very powerful iteration method for strongly nonlinear problems. There are many variants of Newton's method available in the literature (Ostrowski, 1973; Rheinboldt, 1998). We remark that the Crank-Nicholson discretization procedure can be used in (1.97) as well. In the present implicit case, this procedure generates second order accuracy in time. Numerical experience has indicated that the Crank-Nicholson procedure may not be a good choice for nonlinear parabolic equations because it can be unstable for such equations.

1.8.3 Explicit Time Approximations

We end with the application of a forward, explicit time approximation method to (1.90): Find $p_h^n \in V_h$, $n = 1, 2, \dots, N$, such that

$$\begin{aligned}&\left(c\left(p_h^n\right)\frac{p_h^n - p_h^{n-1}}{\Delta t^n}, v\right) + \left(a\left(p_h^{n-1}\right)\nabla p_h^{n-1}, \nabla v\right)\\ &\qquad\qquad = \left(f\left(p_h^{n-1}\right), v\right) \qquad \forall v \in V_h ,\\ &\left(p_h^0, v\right) = (p_0, v) \qquad \forall v \in V_h .\end{aligned} \tag{1.100}$$

In matrix form it is written as follows:

$$\begin{aligned}&\mathbf{C}\left(\mathbf{p}^n\right)\frac{\mathbf{p}^n - \mathbf{p}^{n-1}}{\Delta t^n} + \mathbf{A}\big(\mathbf{p}^{n-1}\big)\mathbf{p}^{n-1} = \mathbf{f}\big(\mathbf{p}^{n-1}\big) ,\\ &\mathbf{B}\mathbf{p}(0) = \mathbf{p}_0 .\end{aligned} \tag{1.101}$$

Note that the only nonlinearity is in matrix **C**. This system can be solved via any standard method (Ostrowski, 1973; Rheinboldt, 1998).

For the explicit method (1.100) to be stable in the sense discussed in Sect. 1.7, a *stability condition* of the following type must be satisfied:

$$\Delta t^n \leq Ch^2, \qquad n = 1, 2, \ldots, N, \tag{1.102}$$

where C now depends on c and a (cf. (1.88)). Unfortunately, this condition on the time steps is very restrictive for long time integration, as noted earlier.

In summary, we have developed linearization, implicit, and explicit time approximation approaches for numerically solving (1.90). In terms of computational effort, the explicit approach is the simplest at each time step; however, it requires an impracticable stability restriction. The linearization approach is more practical, but it reduces the order of accuracy in time for high-order time discretization methods (unless extrapolations are exploited). An efficient method is the fully implicit approach; the extra cost involved at each time step for this implicit method is more than compensated for by the fact that bigger time steps may be taken, particularly when Newton's method with a good initial guess is employed. Modified implicit methods such as *semi-implicit methods* (Aziz-Settari, 1979) can be applied; for a given physical problem, the linearization approach should be applied to weak nonlinearity, while the implicit one should be used for strong nonlinearity (Chen et al., 2000).

1.9 Approximation Theory

1.9.1 Interpolation Errors

It follows from (1.49) that the error $\|p - p_h\|_V$ can be bounded by choosing a suitable function $v \in V_h$ and analyzing $\|p - v\|_V$. We often choose $v = \pi_h p \in V_h$ to be a certain *interpolant* of p in V_h.

We define the *interpolation operator* π_h for three examples. Consider the case where $\Omega \subset \mathbb{R}^2$ is a polygon, K_h is a triangulation of Ω into triangles, $V = H^1(\Omega)$, and (cf. Example 1.6)

$$V_h = \{v \in V : v|_K \in P_1(K),\ K \in K_h\}\ .$$

Let $\{\mathbf{m}_i\}_{i=1}^M$ be the set of vertices in K_h. For any $u \in C^0(\bar{\Omega})$, we define its interpolant $\pi_h u \in V_h$ by

$$\pi_h u(\mathbf{m}_i) = u(\mathbf{m}_i), \quad i = 1, 2, \ldots, M\ . \tag{1.103}$$

Thus $\pi_h u$ is the piecewise linear function that has the same values as u at the nodes of K_h.

For V_h defined in Example 1.7, i.e,

$$V_h = \{v \in H^1(\Omega) : v|_K \in P_2(K),\, K \in K_h\} \,,$$

let now $\{\mathbf{m}_i\}_{i=1}^M$ be the collection of vertices and midpoints of edges in K_h; the interpolant $\pi_h u \in V_h$ of $u \in C^0(\bar{\Omega})$ is defined by the same expression (1.103). That is, $\pi_h u$ is the piecewise quadratic function that has the same values as u at the vertices and midpoints of K_h.

Consider Example 1.13, where

$$V_h = \{v \in H^2(\Omega) : v|_K \in P_2(K),\, K \in K_h\} \,.$$

Let $\{\mathbf{m}_i\}$ and $\{\mathbf{m}^i\}$ be the sets of vertices and midpoints of edges in K_h, respectively. Now, for any $u \in C^1(\bar{\Omega})$, we define $\pi_h u \in V_h$ by

$$\begin{aligned} &D^\alpha(\pi_h u)(\mathbf{m}_i) = D^\alpha u(\mathbf{m}_i), \quad |\alpha| \le 2 \,, \\ &\frac{\partial \pi_h u}{\partial \boldsymbol{\nu}}(\mathbf{m}^i) = \frac{\partial u}{\partial \boldsymbol{\nu}}(\mathbf{m}^i) \,, \end{aligned}$$

where $\boldsymbol{\nu}$ is a unit normal to the edge containing $\mathbf{m}^i$. For other finite element spaces introduced in Sect. 1.4, the interpolation operator π_h can be similarly defined using their respective degrees of freedom.

In this section, we will estimate $\|p - \pi_h p\|_V$. To that end, we introduce some concepts. As noted, a function $f : \Omega \to \mathbb{R}^m$ is *Lipschitz continuous* in Ω if there is a constant $C > 0$ such that

$$\|f(\mathbf{x}) - f(\mathbf{y})\| \le C\|\mathbf{x} - \mathbf{y}\| \qquad \forall \mathbf{x}, \mathbf{y} \in \Omega \subset \mathbb{R}^d \,,$$

where d and m are two positive integers. A hypersurface in $\mathbb{R}^d$ is a *graph* if it can be represented by a function g in the form

$$x_i = g(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d),\ (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d) \in D \,,$$

for some i $(1 \le i \le d)$ and domain $D \subset \mathbb{R}^{d-1}$. A domain $\Omega \subset \mathbb{R}^d$ is termed a *Lipschitz domain* if for each $\mathbf{x}$ in the boundary Γ of Ω, there is an open subset $O_{\mathbf{x}} \subset \mathbb{R}^d$ containing $\mathbf{x}$ such that $O_{\mathbf{x}} \cap \Gamma$ can be represented by the graph of a Lipschitz continuous function.

Lemma 1.4 (Bramble-Hilbert Lemma, 1970). *Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain, and let $\mathcal{F} : H^r(\Omega) \to \mathcal{Y}$ be a bounded linear operator, where $r \ge 1$ and $\mathcal{Y}$ is a normed linear space, such that $P_{r-1}(\Omega)$ is a subset of the kernel of $\mathcal{F}$, where the kernel of $\mathcal{F}$ is defined by $\{v \in H^r(\Omega) : \mathcal{F}(v) = 0\}$. Then there is a positive constant C such that*

$$\|\mathcal{F}(v)\|_{\mathcal{Y}} \le C(\Omega)\|\mathcal{F}\|\, \|v\|_{H^r(\Omega)}, \qquad v \in H^r(\Omega) \,,$$

where $\|\mathcal{F}\|$ is the norm of the operator $\mathcal{F}$.

Lemma 1.5 (Transformation formula). *Let K and $\hat{K}$ be two affine-equivalent open subsets of $\mathbb{R}^d$; i.e., there is a bijective affine mapping*

$$\begin{aligned} &\mathbf{F} : \hat{K} \to K \ , \\ &\mathbf{F}(\hat{\mathbf{x}}) = \mathbf{b} + \mathbf{B}_1 \hat{\mathbf{x}} \ , \end{aligned}$$

where $\mathbf{B}_1$ is a nonsingular matrix and $\mathbf{b} \in \mathbb{R}^d$ is some vector. If $v \in H^r(K)$, then the composite function $\hat{v} = v \circ \mathbf{F} \in H^r(\hat{K})$, and there is a constant $C(d,r)$ such that

$$|\hat{v}|_{H^r(\hat{K})} \le C(d,r) \|\mathbf{B}_1\|^r |\det \mathbf{B}_1|^{-1/2} |v|_{H^r(K)} \ , \tag{1.104}$$

where $\|\mathbf{B}_1\|$ is the matrix norm of $\mathbf{B}_1$. Analogously, it holds that

$$|v|_{H^r(K)} \le C(d,r) \|\mathbf{B}_1^{-1}\|^r |\det \mathbf{B}_1|^{1/2} |\hat{v}|_{H^r(\hat{K})} \ . \tag{1.105}$$

Proof. Since $C^r(\bar{K})$ is dense in $H^r(K)$, it is sufficient to work with the functions $v \in C^r(\bar{K})$. Then $\hat{v} \in C^r(\bar{\hat{K}})$. For any multi-index $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ with $|\alpha| = r$, we use a multilinear form of the derivative of order r:

$$D^\alpha \hat{v}(\hat{\mathbf{x}}) = D^r \hat{v}(\hat{\mathbf{x}})(\mathbf{e}_{\alpha_1}, \mathbf{e}_{\alpha_2}, \ldots, \mathbf{e}_{\alpha_r}) \ ,$$

where the vectors $\mathbf{e}_{\alpha_i}$ $(1 \le i \le r)$ are some of the basis vectors of $\mathbb{R}^d$. For any vectors $\mathbf{y}_i$, $i = 1, 2, \ldots, r$, a multilinear form of the rth-order derivative is defined by

$$D^r v(\mathbf{x})(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_r) = \prod_{i=1}^{r} \left(y_{i1} \frac{\partial}{\partial x_1} + y_{i2} \frac{\partial}{\partial x_2} + \ldots + y_{id} \frac{\partial}{\partial x_d} \right) v(\mathbf{x}) \ ,$$

where $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{id})$, $i = 1, 2, \ldots, r$. Then we see that

$$|D^\alpha \hat{v}(\hat{\mathbf{x}})| \le \|D^r \hat{v}(\hat{\mathbf{x}})\| \equiv \sup_{\|\mathbf{y}_i\|=1, 1 \le i \le r} |D^r \hat{v}(\hat{\mathbf{x}})(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_r)| \ .$$

Consequently, we obtain

$$\begin{aligned} |\hat{v}|^2_{H^r(\hat{K})} &= \int_{\hat{K}} \sum_{|\alpha|=r} |D^\alpha \hat{v}(\hat{\mathbf{x}})|^2 d\hat{\mathbf{x}} \\ &\le d^r \int_{\hat{K}} \|D^r \hat{v}(\hat{\mathbf{x}})\|^2 d\hat{\mathbf{x}} \ . \end{aligned} \tag{1.106}$$

It follows from the chain rule that

$$D^r \hat{v}(\hat{\mathbf{x}})(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_r) = D^r v(\mathbf{x})(\mathbf{B}_1 \mathbf{y}_1, \mathbf{B}_1 \mathbf{y}_2, \ldots, \mathbf{B}_1 \mathbf{y}_r) \ ,$$

so that

$$\|D^r \hat{v}(\hat{\mathbf{x}})\| \le \|\mathbf{B}_1\|^r \|D^r v(\mathbf{x})\| \ .$$

Hence it follows that

$$\int_{\hat{K}} \|D^r \hat{v}(\hat{\mathbf{x}})\|^2 d\hat{\mathbf{x}} \leq \|\mathbf{B}_1\|^{2r} \int_{\hat{K}} \|D^r v\big(\mathbf{F}(\hat{\mathbf{x}})\big)\|^2 d\hat{\mathbf{x}} \,. \tag{1.107}$$

Applying a change of variables yields

$$\int_{\hat{K}} \|D^r v\big(\mathbf{F}(\hat{\mathbf{x}})\big)\|^2 d\hat{\mathbf{x}} = |\det\, \mathbf{B}_1|^{-1} \int_K \|D^r v(\mathbf{x})\|^2 d\mathbf{x} \,. \tag{1.108}$$

Because there is a constant $C(d,r)$ such that

$$\|D^r v(\mathbf{x})\| \leq C(d,r) \max_{|\alpha|=r} |D^\alpha v(\mathbf{x})| \,,$$

we have

$$\int_K \|D^r v(\mathbf{x})\|^2 d\mathbf{x} \leq C(d,r) |v|^2_{H^r(K)} \,. \tag{1.109}$$

Combining (1.106)–(1.109) gives (1.104). Inequality (1.105) can be shown in the same way. □

To use Lemma 1.5, it is necessary to estimate the norms $\|\mathbf{B}_1\|$ and $\|\mathbf{B}_1^{-1}\|$ in terms of geometric quantities. For this, we introduce the parameters (cf. Fig. 1.27)

$$\begin{aligned} &h_K = \text{diam(K)}, \quad \mathrm{h}_{\hat{\mathrm{K}}} = \text{diam}(\hat{\mathrm{K}}), \\ &\rho_K = \text{the diameter of the largest circle inscribed in } K, \\ &\rho_{\hat{K}} = \text{the diameter of the largest circle inscribed in } \hat{K}. \end{aligned}$$

Lemma 1.6. *Let K and $\hat{K}$ be two affine-equivalent open subsets of $\mathbb{R}^d$. Then*

$$\|\mathbf{B}_1\| \leq \frac{h_K}{\rho_{\hat{K}}}, \quad \|\mathbf{B}_1^{-1}\| \leq \frac{h_{\hat{K}}}{\rho_K} \,.$$

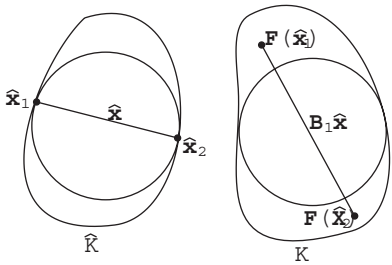

Fig. 1.27. Affine-equivalent sets

Proof. The matrix norm of $\mathbf{B}_1$ can be defined as

$$\|\mathbf{B}_1\| = \frac{1}{\rho_{\hat{K}}} \sup_{\|\hat{\mathbf{x}}\|=\rho_{\hat{K}}} \|\mathbf{B}_1\hat{\mathbf{x}}\| .$$

For a given $\hat{\mathbf{x}}$ satisfying $\|\hat{\mathbf{x}}\| = \rho_{\hat{K}}$, there are two points $\hat{\mathbf{x}}_1$, $\hat{\mathbf{x}}_2 \in \bar{\hat{K}}$ such that $\hat{\mathbf{x}}_1 - \hat{\mathbf{x}}_2 = \hat{\mathbf{x}}$ (cf. Fig. 1.27). Because $\mathbf{B}_1\hat{\mathbf{x}} = \mathbf{F}(\hat{\mathbf{x}}_1) - \mathbf{F}(\hat{\mathbf{x}}_2)$ with $\mathbf{F}(\hat{\mathbf{x}}_1)$, $\mathbf{F}(\hat{\mathbf{x}}_2) \in \bar{K}$, we see that $\|\mathbf{B}_1\hat{\mathbf{x}}\| \le h_K$. Thus the bound on $\|\mathbf{B}_1\|$ follows. The bound on $\|\mathbf{B}_1^{-1}\|$ can be shown similarly. □

We now prove a *local interpolation estimate* of errors.

Theorem 1.7. *For integers $r \ge 0$ and $m \ge 0$ with $r + 1 \ge m$, let $\hat{\pi} : H^{r+1}(\hat{K}) \to H^m(\hat{K})$ be a linear mapping satisfying*

$$\hat{\pi}\hat{w} = \hat{w} \qquad \forall \hat{w} \in P_r(\hat{K}) . \tag{1.110}$$

For any open set K that is affine-equivalent to $\hat{K}$, define the mapping π_K by

$$\widehat{\pi_K v} = \hat{\pi}\hat{v}, \qquad \hat{v} \in H^{r+1}(\hat{K}),\ v \in H^{r+1}(K) . \tag{1.111}$$

Then there is a constant $C = C(\hat{\pi}, \hat{K})$ such that

$$|v - \pi_K v|_{H^m(K)} \le C\, \frac{h_K^{r+1}}{\rho_K^m} |v|_{H^{r+1}(K)}, \quad v \in H^{r+1}(K) . \tag{1.112}$$

Proof. Applying the polynomial invariance (1.110), we deduce that

$$\hat{v} - \hat{\pi}\hat{v} = (I - \hat{\pi})(\hat{v} + \hat{w}), \quad \hat{v} \in H^{r+1}(K),\ \hat{w} \in P_r(\hat{K}) ,$$

where $I : H^{r+1}(\hat{K}) \to H^m(\hat{K})$ is the identity mapping. Then, using Lemma 1.4, we see that

$$\begin{aligned} |\hat{v} - \hat{\pi}\hat{v}|_{H^m(\hat{K})} &\le \|I - \hat{\pi}\| \inf_{\hat{w}\in P_r(\hat{K})} \|\hat{v} + \hat{w}\|_{H^{r+1}(\hat{K})} \\ &\le C(\hat{\pi}, \hat{K}) |\hat{v}|_{H^{r+1}(\hat{K})} . \end{aligned} \tag{1.113}$$

It follows from (1.111) that

$$\widehat{(v - \pi_K v)} = \hat{v} - \hat{\pi}\hat{v} ,$$

so that, by (1.105),

$$|v - \pi_K v|_{H^m(K)} \le C\|\mathbf{B}_1^{-1}\|^m |\det\, \mathbf{B}_1|^{1/2} |\hat{v} - \hat{\pi}\hat{v}|_{H^m(\hat{K})} . \tag{1.114}$$

Next, using (1.104), we see that

$$|\hat{v}|_{H^{r+1}(\hat{K})} \le C\|\mathbf{B}_1\|^{r+1} |\det\, \mathbf{B}_1|^{-1/2} |v|_{H^{r+1}(K)} . \tag{1.115}$$

Finally, combining (1.113)–(1.115) and exploiting Lemma 1.6, we obtain the desired result (1.112). □

As an application of Theorem 1.7, let us consider the case where K_h is a triangulation of a polygon Ω into triangles, and V_h is given by

$$V_h = \{v \in H^1(\Omega) : v|_K \in P_r(K),\ K \in K_h\}, \qquad r \geq 1 .$$

For any $u \in C^0(\bar{\Omega})$, let $\pi_h u \in V_h$ be defined using the degrees of freedom in V_h (cf. Sect. 1.4).

Corollary 1.8. *For $K \in K_h$, assume that $u \in C^0(\bar{K}) \cap H^s(K)$, $1 \leq s \leq r+1$. Then*

$$\begin{aligned} &\|u - \pi_h u\|_{L^2(K)} \leq C h_K^s |u|_{H^s(K)}, \quad 1 \leq s \leq r+1 , \\ &|u - \pi_h u|_{H^1(K)} \leq C \frac{h_K^s}{\rho_K} |u|_{H^s(K)}, \quad 1 \leq s \leq r+1 . \end{aligned} \tag{1.116}$$

In the triangular case, the reference triangle $\hat{K}$ in the $\hat{\mathbf{x}}$-plane has vertices $(0,0)$, $(1,0)$, and $(0,1)$, and any triangle $K \in K_h$ is affine-equivalent to $\hat{K}$ (cf. Exercise 1.32). Then this corollary follows from Theorem 1.7 with $m = 0$ or 1. We emphasize that the constant C in (1.116) depends only the polynomial degree $r \geq 1$, but not on the function u and the mesh size h.

If V_h is given by

$$V_h = \{v \in H^2(\Omega) : v|_K \in P_r(K),\ K \in K_h\} ,$$

then Theorem 1.7 with $m = 2$ can be used to obtain

$$|u - \pi_h u|_{H^2(K)} \leq C \frac{h_K^s}{\rho_K^2} |u|_{H^s(\Omega)}, \quad 2 \leq s \leq r+1,\ K \in K_h , \tag{1.117}$$

provided $u \in C^1(\bar{K}) \cap H^s(K)$, $K \in K_h$, $2 \leq s \leq r+1$.

1.9.2 Error Estimates for Elliptic Problems

It follows from Céa's lemma (Theorem 1.3) that

$$\|p - p_h\|_V \leq C \|p - v\|_V \qquad \forall v \in V_h ,$$

so that, with $v = \pi_h p \in V_h$ (the interpolant of p),

$$\|p - p_h\|_V \leq C \|p - \pi_h p\|_V . \tag{1.118}$$

Theorem 1.9. *For Example* 1.2 *in Sect.* 1.3.3, *with $V = H_0^1(\Omega)$ and*

$$V_h = \{v \in H_0^1(\Omega) : v|_K \in P_r(K),\ K \in K_h\}, \qquad r \geq 1 ,$$

if K_h is a shape-regular triangulation of Ω into triangles, then

$$\|p - p_h\|_{H^1(\Omega)} \le Ch^{s-1}|p|_{H^s(\Omega)}, \quad 2 \le s \le r+1 \,. \tag{1.119}$$

Proof. Note that $\pi_h|_K = \pi_K$. As a result, this theorem follows from (1.118), Corollary 1.8 by piecing all the triangles together, and applying condition (1.52), where the constant β_1 in (1.52) is absorbed into the constant C in (1.119). □

Similarly, for Example 1.5, with $V = H_0^2(\Omega)$ and

$$V_h = \{v \in H_0^2(\Omega) : v|_K \in P_5(K),\, K \in K_h\} \,,$$

it follows from (1.117) and (1.118) that

$$\|p - p_h\|_{H^2(\Omega)} \le Ch^4|p|_{H^6(\Omega)} \,. \tag{1.120}$$

1.9.3 L^2-Error Estimates

An error bound in terms of the H^1-norm is given in (1.119). We now obtain an estimate in the L^2-norm using a *duality argument* that has been called *Aubin-Nitsche's technique* (Aubin, 1967; Nitsche, 1968). For this, we assume that Ω is a convex polygon. In this case, there is a constant C independent of f such that the solution to the Poisson equation (1.16) satisfies (Girault-Raviart, 1981)

$$\|p\|_{H^2(\Omega)} \le C\|f\|_{L^2(\Omega)} \,. \tag{1.121}$$

Namely, if $f \in L^2(\Omega)$, then $p \in H^2(\Omega)$. If Ω is polygonal, convexity is required for (1.121). If the boundary Γ of Ω is a smooth curve (particularly, without corners or cups), convexity is not required. In the smooth case, if $f \in H^s(\Omega)$, then $p \in H^{s+2}(\Omega)$ for $s = 0, 1, \ldots$, and

$$\|p\|_{H^{s+2}(\Omega)} \le C\|f\|_{L^s(\Omega)} \,. \tag{1.122}$$

This property is called *solution regularity* (Girault-Raviart, 1981). If Γ is not smooth, the regularity result (1.122) may not hold, even for $s = 0$. For example, if Ω has a corner, the solution p to (1.16) or its derivatives can have a singularity at the corner even if f is smooth (e.g., $f \in H^s(\Omega)$ for a large s) (Dauge, 1998).

Lemma 1.10. *Let H be a Hilbert space with the norm $\|\cdot\|_H$ and the scalar product $(\cdot,\cdot)$, and let the imbedding $V \hookrightarrow H$ be continuous in the sense that*

$$\|v\|_H \le C\|v\|_V \qquad \forall v \in V \,.$$

Then, under assumptions (1.39)–(1.42)*, if p and p_h are the respective solutions to* (1.38) *and* (1.45)*,*

$$\|p - p_h\|_H \le C\|p - p_h\|_V \sup_{\psi \in H\setminus\{0\}} \left\{ \frac{1}{\|\psi\|_H} \inf_{v \in V_h} \|\varphi_\psi - v\|_V \right\} , \qquad (1.123)$$

where, for given $\psi \in H$, $\varphi_\psi \in V$ *is the solution of the problem*

$$a\,(w, \varphi_\psi) = (w, \psi) \qquad \forall w \in V\ . \qquad (1.124)$$

Proof. By duality, the norm of an element in a Hilbert space H can be calculated as follows:

$$\|w\|_H = \sup_{\psi \in H\setminus\{0\}} \frac{(w, \psi)}{\|\psi\|_H}\ . \qquad (1.125)$$

We recall (1.50):

$$a(p - p_h, v) = 0 \qquad \forall v \in V_h\ .$$

Then, applying (1.40) and (1.124), we see that

$$\begin{aligned}(p - p_h, \psi) &= a\,(p - p_h, \varphi_\psi) \\ &= a\,(p - p_h, \varphi_\psi - v) \le C\|p - p_h\|_V \|\varphi_\psi - v\|_V, \quad v \in V_h\ .\end{aligned}$$

Consequently, we obtain

$$(p - p_h, \psi) \le C\|p - p_h\|_V \inf_{v \in V_h} \|\varphi_\psi - v\|_V\ ,$$

which, together with the definition (1.125), i.e.,

$$\|p - p_h\|_H = \sup_{\psi \in H\setminus\{0\}} \frac{(p - p_h, \psi)}{\|\psi\|_H}\ ,$$

implies the desired result. □

Theorem 1.11. *For Example* 1.2 *in Sect.* 1.3.3, *with* $V = H_0^1(\Omega)$ *and*

$$V_h = \{v \in H_0^1(\Omega) : v|_K \in P_r(K),\ K \in K_h\}, \qquad r \ge 1\ ,$$

if K_h *is a shape-regular triangulation of* Ω *into triangles and the regularity result* (1.121) *holds, then*

$$\|p - p_h\|_{L^2(\Omega)} \le Ch^{r+1}|p|_{H^{r+1}(\Omega)}, \qquad r \ge 1\ . \qquad (1.126)$$

Proof. For Example 1.2, we choose

$$H = L^2(\Omega)\ .$$

Then it follows from the solution regularity (1.121), Lemma 1.10, and the second approximation property in (1.116) with $s = 2$ that

$$\|p - p_h\|_{L^2(\Omega)} \leq Ch\|p - p_h\|_{H^1(\Omega)} ,$$

which, together with (1.119), implies the desired result. □

The estimates in (1.119) and (1.126) do not exclude the possibility that the error is large at certain points. To prevent this, it is necessary to bound the error using the L^∞-norm. Under the condition that the solution $p \in W^{2,\infty}(\Omega)$, $\Omega \subset \mathbb{R}^2$, the following estimates (with $r = 1$) can be shown (Ciarlet, 1978):

$$\begin{aligned} \|p - p_h\|_{L^\infty(\Omega)} &\leq Ch^2|\ln h|^{3/2}|p|_{W^{2,\infty}(\Omega)} , \\ |p - p_h|_{W^{1,\infty}(\Omega)} &\leq Ch|\ln h||p|_{W^{2,\infty}(\Omega)} . \end{aligned} \tag{1.127}$$

1.10 Linear System Solution Techniques

In Sects. 1.1 and 1.7, we have seen that the application of the finite element method to a stationary problem or to an implicit scheme of a transient problem produces a linear system of equations of the form

$$\mathbf{Ap} = \mathbf{f} , \tag{1.128}$$

where the $M \times M$ matrix $\mathbf{A} = (a_{ij})$ is symmetric, positive definite, and sparse. In this section, we review two basic solution techniques for solving (1.128), one based on *Gaussian elimination* or *Cholesky's approach* and the other being the *conjugate gradient algorithm*. These two techniques are sufficient for completing the exercises given in Sect. 1.12 that are related to the numerical solution of sample problems. For more information on solution algorithms for linear systems, refer to the books by Axelsson (1994) and Golub-van Loan (1996), for example.

1.10.1 Gaussian Elimination

A direct method, Gaussian elimination, is studied first for the case where $\mathbf{A}$ is a tridiagonal matrix, and then for the case where $\mathbf{A}$ is a general positive definite matrix.

1.10.1.1 A Tridiagonal Case

In Sect. 1.1.1, we have seen that the matrix $\mathbf{A}$ in the one-dimensional case is tridiagonal; i.e., it has the form

$$\mathbf{A} = \begin{pmatrix} a_1 & b_1 & 0 & \dots & 0 & 0 \\ c_2 & a_2 & b_2 & \dots & 0 & 0 \\ 0 & c_3 & a_3 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & a_{M-1} & b_{M-1} \\ 0 & 0 & 0 & \dots & c_M & a_M \end{pmatrix} .$$

System (1.128) with such a tridiagonal matrix can be solved either by a direct elimination algorithm or by an iterative algorithm. For one-dimensional problems, no known iterative algorithm can compete with direct elimination. Hence we consider only direct elimination for a tridiagonal system.

In general, for a positive definite matrix $\mathbf{A}$, it has a unique *LU-factorization* (Golub-Van Loan, 1996)

$$\mathbf{A} = \mathbf{LQ} \,, \tag{1.129}$$

where $\mathbf{L} = (l_{ij})$ is a *lower triangular* $M \times M$ matrix, i.e., $l_{ij} = 0$ if $j > i$, and $\mathbf{Q} = (q_{ij})$ is an *upper triangular* $M \times M$ matrix, i.e., $q_{ij} = 0$ if $j < i$. For the special tridiagonal matrix under consideration, the matrices $\mathbf{L}$ and $\mathbf{Q}$ are sought to have the form

$$\mathbf{L} = \begin{pmatrix} l_1 & 0 & 0 & \dots & 0 & 0 \\ c_2 & l_2 & 0 & \dots & 0 & 0 \\ 0 & c_3 & l_3 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & l_{M-1} & 0 \\ 0 & 0 & 0 & \dots & c_M & l_M \end{pmatrix} ,$$

and

$$\mathbf{Q} = \begin{pmatrix} 1 & q_1 & 0 & \dots & 0 & 0 \\ 0 & 1 & q_2 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & q_{M-1} \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix} .$$

We note that the lower diagonal of $\mathbf{L}$ is the same as that of $\mathbf{A}$, and the main diagonal of $\mathbf{Q}$ is set to have all ones. The identity (1.129) gives $2M-1$ equations for the unknowns: $l_1, l_2, \dots, l_M$ and $q_1, q_2, \dots, q_{M-1}$. The solution is

$$\begin{aligned} &l_1 = a_1, \\ &q_{i-1} = b_{i-1}/l_{i-1}, \qquad i = 2, 3, \dots, M \,, \\ &l_i = a_i - c_i q_{i-1}, \qquad i = 2, 3, \dots, M \,. \end{aligned}$$

With the factorization (1.129), system (1.128) can be easily solved using *forward elimination* and *backward substitution*:

$$\begin{aligned} &\mathbf{Lv} = \mathbf{f} \,, \\ &\mathbf{Qp} = \mathbf{v} \,. \end{aligned} \tag{1.130}$$

Namely, since $\mathbf{L}$ is lower triangular, the first equation in (1.130) can be solved by forward elimination:

$$v_1 = \frac{f_1}{l_1}, \quad v_i = \frac{f_i - c_i v_{i-1}}{l_i}, \qquad i = 2, 3, \ldots, M \ .$$

Next, since $\mathbf{Q}$ is upper triangular, the second equation in (1.130) can be solved by backward substitution:

$$p_M = v_M, \quad p_i = v_i - q_i p_{i+1}, \qquad i = M-1, M-2, \ldots, 1 \ .$$

As discussed in Sect. 1.1.1, for many practical problems, the matrix $\mathbf{A}$ is symmetric:

$$\mathbf{A} = \begin{pmatrix} a_1 & b_1 & 0 & \ldots & 0 & 0 \\ b_1 & a_2 & b_2 & \ldots & 0 & 0 \\ 0 & b_2 & a_3 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & a_{M-1} & b_{M-1} \\ 0 & 0 & 0 & \ldots & b_{M-1} & a_M \end{pmatrix} .$$

In the symmetric case, $\mathbf{A}$ is factorized by

$$\mathbf{A} = \mathbf{L}\mathbf{L}^T \ ,$$

where $\mathbf{L}^T$ is the transpose of $\mathbf{L}$ and $\mathbf{L}$ now takes the form

$$\mathbf{L} = \begin{pmatrix} l_1 & 0 & 0 & \ldots & 0 & 0 \\ q_1 & l_2 & 0 & \ldots & 0 & 0 \\ 0 & q_2 & l_3 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & l_{M-1} & 0 \\ 0 & 0 & 0 & \ldots & q_{M-1} & l_M \end{pmatrix} .$$

With this factorization, the elements are computed as follows:

$$\begin{aligned} &l_1 = \sqrt{a_1}, \\ &q_i = b_i / l_i, && i = 1, 2, \ldots, M-1 \ , \\ &l_{i+1} = \sqrt{a_{i+1} - q_i^2}, && i = 1, 2, \ldots, M-1 \ . \end{aligned}$$

Now, system (1.128) can be solved in a similar forward elimination and backward substitution fashion.

In using the LU factorization algorithm we must assure that

$$l_i \neq 0, \qquad i = 1, 2, \ldots, M \ .$$

It can be shown that if $\mathbf{A}$ is symmetric positive definite, $l_i > 0, i = 1, 2, \ldots, M$ (Axelsson, 1994; Golub-van Loan, 1996). These quantities l_i are referred to as the *pivots*.

1.10.1.2 A General Case

As noted above, for a general positive definite matrix $\mathbf{A}$, it has the factorization (1.129), where $\mathbf{L} = (l_{ij})$ is a unit lower triangular matrix, i.e., $l_{ii} = 1$ and $l_{ij} = 0$ if $j > i$, and $\mathbf{Q} = (q_{ij})$ is an upper triangular, i.e., $q_{ij} = 0$ if $j < i$. We give the computation of $\mathbf{L}$ and $\mathbf{Q} = \mathbf{A}^{(M)}$ where the matrices $\mathbf{A}^{(k)}$, $k = 1, 2, \ldots, M$, are successively computed using *Gaussian elimination*:

Set $\mathbf{A}^{(1)} = \mathbf{A}$;

Given $\mathbf{A}^{(k)}$ of the form

$$\mathbf{A}^{(k)} = \begin{pmatrix} a_{11}^{(k)} & a_{12}^{(k)} & \ldots & a_{1k}^{(k)} & \ldots & a_{1M}^{(k)} \\ 0 & a_{22}^{(k)} & \ldots & a_{2k}^{(k)} & \ldots & a_{2M}^{(k)} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & a_{kk}^{(k)} & \ldots & a_{kM}^{(k)} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & a_{Mk}^{(k)} & \ldots & a_{MM}^{(k)} \end{pmatrix},$$

set $l_{ik} = -a_{ik}^{(k)}/a_{kk}^{(k)}, \quad i = k+1, k+2, \ldots M$,

calculate $\mathbf{A}^{(k+1)} = \left(a_{ij}^{(k+1)}\right)$ by

$a_{ij}^{(k+1)} = a_{ij}^{(k)}, \quad i = 1, 2, \ldots, k$ or $j = 1, 2, \ldots, k-1$,

$a_{ij}^{(k+1)} = a_{ij}^{(k)} + l_{ik} a_{kj}^{(k)}, \quad i = k+1, \ldots, M,\ j = k, \ldots, M$.

Again, if $\mathbf{A}$ is symmetric positive definite, $a_{kk}^{(k)} > 0$, $k = 1, 2, \ldots, M$.

In the case where $\mathbf{A}$ is symmetric, it can be alternatively factorized as

$$\mathbf{A} = \mathbf{B}\mathbf{B}^T; \tag{1.131}$$

i.e.,

$$\sum_{k=1}^{j} b_{ik} b_{jk} = a_{ij}, \qquad j = 1, 2, \ldots, i,\ i = 1, 2, \ldots, M .$$

In this case, the entries b_{ij} of $\mathbf{B}$ in (1.131) can be computed directly using *Cholesky's approach*, $i = 1, 2, \ldots, M$,

$$b_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} b_{ik}^2} ,$$

$$b_{ij} = \left(a_{ij} - \sum_{k=1}^{j-1} b_{ik} b_{jk}\right) \Big/ b_{jj}, \qquad j = 1, 2, \ldots, i-1 .$$

We note that in the above computation of $\mathbf{B}$, M square root operations are required. To get around this, we can write $\mathbf{B}$ as

$$\mathbf{B} = \tilde{\mathbf{B}}\mathbf{D}, \tag{1.132}$$

where $\tilde{\mathbf{B}}$ is a unit lower triangular matrix (i.e., $\tilde{b}_{ii} = 1$, $i = 1, 2, \ldots, M$) and $\mathbf{D}$ is a diagonal matrix:

$$\mathbf{D} = \operatorname{diag}\left(\sqrt{d_1}, \sqrt{d_2}, \ldots, \sqrt{d_M}\right) .$$

In this factorization we see that

$$\sum_{k=1}^{j} \tilde{b}_{ik} d_k \tilde{b}_{jk} = a_{ij}, \qquad j = 1, 2, \ldots, i, \ i = 1, 2, \ldots, M ,$$

which implies, for $i = 1, \ldots, M$,

$$\begin{aligned} d_i &= a_{ii} - \sum_{k=1}^{i-1} \tilde{b}_{ik}^2 d_k , \\ \tilde{b}_{ij} &= \left(a_{ij} - \sum_{k=1}^{j-1} \tilde{b}_{ik} d_k \tilde{b}_{jk} \right) \Big/ d_j, \qquad j = 1, 2, \ldots, i-1 . \end{aligned} \tag{1.133}$$

The number of arithmetic operations in (1.133) for a symmetric matrix $\mathbf{A}$ is asymptotically of the order $M^3/6$. If the matrix $\mathbf{A}$ is sparse, then one can greatly reduce the number of operations by using the sparsity. This is the case when $\mathbf{A}$ is a *band matrix*. That is, for the ith row, there is an integer m_i such that

$$a_{ij} = 0 \quad \text{if } j < m_i, \quad i = 1, 2, \ldots, M .$$

Note that m_i is the column number of the first nonzero entry in the ith row. Then the *band width* L_i of the ith row satisfies

$$L_i = i - m_i, \qquad i = 1, 2, \ldots, M .$$

We warn the reader that $2L_i + 1$ is sometimes called the band width. It can be checked from (1.133) that $\mathbf{A}$ and $\tilde{\mathbf{B}}$ have the same number m_i. Thus, in the band case, (1.133) can be modified to ($i = 1, 2, \ldots, M$)

$$\begin{aligned} d_i &= a_{ii} - \sum_{k=m_i}^{i-1} \tilde{b}_{ik}^2 d_k , \\ \tilde{b}_{ij} &= \left(a_{ij} - \sum_{k=\max(m_i, m_j)}^{j-1} \tilde{b}_{ik} d_k \tilde{b}_{jk} \right) \Big/ d_j , \\ & \qquad\qquad j = m_i, 2, \ldots, i-1 . \end{aligned} \tag{1.134}$$

We remark that the number of arithmetic operations to factor a band matrix is asymptotically of the order $ML^2/2$, where $L = \max_{1\le i\le M} L_i$; see Exercise 1.33. This number is much smaller than $M^3/6$ if L is smaller than M. In the finite element method, we have

$$a_{ij} = a(\varphi_i, \varphi_j), \qquad i,\ j = 1, 2, \ldots, M\ ,$$

where $\{\varphi_i\}_{i=1}^M$ is a basis of V_h. Then we see that

$$L = \max\{|i-j| : \varphi_i \text{ and } \varphi_j \text{ correspond to degrees of freedom belonging to the same element}\}\ .$$

Consequently, the band width depends on the enumeration of nodes. If direct elimination is used, the nodes should be enumerated in such a way that the band width is as small as possible. For example, with a vertical enumeration of nodes in Fig. 1.28, L is 5 (assuming that one degree of freedom is associated with each node). With a horizontal enumeration, L would be 10.

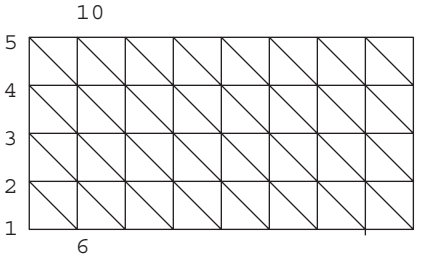

Fig. 1.28. An example of enumeration

Now, we return to (1.128) with the factorization (1.131) of $\mathbf{A}$, where $\mathbf{B}$ is given by (1.132). With this factorization, system (1.128) becomes

$$\begin{aligned} &\tilde{\mathbf{B}}\mathbf{D}^2\mathbf{v} = \mathbf{f}\ , \\ &\tilde{\mathbf{B}}^T\mathbf{p} = \mathbf{v}\ . \end{aligned} \tag{1.135}$$

We emphasize that these two systems are triangular. The first system is

$$\sum_{k=1}^{i} \tilde{b}_{ik} d_k v_k = f_i, \qquad i = 1, 2, \ldots, M\ .$$

Thus forward elimination implies

$$v_1 = \frac{f_1}{d_1}, \quad v_i = \frac{f_i - \sum_{k=1}^{i-1} \tilde{b}_{ik} d_k v_k}{d_i}, \qquad i = 2, 3, \ldots, M\ . \tag{1.136}$$

Similarly, the second system is solved by backward substitution:

$$p_M = v_M, \quad p_i = v_i - \sum_{k=i+1}^{M} \tilde{b}_{ki}\, p_k \ , \\ i = M-1, M-2, \ldots, 1 \ . \tag{1.137}$$

If $\mathbf{A}$ is a band matrix, we apply (1.134) to (1.136) to give

$$v_1 = \frac{f_1}{d_1}, \quad v_i = \frac{f_i - \sum_{k=m_i}^{i-1} \tilde{b}_{ik} d_k v_k}{d_i}, \qquad i = 2, 3, \ldots, M \ .$$

Also, it follows from (1.137) that

$$\begin{aligned} & p_M = v_M \ , \\ & p_{M-1} = v_{M-1} - \tilde{b}_{M,M-1}\, p_M \ , \\ & p_{M-2} = v_{M-2} - \tilde{b}_{M-1,M-2}\, p_{M-1} - \tilde{b}_{M,M-2}\, p_M \ , \\ & \ldots \\ & p_1 = v_1 - \tilde{b}_{2,1}\, p_2 - \tilde{b}_{3,1}\, p_3 - \ldots - \tilde{b}_{M,1}\, p_M \ . \end{aligned}$$

Note that one subtracts $\tilde{b}_{M,k}\, p_M$ from v_k, $k = M-1, M-2, \ldots, 1$. Due to the band structure of $\mathbf{A}$, i.e.,

$$\tilde{b}_{M,k} = 0 \quad \text{if } k < m_M \ ,$$

$\tilde{b}_{M,k}\, p_M$ is subtracted from v_k only when $k \geq m_M$. As a result, one can first find v_k successively by

$$v_k = v_k - \tilde{b}_{ik}\, p_i, \quad k = m_i, m_i + 1, \ldots, i-1, \ i = M, M-1, \ldots, 1 \ ,$$

and then obtain

$$p_i = v_i, \qquad i = M, M-1, \ldots, 1 \ .$$

1.10.2 The Conjugate Gradient Algorithm

We recall that the *condition number* of a *symmetric* matrix $\mathbf{A}$ is defined by

$$\text{cond}(\mathbf{A}) = \frac{\text{the largest eigenvalue of } \mathbf{A}}{\text{the smallest eigenvalue of } \mathbf{A}} \ .$$

For the matrix $\mathbf{A}$ in system (1.128) (for second order problems), it has a condition number proportional to h^{-2} (cf. (1.138)). For the application of the finite element method to a large-scale problem, it would be very expensive to solve the resulting system of equations via a direct method like Gaussian elimination discussed in the previous subsection. Consequently, the usual technique to obtain the solution of a large-scale system is to use an iterative approach. In this section, we consider an application of the *conjugate gradient algorithm* to system (1.128).

1.10.2.1 Condition Numbers

If $\mathbf{A}$ in (1.128) is the stiffness matrix arising from the discretization of an elliptic problem of order $2m$, then the condition number $\text{cond}(\mathbf{A})$ under assumptions (1.52) and (1.78) is estimated by

$$\text{cond}(\mathbf{A}) = \mathcal{O}(h^{-2m}), \quad m \geq 1 \,. \tag{1.138}$$

As an example, we show (1.138) for Example 1.2; i.e., $m = 1$ and the finite element space V_h is the space of piecewise linear polynomials on a triangulation K_h. Let $\{\varphi_i\}_{i=1}^M$ be the basis of V_h introduced in Sect. 1.1.2. For $v \in V_h$, set

$$v = \sum_{i=1}^{M} v_i \varphi_i, \quad \mathbf{v} = (v_1, v_2, \ldots, v_M) \,.$$

Lemma 1.12. *There exist positive constants C, C_1, and C_2, depending only on β_1 and β_2 in* (1.52) *and* (1.78), *such that*

$$\begin{aligned} &\|\nabla v\|_{L^2(\Omega)} \leq C h^{-1} \|v\|_{L^2(\Omega)}, \qquad v \in V_h \,, \\ &C_1 h \|\mathbf{v}\| \leq \|v\|_{L^2(\Omega)} \leq C_2 h \|\mathbf{v}\|, \quad v \in V_h \,, \end{aligned} \tag{1.139}$$

where $\|\mathbf{v}\|^2 = |v_1|^2 + |v_2|^2 + \ldots + |v_M|^2$.

The first inequality is called an *inverse inequality* or *inverse estimate*. The L^2-norm of the gradient of v is bounded by the L^2-norm of v itself at the price of a factor proportional to h^{-1}. We prove only this inequality; the second inequality is proven in the same way.

Proof. It suffices to prove that for each triangle $K \in K_h$,

$$\|\nabla v\|_{L^2(K)} \leq C h_K^{-1} \|v\|_{L^2(K)}, \qquad v \in P_1(K) \,, \tag{1.140}$$

with C independent of K and v; the desired result follows from summation over $K \in K_h$ and (1.78).

We first show (1.140) when $K = \hat{K}$ is the reference triangle with vertices $(1,0)$, $(0,1)$, and $(0,0)$. Let $\hat{\lambda}_1$, $\hat{\lambda}_2$, and $\hat{\lambda}_3$ be the usual basis functions of $P_1(\hat{K})$. For

$$\hat{v}(\hat{x}) = \sum_{i=1}^{3} \hat{v}_i \hat{\lambda}_i(\hat{x}), \qquad \hat{x} \in \hat{K}, \, \hat{\mathbf{v}} = (\hat{v}_1, \hat{v}_2, \hat{v}_3) \,,$$

we define

$$g(\hat{\mathbf{v}}) = \frac{\|\nabla \hat{v}\|_{L^2(\hat{K})}}{\|\hat{v}\|_{L^2(\hat{K})}}, \qquad \hat{v} \in P_1(\hat{K}), \, \|\hat{v}\|_{L^2(\hat{K})} \neq 0 \,.$$

Note that

$$g(\gamma\hat{\mathbf{v}}) = g(\hat{\mathbf{v}}) \qquad \forall\gamma \in \mathbb{R},\ \gamma \neq 0\ ;$$

namely, the function g is *homogeneous of degree zero*. Thus it suffices to prove that there is a constant $C > 0$ such that

$$g(\hat{\mathbf{v}}) \leq C \qquad \forall\hat{\mathbf{v}} \in \mathbf{B}_2 = \left\{\hat{\mathbf{v}} \in \mathbb{R}^3 : \|\hat{\mathbf{v}}\| = 1\right\}\ . \tag{1.141}$$

Because g is continuous on $\mathbf{B}_2$ and $\mathbf{B}_2$ is compact (bounded and closed) in $\mathbb{R}^3$, g achieves a maximum on $\mathbf{B}_2$. This proves (1.141), and thus (1.140) when $K = \hat{K}$.

We now show (1.140) when K is an arbitrary triangle with vertices $\mathbf{m}_i$, $i = 1, 2, 3$; see Fig. 1.29. Introduce the linear mapping $\mathbf{F} : \hat{K} \to K$ by

$$\mathbf{x} = \mathbf{F}(\hat{\mathbf{x}}) = \mathbf{m}_1 + (\mathbf{m}_2 - \mathbf{m}_1)\hat{x}_1 + (\mathbf{m}_3 - \mathbf{m}_1)\hat{x}_2\ ,$$

where $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2)$. For any $v \in P_1(K)$, we define

$$\hat{v}(\hat{\mathbf{x}}) = v\big(F(\hat{\mathbf{x}})\big), \qquad \hat{\mathbf{x}} \in \hat{K}\ .$$

The chain rule gives

$$\frac{\partial v}{\partial x_i} = \frac{\partial}{\partial x_i}\left(\hat{v}\big(\mathbf{F}^{-1}(\mathbf{x})\big)\right) = \frac{\partial\hat{v}}{\partial\hat{x}_1}\frac{\partial\hat{x}_1}{\partial x_i} + \frac{\partial\hat{v}}{\partial\hat{x}_2}\frac{\partial\hat{x}_2}{\partial x_i}\ ,$$

for $i = 1, 2$. Consequently, we see that

$$\nabla v = \mathbf{G}^{-T}\nabla\hat{v}\ ,$$

where $\mathbf{G}^{-T}$ is the transpose of the Jacobian of $\mathbf{F}^{-1}$:

$$\mathbf{G}^{-T} = \begin{pmatrix} \dfrac{\partial\hat{x}_1}{\partial x_1} & \dfrac{\partial\hat{x}_2}{\partial x_1} \\[2ex] \dfrac{\partial\hat{x}_1}{\partial x_2} & \dfrac{\partial\hat{x}_2}{\partial x_2} \end{pmatrix}\ .$$

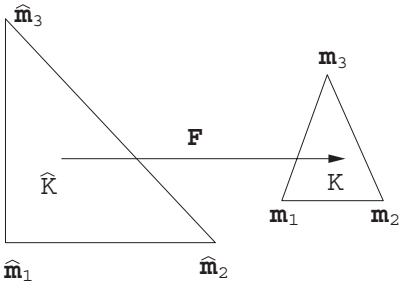

Fig. 1.29. A mapping $\mathbf{F}$

Thus we see that

$$\int_K |\nabla v|^2 \, d\mathbf{x} = \int_{\hat{K}} \left|\mathbf{G}^{-T} \nabla \hat{v}\right|^2 |\det \mathbf{G}| \, d\hat{\mathbf{x}} \, ,$$

where $|\det \mathbf{G}|$ is the absolute value of the determinant of the Jacobian $\mathbf{G}$:

$$\mathbf{G} = \begin{pmatrix} \dfrac{\partial x_1}{\partial \hat{x}_1} & \dfrac{\partial x_1}{\partial \hat{x}_2} \\ \\ \dfrac{\partial x_2}{\partial \hat{x}_1} & \dfrac{\partial x_2}{\partial \hat{x}_2} \end{pmatrix} .$$

Using the facts that $\|\mathbf{m}_i - \mathbf{m}_1\| \leq C h_K$, $i = 2, 3$, and (1.140) holds when $K = \hat{K}$, we obtain

$$\int_K |\nabla v|^2 \, d\mathbf{x} \leq C \int_{\hat{K}} |\nabla \hat{v}|^2 \, d\hat{\mathbf{x}} \leq C \int_{\hat{K}} \hat{v}^2 \, d\hat{\mathbf{x}} \leq C h_K^{-2} \int_K v^2 \, d\mathbf{x} \, ,$$

which implies (1.140) for an arbitrary $K \in K_h$. □

Theorem 1.13. *For Example* 1.2 *in Sect.* 1.3.3, *with* $V = H_0^1(\Omega)$ *and*

$$V_h = \{v \in H_0^1(\Omega) : v|_K \in P_1(K), \, K \in K_h\} \, ,$$

if conditions (1.52) *and* (1.78) *hold, then*

$$\text{cond}(\mathbf{A}) = \mathcal{O}(h^{-2}) \, . \tag{1.142}$$

Proof. For

$$v = \sum_{i=1}^{M} v_i \varphi_i, \quad \mathbf{v} = (v_1, v_2, \ldots, v_M) \, ,$$

we recall that

$$a(v, v) = \mathbf{v}^T \mathbf{A} \mathbf{v} \, .$$

As a result, using (1.139), we see that

$$\frac{\mathbf{v}^T \mathbf{A} \mathbf{v}}{\|\mathbf{v}\|^2} = \frac{a(v, v)}{\|\mathbf{v}\|^2} \leq C h^{-2} \frac{\|v\|_{L^2(\Omega)}^2}{\|\mathbf{v}\|^2} \leq C \, .$$

On the other hand, it follows from Poincaré's inequality (1.36) that

$$\frac{\mathbf{v}^T \mathbf{A} \mathbf{v}}{\|\mathbf{v}\|^2} = \frac{a(v, v)}{\|\mathbf{v}\|^2} \geq \frac{C_1 \|v\|_{L^2(\Omega)}^2}{\|\mathbf{v}\|^2} \geq C_1 h^2 \quad \forall v \in V_h \subset H_0^1(\Omega) \, .$$

Hence the largest eigenvalue of $\mathbf{A}$ is bounded above by C and the smallest eigenvalue of $\mathbf{A}$ is bounded below by $C_1 h^2$. Therefore, $\text{cond}(\mathbf{A}) \leq C h^{-2}$. □

1.10.2.2 The Algorithm

Since $\mathbf{A}$ is symmetric positive definite, it deduces a scalar product $\langle \cdot, \cdot \rangle$ on $\mathbb{R}^M$:

$$\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^T \mathbf{A} \mathbf{w} = \sum_{i,j=1}^{M} v_i a_{ij} w_j, \qquad \mathbf{v},\, \mathbf{w} \in \mathbb{R}^M \ .$$

The norm $\| \cdot \|_{\mathbf{A}}$ corresponding to $\langle \cdot, \cdot \rangle$ is usually called the *energy norm*:

$$\|\mathbf{v}\|_{\mathbf{A}} = \langle \mathbf{v}, \mathbf{v} \rangle^{1/2}, \qquad \mathbf{v} \in \mathbb{R}^M \ .$$

The conjugate gradient algorithm for the solution of (1.128) can be now defined as follows:

Given an initial guess $\mathbf{p}^0 \in \mathbb{R}^M$, set $\mathbf{r}^0 = \mathbf{A}\mathbf{p}^0 - \mathbf{f}$ and $\mathbf{d}^0 = -\mathbf{r}^0$;

For $k = 1, 2, \ldots$, determine $\mathbf{p}^k$ and $\mathbf{d}^k$ by

$$\begin{aligned}
\alpha_{k-1} &= -\frac{\mathbf{r}^{k-1} \cdot \mathbf{d}^{k-1}}{\langle \mathbf{d}^{k-1}, \mathbf{d}^{k-1} \rangle} \ , \\
\mathbf{p}^k &= \mathbf{p}^{k-1} + \alpha_{k-1} \mathbf{d}^{k-1} \ , \\
\mathbf{r}^k &= \mathbf{A}\mathbf{p}^k - \mathbf{f} \ , \\
\beta_{k-1} &= \frac{\langle \mathbf{r}^k, \mathbf{d}^{k-1} \rangle}{\langle \mathbf{d}^{k-1}, \mathbf{d}^{k-1} \rangle} \ , \\
\mathbf{d}^k &= -\mathbf{r}^k + \beta_{k-1} \mathbf{d}^{k-1} \ .
\end{aligned}$$

It can be shown that the conjugate gradient algorithm gives, in the absence of round-off errors, the exact solution after at most M steps; i.e.,

$$\mathbf{A}\mathbf{p}^k = \mathbf{f} \qquad \text{for some } k \le M \ .$$

In practice, the required number of iterations is sometimes smaller than M. In fact, for a given *tolerance* $\epsilon > 0$, to satisfy

$$\|\mathbf{p} - \mathbf{p}^k\|_{\mathbf{A}} \le \epsilon \|\mathbf{p} - \mathbf{p}^0\|_{\mathbf{A}} \ ,$$

it suffices to choose k such that (Axelsson, 1994)

$$k \ge \frac{1}{2} \sqrt{\mathrm{cond}(\mathbf{A})}\, \ln \frac{2}{\epsilon} \ .$$

Hence the required number of iterations for the conjugate gradient algorithm is proportional to $\sqrt{\mathrm{cond}(\mathbf{A})}$. As shown above, in a typical finite element application to a second-order elliptic problem, $\mathrm{cond}(\mathbf{A}) = \mathcal{O}(h^{-2})$, so the required number of iterations is of order $\mathcal{O}(h^{-1})$.

It is possible to reduce the condition number of the problem via a *preconditioning technique*. In fact, one can find a symmetric positive definite matrix $\mathbf{C}_1$ such that

$$\mathbf{C}_1\mathbf{A}\mathbf{p} = \mathbf{C}_1\mathbf{f} \tag{1.143}$$

is much better conditioned than (1.128); i.e., $\text{cond}(\mathbf{C}_1\mathbf{A}) \ll \text{cond}(\mathbf{A})$. The construction of $\mathbf{C}_1$ should be easy. A class of techniques for constructing $\mathbf{C}_1$ are based on *incomplete Cholesky factorization* of $\mathbf{A}$. The resulting ILU preconditioners will make the conjugate gradient algorithm very simple and efficient (Axelsson, 1994). Another class of techniques have been recently developed that are optimal in the sense that the required number of operations is of order $\mathcal{O}(M)$. These techniques are based on the multigrid method (Hackbusch, 1985; Bramble, 1993) and on the domain decomposition method (Smith et al., 1996). Preconditioning techniques will not be studied in this book.

1.11 Bibliographical Remarks

There are numerous books on the finite element method discussed in this chapter (e.g., Strang-Fix, 1973; Ciarlet, 1978; Li et al., 1984; Thomée, 1984; Brenner-Scott, 1994; Johnson, 1994; Braess, 1997; Quarteroni-Valli, 1997). The content of Sects. 1.5 and 1.6 closely follows Johnson (1994). In Sect. 1.7, we briefly treated transient problems. The book by Thomée (1984) exclusively handles time-dependent problems. In Sect. 1.9, we briefly touched on the topic of linear system solution procedures. For more information on this subject, the reader should see the books by Axelsson (1994) and Golub-Van Loan (1996), for example. Finally, for mor information on the finite element theory, the reader should refer to Ciarlet (1978) and Brenner-Scott (1994).

1.12 Exercises

1.1. Consider an elastic bar with tension one, fixed at both ends ($x = 0, 1$) and subject to a transversal load of intensity f (cf. Fig. 1.1). Under the assumption of small displacements, show that the transversal displacement p satisfies problem (1.1).

1.2. Show that if $p \in V = H_0^1(I)$ satisfies (1.3) and if p is twice continuously differentiable, then p satisfies (1.1).

1.3. Write a code to solve the one-dimensional problem (1.1) approximately using the finite element method developed in Sect. 1.1.1. Use the function $f(x) = 4\pi^2 \sin(2\pi x)$ and a uniform partition of $(0, 1)$ with $h = 0.1$. Also, compute the errors

$$\left\| \frac{dp}{dx} - \frac{dp_h}{dx} \right\| = \left(\int_0^1 \left(\frac{dp}{dx} - \frac{dp_h}{dx} \right)^2 dx \right)^{1/2},$$

with $h = 0.1$, 0.01, and 0.001, and compare them. Here p and p_h are the exact and approximate solutions, respectively (cf. Sect. 1.1.1). (If necessary, refer to Sect. 1.10.1.1 for a linear solver.)

1.4. Show Cauchy's inequality (1.10).
1.5. Prove the estimates in (1.13).
1.6. Referring to Sect. 1.1.1, show that the interpolant $\tilde{p}_h \in V_h$ of p defined in (1.12) equals the finite element solution p_h obtained by (1.5).
1.7. Prove Green's formula (1.19) in three space dimensions.
1.8. Carry out the derivation of system (1.22).
1.9. For the following figure:

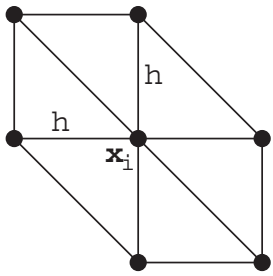

Fig. 1.30. The support of a basis function at node $\mathbf{x}_i$

construct the linear basis function at node $\mathbf{x}_i$ according to the definition in Sect. 1.1.2. Then use this result to show that the stiffness matrix $\mathbf{A}$ in (1.22) for the uniform partition of the unit square $(0,1)\times(0,1)$ given in Fig. 1.7 is determined as in Sect. 1.1.2.

1.10. Write a code to solve the Poisson equation (1.16) approximately using the finite element method developed in Sect. 1.1.2. Use $f(x_1,x_2) = 8\pi^2 \sin(2\pi x_1)$
$\sin(2\pi x_2)$ and a uniform partition of $\Omega = (0,1)\times(0,1)$ as given in Fig. 1.7. Also, compute the errors

$$\|\nabla p - \nabla p_h\| = \left(\int_\Omega |\nabla p - \nabla p_h|^2 \; d\mathbf{x}\right)^{1/2},$$

with $h = 0.1$, 0.01, and 0.001, and compare them. Here p and p_h are the exact and approximate solutions, respectively, and h is the mesh size in the x_1- and x_2-directions. (If necessary, refer to Sect. 1.10.1.2 or Sect. 1.10.2 for a linear solver.)

1.11. Prove (1.26) for (1.25).
1.12. Derive (1.27) from (1.25) in detail.
1.13. Show that for any multi-index α, if $v \in C^{|\alpha|}(\Omega)$, then the weak derivative $D_w^\alpha v$ exists and equals $D^\alpha v$.
1.14. Let $v(x) = 1 - |x|$, $x \in (-1,1)$. Prove that weak derivatives of order greater than one of v do not exist.
1.15. Show the inclusion relations (1.32) and (1.33).
1.16. Let V be a Banach space with norm $\|\cdot\|_V$ and V' be the dual space to V. For $L \in V'$, define

$$\|L\|_{V'} = \sup_{0 \neq v \in V} \frac{L(v)}{\|v\|_V} .$$

Show that $\|\cdot\|_{V'}$ defines a norm on V'.

1.17. Let V_h be a space of piecewise polynomials of degree $r \geq 1$ for a triangulation K_h of a polygon Ω into triangles. Show that $V_h \subset H^1(\Omega)$ if and only if $V_h \subset C^0(\bar{\Omega})$. That is, $V_h \subset H^1(\Omega)$ if and only if the functions in V_h are continuous on $\bar{\Omega}$. Similarly, prove that $V_h \subset H^2(\Omega)$ if and only if $V_h \subset C^1(\bar{\Omega})$; i.e., $V_h \subset H^2(\Omega)$ if and only if the functions in V_h and their first derivatives are continuous on $\bar{\Omega}$.

1.18. Consider the problem with an inhomogeneous boundary condition:

$$\begin{aligned} -\frac{d^2p}{dx^2} &= f(x), \qquad 0 < x < 1 , \\ p(0) &= p_{D0}, \quad p(1) = p_{D1} , \end{aligned}$$

where f is a given real-valued piecewise continuous bounded function on $(0,1)$, and p_{D0} and p_{D1} are real numbers. Write this problem in a variational formulation, and construct a finite element method using piecewise linear functions. Determine the corresponding linear system of algebraic equations for a uniform partition.

1.19. Consider the problem with a Neumann boundary condition at $x = 1$:

$$\begin{aligned} -\frac{d^2p}{dx^2} &= f(x), \qquad 0 < x < 1 , \\ p(0) = \frac{dp}{dx}(1) &= 0 . \end{aligned}$$

Express this problem in a variational formulation, formulate a finite element method using piecewise linear functions, and determine the corresponding linear system of algebraic equations for a uniform partition.

1.20. Give a variational formulation for the problem

$$\begin{aligned} -\Delta p + cp &= f \qquad && \text{in } \Omega , \\ \frac{\partial p}{\partial \boldsymbol{\nu}} &= g && \text{on } \Gamma , \end{aligned}$$

where $c(\mathbf{x}) \geq c_* > 0$, $\mathbf{x} \in \Omega$. Check if conditions (1.39)–(1.41) are satisfied.

1.21. Give a variational formulation for the problem

$$\begin{aligned} -\nabla \cdot (\mathbf{a} \nabla p) + cp &= f \qquad && \text{in } \Omega , \\ p &= g_D && \text{on } \Gamma_D , \\ \gamma p + \mathbf{a} \nabla p \cdot \boldsymbol{\nu} &= g_N && \text{on } \Gamma_N , \end{aligned}$$

where $\mathbf{a}$ is a $d \times d$ matrix ($d = 2$ or 3), c, f, g_D, and g_N are given functions of $\mathbf{x}$, and γ is a constant. Under what conditions on $\mathbf{a}$, c, and γ are the conditions (1.39)–(1.41) satisfied?

1.22. Consider the Poisson equation (1.16) with an inhomogeneous boundary condition, i.e.,

$$\begin{aligned} -\Delta p &= f && \text{in } \Omega \,, \\ p &= g && \text{on } \Gamma \,, \end{aligned}$$

where Ω is a bounded domain in the plane with boundary Γ, and f and g are given. Express this problem in a variational formulation, formulate a finite element method using piecewise linear functions, and determine the corresponding linear system of algebraic equations for a uniform partition of $\Omega = (0,1) \times (0,1)$ as given in Fig. 1.7.

1.23. Consider the problem

$$\begin{aligned} -\Delta p &= f && \text{in } \Omega \,, \\ p &= g_D && \text{on } \Gamma_D \,, \\ \frac{\partial p}{\partial \boldsymbol{\nu}} &= g_N && \text{on } \Gamma_N \,, \end{aligned}$$

where Ω is a bounded domain in the plane with boundary Γ, $\bar{\Gamma} = \bar{\Gamma}_D \cup \bar{\Gamma}_N$, $\Gamma_D \cap \Gamma_N = \emptyset$, and f, g_D, and g_N are given functions. Write down a variational formulation for this problem and formulate a finite element method using piecewise linear functions.

1.24. Set

$$P_r(I) = \{v : \ v \text{ is a polynomial of degree at most } r \text{ on } I\} \ ,$$

where $r = 0, 1, 2, \ldots$ and I is an interval. Show that if v is zero at $r+1$ distinct points on I, then $v \equiv 0$. Hint: If $v \in P_r(I)$ is zero at some point $x_0 \in I$, then $v(x) = (x - x_0)w(x)$, where $w \in P_{r-1}(I)$.

1.25. Let K be a triangle with vertices $\mathbf{m}_i$, $i = 1, 2, 3$. Show that if $v \in P_r(K)$ vanishes on the edge $\mathbf{m}_2\mathbf{m}_3$, then v is of the form

$$v(\mathbf{x}) = \lambda_1(\mathbf{x}) w(\mathbf{x}), \qquad \mathbf{x} \in K \ ,$$

where $w \in P_{r-1}(K)$ and λ_1 is defined as Example 1.6.

1.26. Prove equation (1.58).

1.27. Construct a finite element subspace V_h of $V = H_0^1(I)$ that consists of piecewise quadratic functions on a partition of $I = (0,1)$. How can the parameters (degrees of freedom) be chosen to describe such functions? Find the corresponding basis functions. Then define a finite element method for (1.1) using this space V_h and express the corresponding linear system of algebraic equations for a uniform partition.

1.28. Suppose that Γ is a circle with diameter L and that Γ_h is a polygonal approximation of Γ with vertices on Γ and maximal edge length equal to h. Show that the maximal distance from Γ to Γ_h is of the order $h^2/4L$ (cf. Sect. 1.5).

1.29. Let $\hat{K} = (0,1) \times (0,1)$ be the unit square with vertices $\hat{\mathbf{m}}_i$, $i = 1,2,3,4$, $P(\hat{K}) = Q_1(\hat{K})$, and $\Sigma_{\hat{K}}$ be the degrees of freedom corresponding to the values at $\hat{\mathbf{m}}_i$. If K is a convex quadrilateral, define an appropriate mapping $\mathbf{F} : \hat{K} \to K$ so that an isoparametric finite element $(K, P(K), \Sigma_K)$ can be defined in the form

$$P(K) = \{v : v(\mathbf{x}) = \hat{v}(\mathbf{F}^{-1}(\mathbf{x})),\ \mathbf{x} \in K,\ \hat{v} \in P(\hat{K})\} ,$$
$$\Sigma_K \text{ consists of function values at } \mathbf{m}_i = \mathbf{F}(\hat{\mathbf{m}}_i),\ i = 1,2,3,4 .$$

1.30. Show the stability result (1.82) for Crank-Nicholson's method (1.83) with $f = 0$. What can be shown if $f \neq 0$?

1.31. Consider the time-dependent problem

$$\begin{aligned} &\frac{\partial p}{\partial t} - \nabla \cdot (\mathbf{a} \nabla p) + \boldsymbol{\beta} \cdot \nabla p = f && \text{in } \Omega \times J , \\ &p = 0 && \text{on } \Gamma \times J , \\ &p(\cdot, 0) = p_0 && \text{in } \Omega , \end{aligned}$$

where $\mathbf{a}$ is a $d \times d$ matrix ($d = 2$ or 3), $\boldsymbol{\beta}$ is a constant vector, and f and p_0 are given functions. Extend the methods (1.74), (1.80), (1.83), and (1.85) to this problem and show a stability inequality similar to (1.82) for the method (1.80) in the case $f = 0$.

1.32. Show that for any triangle K in the $\mathbf{x}$-plane, K is affine-equivalent to the reference triangle $\hat{K}$ in the $\hat{\mathbf{x}}$-plane with vertices $(0,0)$, $(1,0)$, and $(0,1)$.

1.33. Prove that the number of operations to factor an $M \times M$ matrix with band width L is $ML^2/2$ (cf. Sect. 1.10.1.2).

2 Nonconforming Finite Elements

In the development of the finite element method for a second-order differential equation problem in the preceding chapter, piecewise polynomials in a finite element space V_h were required to be *continuous* throughout the whole domain Ω. Due to this continuity requirement, the resulting method is called the H^1*-conforming finite element method*. For the discretization of a fourth-order problem, functions in V_h and their first derivatives were required to be continuous on $\bar{\Omega}$. In this case, the finite element method is termed the H^2*-conforming method*. In this chapter, we introduce the *nonconforming finite element method* in which functions in a finite element space V_h for the discretization of a second-order problem are not required continuous on $\bar{\Omega}$; for a fourth-order problem, their derivatives (and even the functions themselves in some cases) are not required continuous on $\bar{\Omega}$.

Compared with the conforming finite element spaces introduced in Chap. 1, finite element spaces used in the nonconforming method (i.e., *nonconforming spaces*) employ fewer degrees of freedom, particularly for a fourth-order differential equation problem. The nonconforming method was initially introduced in the early 1960's (Adini-Clough, 1961). Since then, it has been widely used in computational mechanics and structural engineering; see Chaps. 7 and 8. In this chapter, we discuss its application to second- and fourth-order partial differential equation problems; see Sects. 2.1 and 2.2, respectively. In Sect. 2.3, we briefly present an application of this method to a nonlinear transient problem. Section 2.4 is devoted to theoretical considerations. The reader who is not interested in the theory may skip this section. Finally, in Sect. 2.5, bibliographical information is given.

2.1 Second-Order Problems

As in the preceding chapter, for the purpose of introduction, we consider a stationary problem for the unknown p:

$$\begin{aligned} -\nabla \cdot (\mathbf{a} \nabla p) &= f \qquad && \text{in } \Omega \,, \\ p &= g && \text{on } \Gamma \,, \end{aligned} \tag{2.1}$$

where $\Omega \subset \mathbb{R}^d$ ($d = 2$ or 3) is a bounded two- or three-dimensional domain with boundary Γ, the diffusion tensor $\mathbf{a}$ is assumed to be bounded, symmetric, and uniformly positive-definite in $\mathbf{x}$:

$$0 < a_* \leq |\boldsymbol{\eta}|^2 \sum_{i,j=1}^{d} a_{ij}(\mathbf{x})\eta_i\eta_j \leq a^* < \infty, \quad \mathbf{x} \in \Omega,\ \boldsymbol{\eta} \neq \mathbf{0} \in \mathbb{R}^d,$$

$\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_d)$, and f and g are given real-valued piecewise continuous bounded functions in Ω and Γ, respectively. A typical such problem is heat conduction where one seeks the temperature distribution p in an inhomogeneous plate Ω with conductivity tensor $\mathbf{a}$. This problem corresponds to the stationary case of problem (1.68) studied in Chap. 1.

We recall the scalar-product notation

$$(v, w) = \int_\Omega v(\mathbf{x})w(\mathbf{x})\ d\mathbf{x}\ ,$$

for real-valued functions $v, w \in L^2(\Omega)$, where (cf. Sect. 1.2)

$$L^2(\Omega) = \left\{ v :\ v \text{ is defined on } \Omega \text{ and } \int_\Omega v^2\ d\mathbf{x} < \infty \right\}\ .$$

We will also use the linear space (cf. Sect. 1.2)

$$H^1(\Omega) = \left\{ v \in L^2(\Omega) : \nabla v \in \left(L^2(\Omega)\right)^d \right\}, \quad d = 2 \text{ or } 3\ .$$

Furthermore, set

$$V = H_0^1(\Omega) = \left\{ v \in H^1(\Omega) : v|_\Gamma = 0 \right\}\ .$$

Multiplying the first equation of (2.1) by $v \in V$ and integrating over Ω, we see that

$$-\int_\Omega \nabla \cdot (\mathbf{a}\nabla p)v\ d\mathbf{x} = \int_\Omega fv\ d\mathbf{x}\ .$$

Applying Green's formula (1.19) to this equation and using the boundary condition in the definition of V, we have

$$\int_\Omega (\mathbf{a}\nabla p) \cdot \nabla v\ d\mathbf{x} = \int_\Omega fv\ d\mathbf{x} \qquad \forall v \in V\ ,$$

from which we derive the variational form

$$\text{Find } p \in H^1(\Omega) \text{ such that } a(p, v) = (f, v) \qquad \forall v \in V\ , \tag{2.2}$$

where $p|_\Gamma = g$ and

$$a(p, v) = (\mathbf{a}\nabla p, \nabla v)\ .$$

As happened in Chap. 1, the variational form (2.2) is equivalent to a minimization problem. In subsequent sections, we construct the finite element method for (2.1) that uses various *nonconforming elements*.

2.1.1 Nonconforming Finite Elements on Triangles

Let Ω be a polygonal domain in the plane, and let K_h be a triangulation of Ω into non-overlapping (open) triangles K:

$$\bar{\Omega} = \bigcup_{K \in K_h} \bar{K} \,,$$

such that no vertex of one triangle lies in the interior of an edge of another triangle, where $\bar{\Omega}$ and $\bar{K}$ represent the closure of Ω and K (i.e., $\bar{\Omega} = \Omega \cup \Gamma$ and $\bar{K} = K \cup \partial K$, where ∂K is the boundary of K), respectively. The mesh parameters h_K and h are defined as in the preceding chapter:

$$h_K = \operatorname{diam}(K) \text{ and } h = \max_{K \in K_h} h_K \,,$$

where $\operatorname{diam}(K)$ is the length of the longest edge of $\bar{K}$.

Now, we introduce the finite element spaces on triangles

$$\tilde{V}_h = \{v \in L^2(\Omega) : v|_K \text{ is linear, } K \in K_h; v \text{ is continuous at the midpoints of interior edges}\} \,,$$

and

$$V_h = \{v \in L^2(\Omega) : v|_K \text{ is linear, } K \in K_h; v \text{ is continuous at the midpoints of interior edges and is zero at the midpoints of edges on } \Gamma\} \,.$$

It can be shown that the degrees of freedom (i.e., the function values at the midpoints of edges) for V_h (cf. Fig. 2.1) are legitimate. Namely, a linear function v on each $K \in K_h$ is uniquely determined by them. In fact, by connecting the midpoints of the edges on each triangle $K \in K_h$, we obtain a smaller triangle (cf. Fig. 2.1) on which $v \in P_1$ vanishes at the vertices. Then an argument analogous to that in Example 1.6 applies (cf. Exercise 2.1).

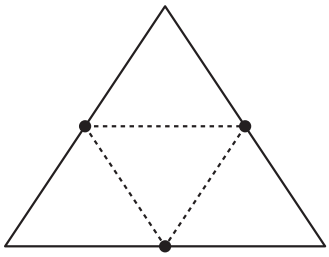

Fig. 2.1. The degrees of freedom for the Crouzeix-Raviart element

In Sect. 1.1.2, functions in the finite element space are required to be continuous across interelement boundaries. In contrast, the functions here are continuous only at the midpoints of interior edges, so $V_h \not\subset V$. In this case, V_h is referred to as a *nonconforming finite element space*. Because of the nonconformity, we introduce the mesh-dependent bilinear form $a_h(\cdot,\cdot) : V_h \times V_h \to \mathbb{R}$

$$a_h(v,w) = \sum_{K\in K_h} (\mathbf{a}\nabla v, \nabla w)_K , \qquad v,\ w \in V_h .$$

Then the *nonconforming finite element method* for (2.1) is formulated as follows:

$$\text{Find } p_h \in \tilde{V}_h \text{ such that } a_h(p_h, v) = (f, v) \qquad \forall v \in V_h , \tag{2.3}$$

where p_h equals g at the midpoints of the edges on the boundary Γ.

Existence and uniqueness of a solution to problem (2.3) can be easily checked. In fact, set $f = g = 0$. Then (2.3) becomes

$$a_h(p_h, v) = 0 \qquad \forall v \in V_h .$$

With $v = p_h$ in this equation, we see that p_h is a constant on each $K \in K_h$. Due to the continuity of p_h at interior midpoints, p_h is a constant on Ω. Consequently, the zero boundary condition implies $p_h = 0$. Uniqueness also yields existence since (2.3) is a finite-dimensional linear system.

Denote the midpoints (*nodes*) of edges in K_h by $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{\tilde{M}}$. The basis functions φ_i in $\tilde{V}_h$, $i = 1, 2, \ldots, \tilde{M}$, are defined as follows:

$$\varphi_i(\mathbf{x}_j) = \begin{cases} 1 & \text{if } i = j , \\ 0 & \text{if } i \neq j . \end{cases}$$

The *support* of φ_i, i.e., the set of $\mathbf{x}$ where $\varphi_i(\mathbf{x}) \neq 0$, consists of the triangles with the common node $\mathbf{x}_i$; see Fig. 2.2. In the present case, the support of each φ_i consists of at most two triangles.

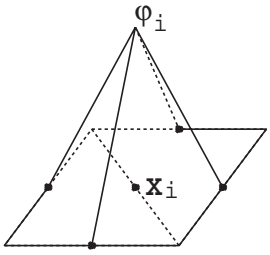

Fig. 2.2. A basis function in two dimensions

Let M ($M < \tilde{M}$) be the number of interior nodes in K_h. For notational convenience, the interior nodes are chosen to be the first M nodes. Then any function $v \in V_h$ has the unique representation

$$v(\mathbf{x}) = \sum_{i=1}^{M} v_i \varphi_i(\mathbf{x}), \qquad \mathbf{x} \in \Omega \ ,$$

where $v_i = v(\mathbf{x}_i)$. Also, the solution to (2.3) is given by

$$p_h = \sum_{i=1}^{M} p_i \varphi_i + \sum_{k=M+1}^{\tilde{M}} g_k \varphi_k \ , \tag{2.4}$$

where $g_k = g(\mathbf{x}_k)$.

For each j, we take $v = \varphi_j$ in (2.3) to see that

$$a_h(p_h, \varphi_j) = (f, \varphi_j), \qquad j = 1, 2, \ldots, M \ .$$

Substituting (2.4) into this equation, we have

$$\sum_{i=1}^{M} a_h(\varphi_i, \varphi_j) p_i = (f, \varphi_j) - \sum_{k=M+1}^{\tilde{M}} a_h(\varphi_k, \varphi_j) g_k, \qquad j = 1, 2, \ldots, M \ .$$

This is a linear system of M algebraic equations in the M unknowns p_1, $p_2, \ldots, p_M$. It can be written in matrix form

$$\mathbf{A}\mathbf{p} = \mathbf{f}, \tag{2.5}$$

where the matrix $\mathbf{A}$ and vectors $\mathbf{p}$ and $\mathbf{f}$ are given by

$$\mathbf{A} = (a_{ij}) \,, \quad \mathbf{p} = (p_j) \,, \quad \mathbf{f} = (f_j) \ ,$$

with, i, $j = 1, 2, \ldots, M$,

$$a_{ij} = a_h(\varphi_i, \varphi_j) \,, \quad f_j = (f, \varphi_j) - \sum_{k=M+1}^{\tilde{M}} a_h(\varphi_k, \varphi_j) g_k \ .$$

Symmetry of $\mathbf{A}$ can be seen from the definition of a_{ij}: $a_{ij} = a_{ji}$. Positive definiteness can be checked as in Chap. 1: With

$$\eta = \sum_{i=1}^{M} \eta_i \varphi_i \in V_h, \quad \boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_M) \ ,$$

we see that

$$\sum_{i,j=1}^{M} \eta_i a_{ij} \eta_j = \sum_{i,j=1}^{M} \eta_i a_h(\varphi_i, \varphi_j)\, \eta_j$$
$$= a_h\left(\sum_{i=1}^{M} \eta_i \varphi_i, \sum_{j=1}^{M} \eta_j \varphi_j\right) = a_h(\eta, \eta) \geq 0\ ,$$

so, as for (2.3), the equality holds only for $\eta \equiv 0$ since a constant function η must be zero because of the boundary condition in V_h. Particularly, positive definiteness implies that $\mathbf{A}$ is nonsingular. As a result, (2.5) has a unique solution. This is another way to show that (2.3) has a unique solution. System (2.5) can be solved using the linear system solution techniques discussed in Sect. 1.10.

We have considered a Dirichlet boundary value problem in this section. A Neumann or more general boundary value problem can be treated in a similar manner; see Sect. 1.1.3. An error analysis for the nonconforming finite element method (2.3) is delicate. A general theory for this method will be presented in Sect. 2.4. Here we just state the error estimate: If the solution p to (2.1) is in $H^2(\Omega)$, then

$$\|p - p_h\|_{L^2(\Omega)} + h\left(\sum_{K \in K_h} \|\nabla(p - p_h)\|^2_{(L^2(K))^2}\right)^{1/2}$$
$$\leq Ch^2 \|p\|_{H^2(\Omega)}\ , \tag{2.6}$$

where p_h is the solution of (2.3), C is a constant independent of h, and the triangulation K_h is assumed to be regular (see the definition of regularity on a triangulation in (1.52)). For the definition of the norms used in (2.6), refer to Sect. 1.2. The norm $\|\nabla(p - p_h)\|_{(L^2(K))^2}$ will be often denoted by $\|\nabla(p - p_h)\|_{\mathbf{L}^2(K)}$.

The nonconforming finite element under consideration is the linear Crouzeix-Raviart (1973) element, which is the simplest nonconforming element on triangles (also called the P_1-*nonconforming element*). For a quadratic nonconforming element on triangles, refer to Fortin-Soulie (1983). For general high-order nonconforming elements on triangles, see Arbogast-Chen (1995).

2.1.2 Nonconforming Finite Elements on Rectangles

We now consider the case where Ω is a rectangular domain and K_h is a partition of Ω into rectangles such that the horizontal and vertical edges of rectangles are parallel to the x_1- and x_2-coordinate axes, respectively, and adjacent elements completely share their common edge. Associated with K_h, we define the nonconforming finite element spaces on rectangles

$$\tilde{V}_h = \{v \in L^2(\Omega) : v|_K = a_K^1 + a_K^2 x_1 + a_K^3 x_2 + a_K^4 (x_1^2 - x_2^2) \,, \\ a_K^i \in \mathbb{R},\ K \in K_h;\ v \text{ is continuous at the} \\ \text{midpoints of interior edges}\} \,,$$

and

$$V_h = \{v \in L^2(\Omega) : v|_K = a_K^1 + a_K^2 x_1 + a_K^3 x_2 + a_K^4 (x_1^2 - x_2^2) \,, \\ a_K^i \in \mathbb{R},\ K \in K_h;\ v \text{ is continuous at the} \\ \text{midpoints of interior edges and is zero at} \\ \text{the midpoints of edges on } \Gamma\} \,.$$

The degrees of freedom for $\tilde{V}_h$ can be the function values at the midpoints of edges in K_h (cf. Fig. 2.3), and they are legitimate (cf. Exercise 2.5). With this definition, a linear system similar to (2.5) can be derived, and the error estimate (2.6) remains the same.

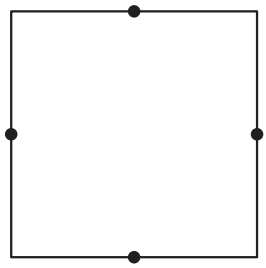

Fig. 2.3. The degrees of freedom for the rotated Q_1 element

This rectangular nonconforming element is termed the *rotated Q_1 element* (Rannacher-Turek, 1992; Chen, 1993B) because of the fact that $x_1^2 - x_2^2$ can be generated from $x_1 x_2$ by a rotation of 45°. The degrees of freedom for this element can be chosen in a different way (cf. Exercise 2.7):

$$\tilde{V}_h = \Big\{ v \in L^2(\Omega) : v|_K = a_K^1 + a_K^2 x_1 + a_K^3 x_2 + a_K^4 (x_1^2 - x_2^2), \\ a_K^i \in \mathbb{R},\ K \in K_h;\ \text{if } K_1 \text{ and } K_2 \text{ share an} \\ \text{edge } e, \text{ then } \int_e v|_{\partial K_1}\, d\ell = \int_e v|_{\partial K_2}\, d\ell \Big\} \,,$$

and

$$V_h = \Bigg\{ v \in L^2(\Omega) : v|_K = a_K^1 + a_K^2 x_1 + a_K^3 x_2 + a_K^4 (x_1^2 - x_2^2),$$
$$a_K^i \in \mathbb{R},\ K \in K_h;\ \text{if } K_1 \text{ and } K_2 \text{ share an}$$
$$\text{edge } e, \text{ then } \int_e v|_{\partial K_1}\, d\ell = \int_e v|_{\partial K_2}\, d\ell;$$
$$\int_{e \cap \Gamma} v|_e\, d\ell = 0 \Bigg\}\,.$$

The determination of the basis functions must be modified accordingly. Denote the set of edges in K_h by $e_1, e_2, \ldots, e_{\tilde{M}}$. The basis functions φ_i in $\tilde{V}_h$, $i = 1, 2, \ldots, \tilde{M}$, are determined by

$$\frac{1}{|e_j|} \int_{e_j} \varphi_i\, d\ell = \begin{cases} 1 & \text{if } i = j\,, \\ 0 & \text{if } i \neq j\,, \end{cases}$$

where $|e_j|$ represents the length of the edge e_j.

Although the degrees of freedom are different in these two spaces, they exhibit the same convergence rate as in (2.6). In terms of implementation, the linear system of equations from the second definition seems better conditioned (Chen-Oswald, 1998).

The rotated Q_1 nonconforming element is the simplest available on rectangles. The next simplest element is the *Wilson nonconforming element* (Wilson's rectangle), which is defined by

$$V_h = \Bigg\{ v : v|_K \in P_2(K),\ K \in K_h;\ v \text{ is determined}$$
$$\text{by its values at the vertices of } K \text{ and}$$
$$\text{the mean values of its second derivatives } \frac{\partial^2 v}{\partial x_1^2}$$
$$\text{and } \frac{\partial^2 v}{\partial x_2^2} \text{ over } K;\ v = 0 \text{ at the vertices on } \Gamma \Bigg\}\,,$$

where the mean value of $\dfrac{\partial^2 v}{\partial x_1^2}$ over K is defined by

$$\frac{1}{|K|} \int_K \frac{\partial^2 v}{\partial x_1^2}\, d\mathbf{x}\,,$$

with $|K|$ being the area of K. Using this space in (2.3), the error estimate (2.6) remains valid. That is, while more degrees of freedom are exploited in the Wilson element, the convergence rate is the same as in the rotated Q_1 element. For general high-order nonconforming elements on rectangles, refer to Arbogast-Chen (1995). We remark that although rectangular elements have been presented, an extension to general quadrilaterals can be made through change of variables from a reference rectangular element to quadrilaterals; refer to Sect. 1.5.

2.1.3 Nonconforming Finite Elements on Tetrahedra

Let K_h be a partition of $\Omega \subset \mathbb{R}^3$ into tetrahedra such that adjacent elements completely share their common face. In three dimensions, P_r is now the space of polynomials of degree r in three variables x_1, x_2, and x_3. The following space is the three-dimensional analogue of the Crouzeix-Raviart space on triangles:

$$V_h = \{v \in L^2(\Omega) : v|_K \in P_1(K),\ K \in K_h; v \text{ is continuous at the centroids of interior faces and is zero at the centroids of the faces on } \Gamma\} .$$

The degrees of freedom are the function values at the centroids of faces in K_h (cf. Fig. 2.4). With this definition in (2.3), the nonconforming finite element method and its analysis can be given in a similar fashion as in Sect. 2.1.1. Moreover, estimate (2.6) holds.

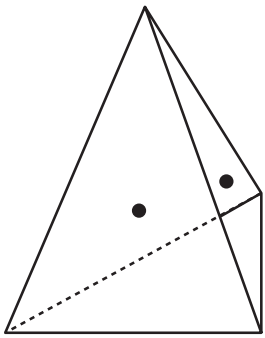

Fig. 2.4. The three-dimensional Crouzeix-Raviart element

2.1.4 Nonconforming Finite Elements on Parallelepipeds

Let $\Omega \subset \mathbb{R}^3$ be a rectangular domain and K_h be a partition of Ω into rectangular parallelepipeds such that their faces are parallel to the coordinate axes and adjacent elements completely share their common face. As in the two-dimensional case, the rotated Q_1 nonconforming element in three dimensions can be defined using two different sets of degrees of freedom. Namely, it can be defined either in terms of nodal values (cf. Fig. 2.5):

$$V_h = \{v \in L^2(\Omega) : v|_K = a_K^1 + a_K^2 x_1 + a_K^3 x_2 + a_K^4 x_3 + a_K^5 (x_1^2 - x_2^2) + a_K^6 (x_1^2 - x_3^2),\ a_K^i \in \mathbb{R},\ K \in K_h;\ v \text{ is continuous at the centroids of interior faces and is zero at the centroids of the faces on } \Gamma\} ,$$

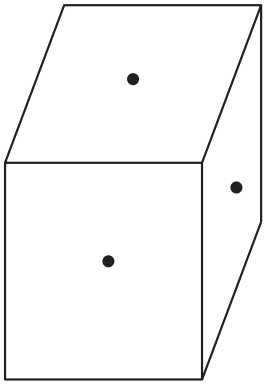

Fig. 2.5. The three-dimensional rotated Q_1 element

or in terms of the mean values over faces:

$$V_h = \Big\{ v \in L^2(\Omega) : v|_K = a_K^1 + a_K^2 x_1 + a_K^3 x_2 + a_K^4 x_3 + a_K^5 (x_1^2 - x_2^2) + a_K^6 (x_1^2 - x_3^2),\ a_K^i \in \mathbb{R},\ K \in K_h;\ \text{if } K_1 \text{ and } K_2 \text{ share a face } e,\ \text{then } \int_e v|_{\partial K_1}\, d\ell = \int_e v|_{\partial K_2}\, d\ell;\ \int_{e \cap \Gamma} v|_e\, d\ell = 0 \Big\}.$$

Again, they produce the same convergence rate as in (2.6), but the second definition seems to yield a better conditioned stiffness system (Chen-Oswald, 1998).

The three-dimensional analogue of the Wilson nonconforming element is called the *Wilson brick* (Ciarlet, 1978):

$$V_h = \Big\{ v : v|_K \in P_2(K) \oplus \operatorname{span}\{x_1 x_2 x_3\},\ K \in K_h;\ v \text{ is determined by its values at the vertices of } K \text{ and the mean values of its second derivatives } \frac{\partial^2 v}{\partial x_1^2},\ \frac{\partial^2 v}{\partial x_2^2},\ \text{and } \frac{\partial^2 v}{\partial x_3^2} \text{ on } K;\ v = 0 \text{ at the vertices on } \Gamma \Big\}.$$

Equivalently, the Wilson brick can be defined as

$$V_h = \Bigg\{ v : v|_K \in Q_1(K) \oplus \operatorname{span}\{x_1^2, x_2^2, x_3^2\},\ K \in K_h;$$
$$v \text{ is determined by its values at the vertices of } K$$
$$\text{and the mean values of its second derivatives } \frac{\partial^2 v}{\partial x_1^2},$$
$$\frac{\partial^2 v}{\partial x_2^2}, \text{ and } \frac{\partial^2 v}{\partial x_3^2} \text{ on } K;\ v = 0 \text{ at the vertices on } \Gamma \Bigg\},$$

where $Q_1(K)$ is the space of trilinear functions on K (cf. Sect. 1.4.3). The Wilson brick has more degrees of freedom than the three-dimensional rotated Q_1 element, but has the same convergence rate.

2.1.5 Nonconforming Finite Elements on Prisms

Let $\Omega \subset \mathbb{R}^3$ be a domain of the form $\Omega = G \times (l_1, l_2)$, where $G \subset \mathbb{R}^2$ and l_1 and l_2 are real numbers. Let K_h be a partition of Ω into prisms such that their bases are triangles in the (x_1, x_2)-plane with three vertical edges parallel to the x_3-axis and adjacent prisms completely share their common face. The nonconforming finite elements on prisms are analogues of those on rectangular parallelepipeds. Hence they can be defined using two different sets of degrees of freedom: nodal values (cf. Fig. 2.6) or mean values over faces. As an example, we present them in terms of the latter:

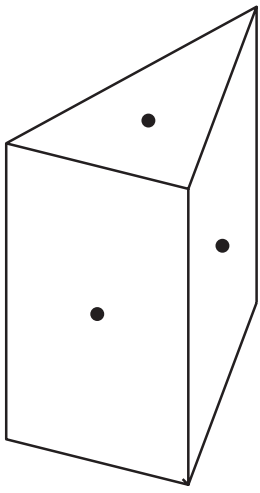

Fig. 2.6. The prismatic nonconforming element

$$V_h = \Big\{ v \in L^2(\Omega) : v|_K = a_K^1 + a_K^2 x_1 + a_K^3 x_2 + a_K^4 x_3 + a_K^5 (x_1^2 + x_2^2 - 2x_3^2),\ a_K^i \in \mathbb{R},\ K \in K_h;\ \text{if } K_1 \text{ and } K_2 \text{ share a face } e,\ \text{then } \int_e v|_{\partial K_1}\, d\ell = \int_e v|_{\partial K_2}\, d\ell;\ \int_{e \cap \Gamma} v|_e\, d\ell = 0 \Big\} .$$

Estimate (2.6) holds for this prismatic element.

In summary, we have presented the simplest nonconforming finite elements on triangles, rectangles, tetrahedra, rectangular parallelepipeds, and prisms. In practice, these elements are the most often used nonconforming elements. For corresponding higher-order nonconforming elements, the reader should refer to Arbogast-Chen (1995). An error analysis for these elements will be carried out in Sect. 2.4.

2.2 Fourth-Order Problems

We now extend the nonconforming finite element method to the fourth-order problem

$$\begin{aligned} \Delta^2 p &= f && \text{in } \Omega\ , \\ p = \frac{\partial p}{\partial \boldsymbol{\nu}} &= 0 && \text{on } \Gamma\ , \end{aligned} \tag{2.7}$$

where $\Omega \subset \mathbb{R}^2$, $\Delta^2 = \Delta\Delta$, and $\boldsymbol{\nu}$ is the outward unit normal to boundary Γ. This problem was briefly studied in Sect. 1.3.3 using the conforming finite element method. It models the displacement of a thin elastic plate under a transversal load of intensity f. The first boundary condition $p|_\Gamma = 0$ says that the displacement p is held fixed (at the zero height) at the boundary Γ, while the second condition $\partial p/\partial \boldsymbol{\nu}|_\Gamma = 0$ means that the rotation of the plate is also prescribed at Γ. These boundary conditions thus imply that the plate is *clamped*. In this section, we examine various nonconforming finite elements for (2.7).

We use the linear space

$$V = H_0^2(\Omega) = \left\{ v \in H^2(\Omega) : v = \frac{\partial v}{\partial \boldsymbol{\nu}} = 0 \text{ on } \Gamma \right\} ,$$

with the norm

$$\|v\|_V = \|v\|_{H^2(\Omega)} .$$

By Green's formula (1.19) and the boundary conditions in V, we see that

$$\begin{aligned}(\Delta^2 p, v) &= \left(\frac{\partial \Delta p}{\partial \boldsymbol{\nu}}, v\right)_\Gamma - (\nabla \Delta p, \nabla v) \\ &= -\left(\Delta p, \frac{\partial v}{\partial \boldsymbol{\nu}}\right)_\Gamma + (\Delta p, \Delta v) \\ &= (\Delta p, \Delta v)\,, \qquad v \in V\,,\end{aligned} \tag{2.8}$$

which is the bilinear form used Example 1.5 for the conforming finite element method. To have a well-posed problem for the nonconforming method for (2.7), this bilinear form needs to be modified. Toward that end, let $\mathbf{t} = (t_1, t_2)$ denote the unit tangential vector along Γ, oriented in the usual way. In addition to the outward normal derivative operator $\partial/\partial\boldsymbol{\nu}$, we also use the differential operators

$$\begin{aligned}\frac{\partial v}{\partial \mathbf{t}} &= \nabla v \cdot \mathbf{t}\,, \\ \frac{\partial^2 v}{\partial \boldsymbol{\nu} \partial \mathbf{t}} &= \sum_{i,j=1}^{2} \nu_i t_j \frac{\partial^2 v}{\partial \nu_i \partial t_j}\,, \\ \frac{\partial^2 v}{\partial \mathbf{t}^2} &= \sum_{i,j=1}^{2} t_i t_j \frac{\partial^2 v}{\partial t_i \partial t_j}\,.\end{aligned}$$

Then one can prove *Green's second formula*

$$\begin{aligned}&\int_\Omega \left(2\frac{\partial^2 v}{\partial x_1 \partial x_2}\frac{\partial^2 w}{\partial x_1 \partial x_2} - \frac{\partial^2 v}{\partial x_1^2}\frac{\partial^2 w}{\partial x_2^2} - \frac{\partial^2 v}{\partial x_2^2}\frac{\partial^2 w}{\partial x_1^2}\right) d\mathbf{x} \\ &\quad = \int_\Gamma \left(\frac{\partial^2 v}{\partial \boldsymbol{\nu} \partial \mathbf{t}}\frac{\partial w}{\partial \mathbf{t}} - \frac{\partial^2 v}{\partial \mathbf{t}^2}\frac{\partial w}{\partial \boldsymbol{\nu}}\right) d\ell, \quad v \in H^3(\Omega),\ w \in H^2(\Omega)\,.\end{aligned} \tag{2.9}$$

The proof of this formula is left as an exercise (Exercise 2.9).

Note that

$$\int_\Gamma \left(\frac{\partial^2 v}{\partial \boldsymbol{\nu} \partial \mathbf{t}}\frac{\partial w}{\partial \mathbf{t}} - \frac{\partial^2 v}{\partial \mathbf{t}^2}\frac{\partial w}{\partial \boldsymbol{\nu}}\right) d\ell = 0, \quad v \in H^3(\Omega),\ w \in H_0^2(\Omega)\,, \tag{2.10}$$

by the definition of $H_0^2(\Omega)$. Because of (2.10), we introduce a new bilinear form:

$$\begin{aligned}a(p, v) = (\Delta p, \Delta v) + (1 - \sigma)\Bigg[&2\left(\frac{\partial^2 p}{\partial x_1 \partial x_2}, \frac{\partial^2 v}{\partial x_1 \partial x_2}\right) \\ &- \left(\frac{\partial^2 p}{\partial x_1^2}, \frac{\partial^2 v}{\partial x_2^2}\right) - \left(\frac{\partial^2 p}{\partial x_2^2}, \frac{\partial^2 v}{\partial x_1^2}\right)\Bigg]\,,\end{aligned}$$

where σ is a physical constant known as *Poisson's ratio* (Ciarlet, 1978). In the model for the bending of plates, it satisfies $0 < \sigma \leq 1/2$. Using (2.8)

and (2.9), equation (2.7) can be written in the variational form (cf. Exercise 2.10):

$$\text{Find } p \in V \text{ such that } a(p, v) = (f, v) \qquad \forall v \in V. \tag{2.11}$$

We emphasize that the introduction of the constant σ in the bilinear form $a(\cdot, \cdot)$ is for the well-posedness of the discrete problem using the nonconforming method for (2.7) (cf. (2.15)).

To distinguish the bilinear form used for (1.57) and that for (2.7), we refer to a fourth-order problem associated with the bilinear form for the former as a *biharmonic problem*, while we refer to the same problem corresponding to the bilinear form for the latter as a *plate problem*. The latter concept comes from the observation that (2.11) corresponds to the variational formulation of the (*clamped*) *plate problem*, which concerns the equilibrium position of a plate of constant thickness under the action of a transverse force; see Chap. 7 for more details.

It follows from the definition of the bilinear form $a(\cdot, \cdot)$ that

$$a(v, v) = \sigma \|\Delta v\|^2_{L^2(\Omega)} + (1 - \sigma)|v|^2_{H^2(\Omega)}, \quad v \in H^2(\Omega). \tag{2.12}$$

Thus we see that $a(\cdot, \cdot)$ is V-elliptic (cf. Sect. 1.3.1). Also, it is easy to see that it is continuous in the norm $\|\cdot\|_{H^2(\Omega)}$. Therefore, (2.11) has a unique solution in V (Theorem 1.1 in Sect. 1.3.1). Moreover, it is known that $p \in H^3(\Omega) \cap H^2_0(\Omega)$ if Ω is a convex polygon or a smooth domain (Kondratev, 1967).

If we define the functional (*total potential energy of the plate*)

$$F(v) = \frac{1}{2}(\Delta v, \Delta v) + (1 - \sigma)\left(\left(\frac{\partial^2 v}{\partial x_1 \partial x_2}, \frac{\partial^2 v}{\partial x_1 \partial x_2}\right) - \left(\frac{\partial^2 v}{\partial x_1^2}, \frac{\partial^2 v}{\partial x_2^2}\right)\right) - (f, v)\,,$$

then (2.11) is equivalent to the minimization problem

$$\text{Find } p \in V \text{ such that } F(p) \le F(v) \qquad \forall v \in V. \tag{2.13}$$

The proof of this equivalence can be given in the same manner as for the equivalence between (1.2) and (1.3) (cf. Sect. 1.1.1).

2.2.1 The Morley Element

Let Ω be a polygonal domain in the plane, and let K_h be a triangulation of Ω into triangles, as in Sect. 2.1.1. The *Morley element* (Morley, 1968) on triangles is defined as follows: On each triangle $K \in K_h$, the shape function is in $P_2(K)$, and the degrees of freedom are the values of the function at the vertices of the triangle and the values of the first normal derivatives at the midpoints of the edges of the triangle (cf. Fig. 2.7; also refer to Exercise 2.12). Thus, for problem (2.7) the Morley finite element space is given by

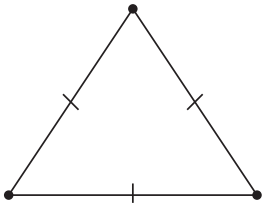

Fig. 2.7. The degrees of freedom for the Morley element

$$V_h = \{v \in L^2(\Omega) : \ v|_K \in P_2(K) \text{ for all } K \in K_h;\ v \text{ is continuous at the interior vertices and vanishes at the vertices on } \Gamma;\ \partial v/\partial \boldsymbol{\nu} \text{ is continuous at the midpoints of interior edges and vanishes at the midpoints of the edges on } \Gamma\}\ .$$

Note that functions in V_h are not continuous in Ω, and thus $V_h \not\subset V$. Compared with the H^2-conforming finite element studied in Example 1.5, the Moley element uses far fewer degrees of freedom.

As in the previous section, we introduce the mesh-dependent bilinear form $a_h(\cdot,\cdot) : V_h \times V_h \to \mathbb{R}$ by

$$a_h(v,w) = \sum_{K\in K_h} \left\{ (\Delta v, \Delta w)_K + (1-\sigma)\left[2\left(\frac{\partial^2 v}{\partial x_1 \partial x_2}, \frac{\partial^2 w}{\partial x_1 \partial x_2} \right)_K - \left(\frac{\partial^2 v}{\partial x_1^2}, \frac{\partial^2 w}{\partial x_2^2} \right)_K - \left(\frac{\partial^2 v}{\partial x_2^2}, \frac{\partial^2 w}{\partial x_1^2} \right)_K \right] \right\}, \quad v,\ w \in V_h\ .$$

Now, based on the Morley element, the nonconforming finite element method for (2.7) is formulated as follows:

$$\text{Find } p_h \in V_h \text{ such that } a_h(p_h, v) = (f, v) \qquad \forall v \in V_h. \tag{2.14}$$

Setting

$$\|v\|_h^2 = \sum_{K\in K_h} \left\{ \left(\frac{\partial^2 v}{\partial x_1^2}, \frac{\partial^2 v}{\partial x_1^2} \right)_K + 2\left(\frac{\partial^2 v}{\partial x_1 \partial x_2}, \frac{\partial^2 v}{\partial x_1 \partial x_2} \right)_K + \left(\frac{\partial^2 v}{\partial x_2^2}, \frac{\partial^2 v}{\partial x_2^2} \right)_K \right\}, \qquad v \in V_h\ ,$$

then we see that

$$a_h(v,v) \geq (1-\sigma)\|v\|_h^2, \qquad \forall v \in V_h. \tag{2.15}$$

Hence, if $\|\cdot\|_h$ is a norm on V_h, (2.14) has a unique solution $p_h \in V_h$. That $\|\cdot\|_h$ is indeed a norm on V_h can be seen as follows: Let $\|v\|_h = 0$, $v \in V_h$. Then the first partial derivatives of v are constant on each $K \in K_h$. Since

Again, functions in V_h are not continuous in Ω. For this element, inequality (2.15) still holds, and the results on existence, uniqueness, and convergence of a solution to (2.14) given in the previous subsection remain valid.

2.2.3 The Zienkiewicz Element

Another cubic nonconforming element on triangles is the *Zienkiewicz element* (Bazeley et al., 1965). This element has the same shape functions as the Fraeijs de Veubeke element, but utilizes different degrees of freedom. It is defined as follows: On each triangle $K \in K_h$, the shape function is in $P_3(K)$, and the degrees of freedom are the values of the function and of its first partial derivatives at the vertices of the triangle (cf. Fig. 2.9). The problem of determining a complete cubic function by these degrees of freedom does not have a unique solution, unless an additional independent relation is added, such as the following one:

$$6v(\mathbf{m}_0) - 2\sum_{i=1}^{3} v(\mathbf{m}_i) + \sum_{i=1}^{3} (\mathbf{m}_i - \mathbf{m}_0) \cdot \nabla v(\mathbf{m}_i) = 0 \,, \tag{2.18}$$

where $\mathbf{m}_0$ and $\mathbf{m}_i$ are the centroid and vertices of the triangle K, respectively. Now, the Zienkiewicz finite element space is given by

$$\begin{aligned} V_h = \{ v \in L^2(\Omega) : \; & v|_K \in P_3(K) \text{ for all } K \in K_h; \; v, \partial v/\partial x_1, \\ & \text{and } \partial v/\partial x_2 \text{ are continuous at the interior} \\ & \text{vertices and vanish at the vertices on } \Gamma; \\ & v \text{ on each } K \text{ satisfies (2.18)} \} \,. \end{aligned}$$

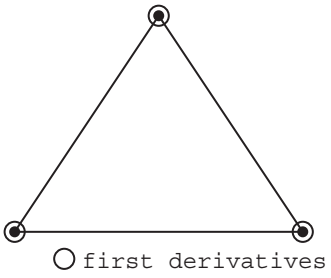

Fig. 2.9. The degrees of freedom for the Zienkiewicz element

Observe that for $v \in V_h$, the restriction $v|_e$ on each edge of $K \in K_h$ is a polynomial of degree at most three in a single variable. Because such a polynomial is uniquely determined by its values and the values of its first derivative at the end points of e, we see that v is continuous on $\bar{\Omega}$. That is,

the functions in V_h are continuous on $\bar{\Omega}$; however, they are not continuously differentiable.

Again, inequality (2.15) holds for the Zienkiewicz element. To see that $\|\cdot\|_h$ is a norm on V_h, let $\|v\|_h = 0$, $v \in V_h$. Then the first partial derivatives of v are constant on each $K \in K_h$. Since they are continuous at the interior vertices and equal zero at the vertices on Γ, v is constant on each $K \in K_h$. Also, because v is continuous at the interior vertices and equal zero at the vertices on Γ, we have $v = 0$. Therefore, $\|\cdot\|_h$ is a norm on V_h, and by (2.15), problem (2.14) has a unique solution when V_h is the Zienkiewicz finite element space.

To state a convergence rate for the Zienkiewicz element, we need an assumption on the triangulation K_h. We assume that all triangles in K_h have their edges parallel to three given directions (cf. Fig. 2.10). Then, if $p \in H^3(\Omega)$ and $p_h \in V_h$ is the solution of (2.14), the following error estimate holds (Lascaux-LeSaint, 1975):

$$\|p - p_h\|_{H^1(\Omega)} + h\|p - p_h\|_h \le Ch^2|p|_{H^3(\Omega)} \,. \tag{2.19}$$

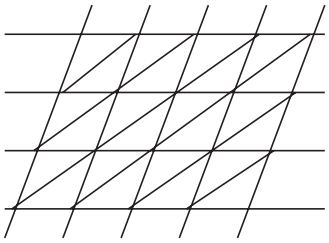

Fig. 2.10. Triangles with edges parallel to three given directions

2.2.4 The Adini Element

Let Ω be a rectangular domain, and K_h be a partition of Ω into rectangles, as in Sect. 2.1.2. We now introduce a nonconforming finite element on rectangles, the *Adini element* (Adini-Clough, 1961): On each rectangle $K \in K_h$, the shape function is in $P_3(K) \oplus \operatorname{span}\{x_1^3 x_2, x_1 x_2^3\}$, and the degrees of freedom are the values of the function and of its first partial derivatives at the vertices of the rectangle (cf. Fig. 2.11). Then the Adini finite element space is defined by

$$\begin{aligned} V_h = \{v \in L^2(\Omega) :\ & v|_K \in P_3(K) \oplus \operatorname{span}\{x_1^3 x_2, x_1 x_2^3\},\ K \in K_h; \\ & v,\ \partial v/\partial x_1, \text{ and } \partial v/\partial x_2 \text{ are continuous} \\ & \text{at the interior vertices and vanish at the} \\ & \text{vertices on } \Gamma\} \,. \end{aligned}$$

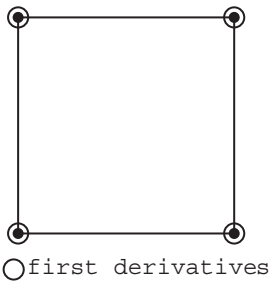

Fig. 2.11. The degrees of freedom for the Adini element

As seen as for the Zienkiewicz element, the functions in the Adini space V_h are continuous on $\bar{\Omega}$, but not continuously differentiable. Furthermore, the results on existence, uniqueness, and convergence of a solution to (2.14) for the Zienkiewicz element remain true here. In addition, if the partition K_h is uniform in both x_1- and x_2-directions, it holds that (Lascaux-LeSaint, 1975; Ciarlet, 1978)

$$\|p - p_h\|_h \leq Ch^2 |p|_{H^4(\Omega)} \ . \tag{2.20}$$

2.3 Nonlinear Problems

The nonconforming finite element method has been developed for stationary problems in the previous two sections. It can be also applied to the discretization of time-dependent transient problems as in Chap. 1 using the conforming method. As an example, we briefly present an application to a more general transient problem, the following nonlinear transient problem:

$$\begin{aligned} &c(p)\frac{\partial p}{\partial t} - \nabla \cdot \big(a(p)\nabla p\big) = f(p) && \text{in } \Omega \times J \ , \\ &p = 0 && \text{on } \Gamma \times J \ , \\ &p(\cdot, 0) = p_0 && \text{in } \Omega \ , \end{aligned} \tag{2.21}$$

where $c(p) = c(\mathbf{x}, t, p)$, $a(p) = a(\mathbf{x}, t, p)$, $f(p) = f(\mathbf{x}, t, p)$, $J = (0, T]$ $(T > 0)$, and $\Omega \subset \mathbb{R}^d$, $d = 2$ or 3. This problem has been studied for the conforming finite element method in the preceding chapter. We assume that (2.21) admits a unique solution. Furthermore, we assume that the coefficients $c(p)$, $a(p)$, and $f(p)$ are *globally Lipschitz continuous* in p; i.e., for some constants C_ξ, they satisfy

$$|\xi(p_1) - \xi(p_2)| \leq C_\xi |p_1 - p_2|, \qquad p_1, p_2 \in \mathbb{R},\ \xi = c,\ a,\ f \ . \tag{2.22}$$

Let $V = H_0^1(\Omega)$. Then problem (2.21) can be written in the variational form: Find $p : J \to V$ such that

$$\begin{aligned} &\left(c(p)\frac{\partial p}{\partial t}, v\right) + \big(a(p)\nabla p, \nabla v\big) = \big(f(p), v\big) \qquad \forall v \in V,\, t \in J\ , \\ &p(\mathbf{x}, 0) = p_0(\mathbf{x}) \qquad\qquad\qquad\qquad\quad \forall \mathbf{x} \in \Omega\ . \end{aligned} \tag{2.23}$$

Let V_h be one of the nonconforming finite element spaces introduced in Sect. 2.1. Then the nonconforming finite element method for (2.21) is: Find $p_h : J \to V_h$ such that

$$\begin{aligned} &\left(c(p_h)\frac{\partial p_h}{\partial t}, v\right) + \sum_{K \in K_h} \big(a(p_h)\nabla p_h, \nabla v\big)_K \\ &\qquad\qquad = \big(f(p_h), v\big) \qquad \forall v \in V_h\ , \\ &(p_h(\cdot, 0), v) = (p_0, v) \qquad \forall v \in V_h\ . \end{aligned} \tag{2.24}$$

As for (1.93) (also see (2.5)), after introduction of basis functions in V_h, system (2.24) can be restated in matrix form

$$\begin{aligned} &\mathbf{C}(\mathbf{p})\frac{d\mathbf{p}}{dt} + \mathbf{A}(\mathbf{p})\mathbf{p} = \mathbf{f}(\mathbf{p}), \qquad t \in J\ , \\ &\mathbf{B}\mathbf{p}(0) = \mathbf{p}_0\ . \end{aligned} \tag{2.25}$$

Under the assumption that the coefficient $c(p)$ is bounded below by a positive constant, this nonlinear system of ODEs locally has a unique solution. In fact, because of assumption (2.22) on c, a, and f, the solution $\mathbf{p}(t)$ exists for all t. The various solution approaches (e.g., linearization, implicit time approximation, and explicit time approximation) developed in Sect. 1.8 for the conforming finite element method can be applied to (2.25) in the same fashion.

2.4 Theoretical Considerations

In this section, we present a convergence theory for the nonconforming finite element method. First, we give an abstract formulation for this method. Then, as an example, we apply this formulation to second-order partial differential equations. For an application to fourth-order equations, the reader should refer to Lascaux-LeSaint (1975).

2.4.1 An Abstract Formulation

In general, if the finite element space used in the discretization of an H^m-elliptic problem ($m \geq 1$) is not a subspace of the Sobolev space $H^m(\Omega)$,

the finite element space is termed a *nonconforming finite element* space. In Sect. 2.1, $m = 1$, while, in Sect. 2.2, $m = 2$. In the nonconforming case, a convergence analysis is by no means obvious. The convergence theory in Sect. 1.9 for the conforming method must be extended. In particular, Céa's Lemma (Theorem 1.3) must be generalized. The notation in Sect. 1.9 will be utilized.

Suppose that V is a Hilbert space such that $H_0^r(\Omega) \subset V \subset H^r(\Omega)$ for some integer $r > 0$. Let $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ be a bilinear form and $L : V \to \mathbb{R}$ be a linear functional. Then we consider the abstract variational problem

$$\text{Find } p \in V \text{ such that } a(p,v) = L(v) \qquad \forall v \in V \ . \tag{2.26}$$

Under the assumptions of Theorem 1.1, this problem has unique solution.

Suppose that V_h is a finite dimensional space with norm $\|\cdot\|_h$. Let $a_h(\cdot,\cdot) : V_h \times V_h \to \mathbb{R}$ be a discrete bilinear form and $L_h : V_h \to \mathbb{R}$ be a linear functional. We also consider the discrete problem

$$\text{Find } p_h \in V_h \text{ such that } a_h(p_h,v) = L_h(v) \qquad \forall v \in V_h \ . \tag{2.27}$$

Assume that $a_h(\cdot,\cdot)$ is well defined on $V \times V$, it is continuous in the sense that

$$|a_h(v,w)| \le a^* \|v\|_h \|w\|_h \qquad \forall v,\ w \in V \cup V_h \ , \tag{2.28}$$

and it is V_h-elliptic:

$$|a_h(v,v)| \ge a_* \|v\|_h^2 \qquad \forall v \in V_h \ , \tag{2.29}$$

where a^* and a_* are positive constants independent of h. Under properties (2.28) and (2.29), problem (2.27) has a unique solution $p_h \in V_h$.

The next lemma, *Strang's Second Lemma*, is a generalization of Céa's Lemma from the conforming finite element method to the nonconforming method.

Lemma 2.1. *Let the bilinear form $a_h(\cdot,\cdot)$ satisfy* (2.28) *and* (2.29). *Then, for the respective solutions p and p_h of* (2.26) *and* (2.27), *there exists a constant $C > 0$, independent of h, such that*

$$\|p - p_h\|_h \le C \left(\inf_{v \in V_h} \|p - v\|_h + \sup_{w \in V_h \setminus \{0\}} \frac{|a_h(p,w) - L_h(w)|}{\|w\|_h} \right) . \tag{2.30}$$

Proof. For any $v \in V_h$, it follows from (2.27) and (2.29) that

$$\begin{aligned} a_* \|p_h - v\|_h^2 &\le a_h(p_h - v, p_h - v) \\ &= a_h(p - v, p_h - v) + [L_h(p_h - v) - a_h(p, p_h - v)] \ . \end{aligned}$$

Dividing this inequality by $\|p_h - v\|_h$, setting $w = p_h - v$, and using (2.28), we see that

$$a_*\|p_h - v\|_h \le a^*\|p - v\|_h + \frac{|L_h(w) - a_h(p,w)|}{\|w\|_h} ,$$

which, together with the triangle inequality

$$\|p - p_h\|_h \le \|p - v\|_h + \|v - p_h\|_h ,$$

implies (2.30). □

In (2.30), the first term in the right-hand side is referred to as the *approximation error*, and the second term is called the *consistency error*. The latter error stems from nonconformity.

The *duality argument* developed in Sect. 1.9.3 also needs to be generalized to the nonconforming method; the next lemma extends the Aubin-Nitsche technique in the conforming method.

Lemma 2.2. *Let H be a Hilbert space with the norm $\|\cdot\|_H$ and the scalar product $(\cdot,\cdot)$, and let $V_h \subset H$ and the imbedding $V \hookrightarrow H$ be continuous in the sense that*

$$\|v\|_H \le C\|v\|_V \qquad \forall v \in V .$$

Then, under (2.28)*, we have*

$$\begin{aligned}\|p - p_h\|_H \le \sup_{\psi\in H\setminus\{0\}} \frac{1}{\|\psi\|_H}\big\{ & a^*\|p - p_h\|_h\|\varphi_\psi - \varphi_h\|_h \\ & + |a_h(p - p_h, \varphi_\psi) - (p - p_h, \psi)| \\ & + |a_h(p, \varphi_\psi - \varphi_h) - L(\varphi_\psi - \varphi_h)|\big\} ,\end{aligned}$$

where, for given $\psi \in H$, $\varphi_\psi \in V$ is the solution of the problem

$$a(w, \varphi_\psi) = (w, \psi) \qquad \forall w \in V ,$$

and φ_h is the corresponding nonconforming finite element solution.

Proof. For any $\psi \in H$, it follows from the definition of p_h, φ_ψ, and φ_h that

$$\begin{aligned}(p - p_h, \psi) &= a_h(p, \varphi_\psi) - a_h(p_h, \varphi_h) \\ &= a_h(p - p_h, \varphi_\psi - \varphi_h) + a_h(p_h, \varphi_\psi - \varphi_h) + a_h(p - p_h, \varphi_h) \\ &= a_h(p - p_h, \varphi_\psi - \varphi_h) - [a_h(p - p_h, \varphi_\psi) - (p - p_h, \psi)] \\ &\quad - [a_h(p, \varphi_\psi - \varphi_h) - L(\varphi_\psi - \varphi_h)] ,\end{aligned}$$

which, together with (2.28) and the identity

$$\|p - p_h\|_H = \sup_{\psi\in H\setminus\{0\}} \frac{(p - p_h, \psi)}{\|\psi\|_H} ,$$

yields the desired result. □

2.4.2 Applications

We present an application of the theory to the second-order problem (2.1). For simplicity, we consider the model problem

$$\begin{aligned} -\Delta p &= f \qquad \text{in } \Omega \,, \\ p &= 0 \qquad \text{on } \Gamma \,. \end{aligned} \tag{2.31}$$

Define the spaces

$$V = H_0^1(\Omega), \quad H = L^2(\Omega) \,,$$

and the bilinear form

$$a(v,w) = (\nabla v, \nabla w), \qquad v,\ w \in V \,.$$

The norm and scalar product of H are given by

$$\|v\|_H = \|v\|_{L^2(\Omega)}, \qquad (v,w) = \int_\Omega v(\mathbf{x}) w(\mathbf{x})\ d\mathbf{x} \,.$$

As an example, we carry out in detail the analysis for the Crouzeix-Raviart nonconforming element. The finite element space V_h is given by

$$\begin{aligned} V_h = \{ v \in L^2(\Omega) : v|_K \text{ is linear, } K \in K_h; v \text{ is continuous} \\ \text{at the midpoints of interior edges and} \\ \text{is zero at the midpoints of edges on } \Gamma \} \,, \end{aligned}$$

and the bilinear form $a_h(\cdot,\cdot) : V_h \times V_h \to \mathbb{R}$ is

$$a_h(v,w) = \sum_{K \in K_h} (\nabla v, \nabla w)_K \,, \qquad v,\ w \in V_h \,.$$

The norm $\|\cdot\|_h$ on V_h is

$$\|v\|_h = a_h^{1/2}(v,v), \qquad v \in V_h \,.$$

The linear functionals $L : V \to \mathbb{R}$ and $L_h : V_h \to \mathbb{R}$ are given by

$$L(v) = (f,v), \quad v \in V, \quad L_h(v) = (f,v), \quad v \in V_h \,.$$

Let Ω be a convex polygon. Then the H^2-regularity result holds (cf. (1.121))

$$\|p\|_{H^2(\Omega)} \le C \|f\|_{L^2(\Omega)} \,. \tag{2.32}$$

We now apply Lemmas 2.1 and 2.2 to obtain error estimates for the Crouzeix-Raviart element. For this, we need the conforming finite element space

$$W_h = \{v : v \text{ is a continuous function on } \Omega, v \text{ is linear on each triangle } K \in K_h, \text{ and } v = 0 \text{ on } \Gamma\} .$$

For $v \in H^2(\Omega)$, let $\pi_h v$ be the interpolant of v in W_h; i.e., v and $\pi_h v$ have the same values at the vertices in K_h.

We say that the domain Ω satisfies a *cone condition* if for every point on the boundary Γ, there exists a cone with a positive angle such that this cone can be positioned in Ω with its vertex at that point (cf. Fig. 2.12).

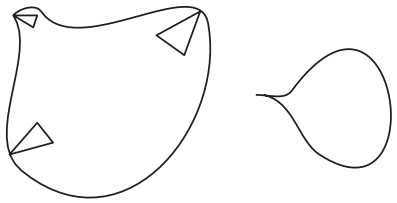

Fig. 2.12. The *left* domain satisfies the cone condition. The *right* domain does not satisfy the cone condition

We need the following *trace theorem* (Lions-Magenes, 1972):

Lemma 2.3. *Let the bounded domain Ω have a piecewise smooth boundary and satisfy the cone condition. Then there is a bounded linear mapping $\gamma : H^1(\Omega) \to L^2(\Gamma)$ such that*

$$\|\gamma(v)\|_{L^2(\Gamma)} \le C\|v\|_{H^1(\Omega)}, \qquad v \in H^1(\Omega) ,$$

and $\gamma(v) = v|_\Gamma$ for all $v \in C^1(\bar{\Omega})$.

It follows from this theorem that $\gamma(v)$ is the *trace* of v on boundary Γ, i.e., the restriction of v to Γ. The evaluation of a function in $H^1(\Omega)$ at a point on Γ does not always make sense. This theorem implies that the trace of v on Γ is at least in $L^2(\Gamma)$.

We also need the *Bramble-Hilbert Lemma* (Bramble-Hilbert, 1970), which is Lemma 1.4 in Chap. 1.

Lemma 2.4. *Let $\Omega \subset \mathbb{R}^2$ have a Lipschitz continuous boundary Γ, and $\mathcal{F} : H^r(\Omega) \to \mathcal{Y}$ be a bounded linear operator, where $r \ge 1$ and $\mathcal{Y}$ is a normed linear space, such that $P_{r-1}(\Omega)$ is a subset of the kernel of $\mathcal{F}$. Then there is a positive constant C such that*

$$\|\mathcal{F}(v)\|_{\mathcal{Y}} \le C(\Omega)\|\mathcal{F}\| \, \|v\|_{H^r(\Omega)}, \qquad v \in H^r(\Omega) ,$$

where $\|\mathcal{F}\|$ is the norm of the operator $\mathcal{F}$.

We are now in a position to apply Lemmas 2.1 and 2.2 to the Crouzeix-Raviart nonconforming element. Again, Ω is assumed to be a convex polygon. The following result also holds when it has a smooth boundary Γ.

Theorem 2.5. *Suppose that $\Omega \subset \mathbb{R}^2$ is convex. Then, if the solution p to (2.31) is in $H^2(\Omega)$ and p_h is the Crouzeix-Raviart nonconforming finite element solution, it holds that*

$$\|p - p_h\|_{L^2(\Omega)} + h\|p - p_h\|_h \leq Ch^2|p|_{H^2(\Omega)} ,$$

provided the triangulation K_h is shape-regular in the sense (1.52).

Proof. Since W_h is a subspace of V_h, the approximation error in (2.30) can be estimated (cf. Sect. 1.9):

$$\inf_{v \in V_h} \|p - v\|_h \leq Ch|p|_{H^2(\Omega)} . \tag{2.33}$$

It thus suffices to bound the consistency error in (2.30). Using Green's formula (1.19) and (2.31), we see that

$$\begin{aligned} a_h(p,w) - L_h(w) &= a_h(p,w) - (f,w) \\ &= \sum_{K \in K_h} (\nabla p, \nabla w)_K - (f,w) \\ &= \sum_{K \in K_h} \left[\left(\frac{\partial p}{\partial \boldsymbol{\nu}}, w \right)_{\partial K} - (\Delta p, w)_K \right] - (f,w) \\ &= \sum_{K \in K_h} \left(\frac{\partial p}{\partial \boldsymbol{\nu}}, w \right)_{\partial K} . \end{aligned} \tag{2.34}$$

For each $e \in \partial K$, we define the mean value of w on e

$$\bar{w}_e = \frac{1}{|e|} \int_e w|_K \, d\ell . \tag{2.35}$$

Note that each interior edge appears twice in the sum of (2.34), and $\bar{w}_e$ is a constant. Then it follows from (2.34) that

$$a_h(p,w) - L_h(w) = \sum_{K \in K_h} \sum_{e \in \partial K} \left(\frac{\partial p}{\partial \boldsymbol{\nu}}, w - \bar{w}_e \right)_e . \tag{2.36}$$

By (2.35), we have

$$\int_e (w - \bar{w}_e) \, d\ell = 0 .$$

Using the definition of $\pi_h p$, we see that

$$a_h(p,w) - L_h(w) = \sum_{K \in K_h} \sum_{e \in \partial K} \left(\frac{\partial p}{\partial \boldsymbol{\nu}} - \frac{\partial(\pi_h p)}{\partial \boldsymbol{\nu}}, w - \bar{w}_e \right)_e ,$$

so that, by Cauchy's inequality (1.10),

$$|a_h(p,w) - L_h(w)| \le \sum_{K\in K_h} \sum_{e\in\partial K} \|\nabla(p-\pi_h p)\|_{\mathbf{L}^2(e)} \|w - \bar{w}_e\|_{L^2(e)} \ . \quad (2.37)$$

We now estimate the right-hand side of (2.37). By Lemmas 2.3 and 2.4, for the reference triangle $\hat{K}$ with vertices $(1,0)$, $(0,1)$, and $(0,0)$, we get

$$\begin{aligned} \|\nabla(p-\pi_h p)\|_{\mathbf{L}^2(\partial\hat{K})} &\le C\|\nabla(p-\pi_h p)\|_{\mathbf{H}^1(\hat{K})} \\ &\le C\|p-\pi_h p\|_{H^2(\hat{K})} \\ &\le C|p|_{H^2(\hat{K})} \ . \end{aligned} \quad (2.38)$$

Applying a *scaling argument* (Dupont-Scott, 1980; also see Lemmas 1.5 and 1.6) to (2.38), we obtain, for $K \in K_h$,

$$\|\nabla(p-\pi_h p)\|_{\mathbf{L}^2(\partial K)} \le Ch^{1/2}|p|_{H^2(K)} \ . \quad (2.39)$$

In a similar way, we have, for $e \in \partial\hat{K}$,

$$\|w - \bar{w}_e\|_{L^2(e)} \le C|w|_{H^1(\hat{K})} \ ,$$

and, for $e \in \partial K$, $K \in K_h$,

$$\|w - \bar{w}_e\|_{L^2(e)} \le Ch^{1/2}|w|_{H^1(K)} \ . \quad (2.40)$$

We substitute (2.39) and (2.40) into (2.37) to have

$$\begin{aligned} &|a_h(p,w) - L_h(w)| \\ &\le\ Ch \sum_{K\in K_h} |p|_{H^2(K)}|w|_{H^1(K)} \\ &\le\ Ch\Big(\sum_{K\in K_h} |p|^2_{H^2(K)}\Big)^{1/2}\Big(\sum_{K\in K_h} |w|^2_{H^1(K)}\Big)^{1/2} \\ &=\ Ch|p|_{H^2(\Omega)}\|w\|_h \ , \end{aligned} \quad (2.41)$$

which, together with Lemma 2.1 and (2.33), yields

$$\|p - p_h\|_h \le Ch|p|_{H^2(\Omega)} \ . \quad (2.42)$$

We now apply Lemma 2.2 to estimate $p - p_h$ in the L^2-norm. It follows from (2.42) that

$$\|\varphi_\psi - \varphi_h\|_h \le Ch|\varphi_\psi|_{H^2(\Omega)} \ . \quad (2.43)$$

As for (2.41), we can show that

$$\begin{aligned} &|a_h\,(p-p_h,\varphi_\psi) - (p-p_h,\psi)| \le Ch|\varphi_\psi|_{H^2(\Omega)}\|p-p_h\|_h \ , \\ &|a_h\,(p,\varphi_\psi-\varphi_h) - L\,(\varphi_\psi-\varphi_h)| \le Ch|p|_{H^2(\Omega)}\|\varphi_\psi-\varphi_h\|_h \ . \end{aligned} \quad (2.44)$$

Finally, we combine Lemma 2.2, (2.32), (2.43), and (2.44) to obtain

$$\|p - p_h\|_{L^2(\Omega)} \le Ch^2|p|_{H^2(\Omega)} \ . \quad (2.45)$$

Inequalities (2.42) and (2.45) imply the desired result. □

2.5 Bibliographical Remarks

The nonconforming P_1 element in Sect. 2.1.1 was first introduced by Crouzeix-Raviart (1973). The rotated Q_1 element in Sect. 2.1.2 was developed independently by Rannacher-Turek (1992) and Chen (1993B). In Rannacher-Turek (1992), this element was applied to the numerical solution of a Stokes problem, and it was shown that this nonconforming element provides the simplest example of discretely divergence-free nonconforming elements on quadrilaterals. In Chen (1993B), this element was derived from a mixed finite element (i.e., the lowest-order Raviart-Thomas mixed element on rectangles; see Chap. 3). The extension of these two nonconforming elements to tetrahedra, rectangular parallelepipeds, and prisms in Sects. 2.1.3–2.1.5 was considered by Arbogast-Chen (1995).

For the Morley, Fraeijs de Veubeke, Zienkiewicz, and Adini elements for the fourth-order problems in Sect. 2.2, the reader refers to Morley (1968), Fraeijs de Veubeke (1974), Bazeley et al. (1965), and Adini-Clough (1961), respectively. For the stability and convergence analysis of these elements, the reader should see Lascaux-LeSaint (1975) and Ciarlet (1978). The theoretical development in Sect. 2.4 follows Braess (1997).

2.6 Exercises

2.1. For $v \in P_1(K)$, where K is a triangle, show that v is uniquely determined by its values at the midpoints of the three edges of K (refer to Sect. 2.1.1).

2.2. Use Fig. 2.13 to construct the linear basis functions φ_i at the three nodes $\mathbf{x}_i$ according to the definition in Sect. 2.1.1. Then use this result to determine the stiffness matrix $\mathbf{A}$ in (2.5) for problem (2.1), with

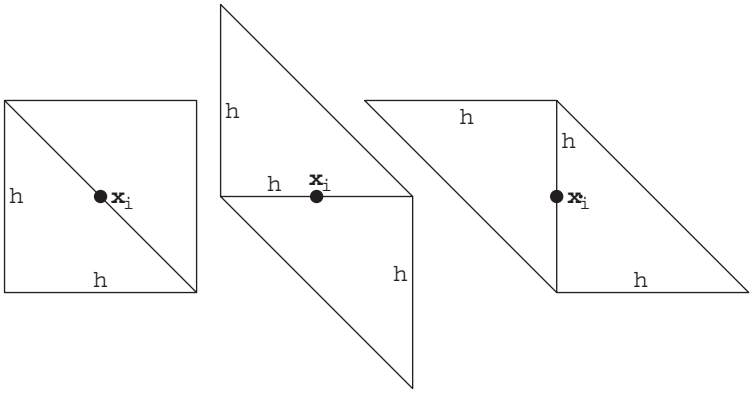

Fig. 2.13. The support of a basis function at node $\mathbf{x}_i$

$\mathbf{a} = \mathbf{I}$ (the identity tensor), $g = 0$, and a uniform partition of the unit square $(0,1) \times (0,1)$ as given in Fig. 1.7.

2.3. Write a code to solve problem (2.1) approximately using the nonconforming finite element method developed in Sect. 2.1.1. Use $\mathbf{a} = \mathbf{I}$ (the identity tensor), $f(x_1, x_2) = 8\pi^2 \sin(2\pi x_1) \sin(2\pi x_2)$, $g = 0$, and a uniform partition of $\Omega = (0,1) \times (0,1)$ as given in Fig. 1.7. Also, compute the errors

$$\|p - p_h\| = \left(\int_\Omega (p - p_h)^2 \, d\mathbf{x} \right)^{1/2} ,$$

with $h = 0.1$, 0.01, and 0.001, and compare them. Here p and p_h are the exact and approximate solutions, respectively, and h is the mesh size in the x_1- and x_2-directions. (If necessary, refer to Sect. 1.10.1.2 or Sect. 1.10.2 for a linear solver.)

2.4. Consider the problem

$$\begin{aligned} -\Delta p &= f && \text{in } \Omega , \\ p &= g_D && \text{on } \Gamma_D , \\ \frac{\partial p}{\partial \boldsymbol{\nu}} &= g_N && \text{on } \Gamma_N , \end{aligned}$$

where Ω is a bounded domain in the plane with boundary Γ, $\bar{\Gamma} = \bar{\Gamma}_D \cup \bar{\Gamma}_N$, $\Gamma_D \cap \Gamma_N = \emptyset$, and f, g_D, and g_N are given functions. Write down a variational formulation for this problem and formulate a nonconforming finite element method using the P_1-nonconforming element discussed in Sect. 2.1.1.

2.5. Let $v = a^1 + a^2 x_1 + a^3 x_2 + a^4 (x_1^2 - x_2^2)$, where $a^i \in \mathbb{R}$ $(i = 1, 2, 3, 4)$ and K is a rectangle. Prove that v is uniquely defined by its values at the midpoints of the four edges of K (refer to Sect. 2.1.2).

2.6. Let $v = a^1 + a^2 x_1 + a^3 x_2 + a^4 x_1 x_2$, where $a^i \in \mathbb{R}$ $(i = 1, 2, 3, 4)$ and K is a rectangle. Is v uniquely defined by its values at the midpoints of the four edges of K? Why?

2.7. Let $v = a^1 + a^2 x_1 + a^3 x_2 + a^4 (x_1^2 - x_2^2)$, where $a^i \in \mathbb{R}$ $(i = 1, 2, 3, 4)$ and K is a rectangle. Show that v is uniquely defined by its four integral values $\int_e v \, d\ell$, $e \in \partial K$ (refer to Sect. 2.1.2).

2.8. For Fig. 2.14, construct the rotated Q_1 basis functions associated with the edges e according to the definition in Sect. 2.1.2 (using the mean values over edges). Then use this result to determine the stiffness matrix $\mathbf{A}$ in (2.5) for problem (2.1) generated by the rotated Q_1 nonconforming method, with $\mathbf{a} = \mathbf{I}$ (the identity tensor), $g = 0$, and a uniform partition of the unit square $(0,1) \times (0,1)$ into rectangles with the mesh size h.

2.9. Prove Green's second formula (2.9).

2.10. Use (2.9) and (2.10) to show that problem (2.7) can be written in the variational form (2.11).

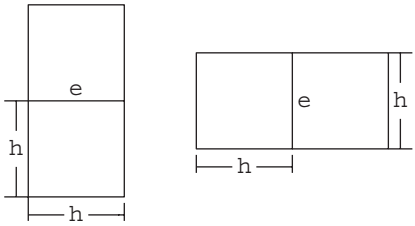

Fig. 2.14. The support of a basis function associated with edge e

2.11. Let $V = H_0^2(\Omega)$, and define the bilinear form $V \times V \to \mathbb{R}$:

$$a(p,v) = (\Delta p, \Delta v) + (1-\sigma)\left[2\left(\frac{\partial^2 p}{\partial x_1 \partial x_2}, \frac{\partial^2 v}{\partial x_1 \partial x_2}\right) - \left(\frac{\partial^2 p}{\partial x_1^2}, \frac{\partial^2 v}{\partial x_2^2}\right) - \left(\frac{\partial^2 p}{\partial x_2^2}, \frac{\partial^2 v}{\partial x_1^2}\right)\right].$$

Prove that $a(\cdot,\cdot)$ is V-elliptic (see Sect. 1.3.1 for the definition of V-ellipticness).

2.12. For $v \in P_2(K)$, where K is a triangle, show that v is uniquely determined by its values at the three vertices of K and the values of its normal derivatives at the midpoints of the three edges of K (refer to Sect. 2.2.1).

2.13. Prove the V_h-elliptic property (2.15).

2.14. Give a variational formulation for the problem

$$\begin{aligned} -\Delta p + p &= f && \text{in } \Omega\,, \\ \frac{\partial p}{\partial \boldsymbol{\nu}} &= g && \text{on } \Gamma\,, \end{aligned}$$

and formulate a nonconforming finite element method using the P_1-nonconforming space (cf. Sect. 2.1.1). Check if conditions (2.28) and (2.29) are satisfied.

2.15. Give a variational formulation for the problem

$$\begin{aligned} -\nabla \cdot (\mathbf{a}\nabla p) + cp &= f && \text{in } \Omega\,, \\ p &= g_D && \text{on } \Gamma_D\,, \\ \gamma p + \mathbf{a}\nabla p \cdot \boldsymbol{\nu} &= g_N && \text{on } \Gamma_N\,, \end{aligned}$$

where $\mathbf{a}$ is a 2×2 matrix, c, f, g_D, and g_N are given functions of $\mathbf{x}$, and γ is a constant. Formulate the nonconforming finite element method for this problem using the P_1-nonconforming space (cf. Sect. 2.1.1). Under what conditions on $\mathbf{a}$, c, and γ are the conditions (2.28) and (2.29) satisfied?

2.16. Let $\Omega \subset \mathbb{R}^2$ be a convex polygonal domain, and $\mathbf{a}$ and f be given as in the second-order problem (2.1). Formulate the nonconforming finite element method for the following problem using the P_1-nonconforming space (cf. Sect. 2.1.1):

$$\begin{aligned} -\nabla \cdot (\mathbf{a} \nabla p) + p &= f && \text{in } \Omega \,, \\ p &= 0 && \text{on } \Gamma \,. \end{aligned}$$

Under the assumption that $p \in H^2(\Omega)$, prove Theorem 2.5 for the resulting nonconforming method.

3 Mixed Finite Elements

In this chapter, we study the *mixed finite element method*, which generalizes the finite element methods discussed in the preceding chapters. This method was initially introduced by engineers in the 1960's (Fraeijs de Veubeke, 1965; Hellan, 1967; Hermann, 1967) for solving problems in solid continua. Since then, it has been applied to many areas such as solid and fluid mechanics; see Chaps. 7 and 8. In this chapter, we discuss its application to second-order partial differential equation problems. The reason for using the mixed method is, among others, that in some applications a vector variable (e.g., a fluid velocity) is the primary variable in which one is interested. Then the mixed method is developed to approximate both this variable and a scalar variable (e.g., a pressure) simultaneously and to give a high order approximation of both variables. Instead of a single finite element space used in the standard finite element method, the mixed finite element method employs two different spaces, which suggests the name *mixed*. These two spaces must satisfy an *inf-sup* condition for the mixed method to be stable. Raviart-Thomas (1977) introduced the first family of mixed finite element spaces for second-order elliptic problems in the two-dimensional case. Somewhat later, Néd'elec (1980) extended these spaces to three-dimensional problems. Motivated by these two papers, there are now many mixed finite element spaces available in the literature; see Brezzi et al. (1985, 1987A, 1987B) and Chen-Douglas (1989).

As an introduction, in Sect. 3.1, we first describe the mixed finite element method for a one-dimensional model problem. Then we generalize it to a two-dimensional model problem in Sect. 3.2. In Sect. 3.3, we consider the method for general boundary conditions. In Sect. 3.4, we present various mixed finite element spaces, and, in Sect. 3.5, we state the approximation properties of these spaces. In Sect. 3.6, we briefly present an application of the mixed method to a nonlinear transient problem. We also discuss solution techniques for solving the linear algebraic systems arising from this method in Sect. 3.7. Section 3.8 is devoted to theoretical considerations of this method. Finally, bibliographical information is given in Sect. 3.9. Here the mixed method is developed in a simple setting. The book by Brezzi-Fortin (1991) should be consulted for a thorough treatment of the subject.

3.1 A One-Dimensional Model Problem

As in Chap. 1, for the purpose of demonstration, we consider a stationary problem for p in one dimension

$$\begin{aligned} -\frac{d^2p}{dx^2} &= f(x), \quad 0 < x < 1 \,, \\ p(0) &= p(1) = 0 \,, \end{aligned} \tag{3.1}$$

where the function $f \in L^2(I)$ is given, with $I = (0,1)$ and

$$L^2(I) = \left\{ v : \ v \text{ is defined on } I \text{ and } \int_I v^2 \, dx < \infty \right\} \,.$$

We recall the scalar-product notation in $L^2(I)$:

$$(v,w) = \int_0^1 v(x)w(x) \, dx \,,$$

for real-valued functions $v, w \in L^2(I)$ (cf. Sect. 1.2). We will also use the linear space (cf. Sect. 1.2)

$$H^1(I) = \left\{ v \in L^2(I) : \frac{dv}{dx} \in L^2(I) \right\} \,.$$

Set

$$V = H^1(I), \quad W = L^2(I) \,.$$

Observe that the functions in W are not required to be continuous on the interval I.

After introducing the variable

$$u = -\frac{dp}{dx} \,, \tag{3.2}$$

equation (3.1) can be recast in the form

$$\frac{du}{dx} = f \,. \tag{3.3}$$

Multiplying (3.2) by any function $v \in V$ and integrating over I, we see that

$$(u,v) = -\left(\frac{dp}{dx}, v \right) \,.$$

Application of integration by parts to the right-hand side of this equation leads to

$$(u,v) = \left(p, \frac{dv}{dx} \right) \,,$$

where we use the boundary conditions $p(0) = p(1) = 0$ from (3.1). Also, we multiply (3.3) by any function $w \in W$ to give

$$\left(\frac{du}{dx}, w\right) = (f, w) \ .$$

Therefore, we see that the pair of functions u and p satisfies the system

$$\begin{aligned} (u, v) - \left(\frac{dv}{dx}, p\right) &= 0, \qquad v \in V \ , \\ \left(\frac{du}{dx}, w\right) &= (f, w), \qquad w \in W \ . \end{aligned} \tag{3.4}$$

This system is referred to as a *mixed variational* (or *weak*) *form* of (3.1). If the pair of functions u and p is a solution to (3.2) and (3.3), then this pair also satisfies (3.4). The converse also holds if p is sufficiently smooth (e.g., if $p \in H^2(I)$); see Exercise 3.1.

We introduce the functional $F : V \times W \to \mathbb{R}$ by

$$F(v, w) = \frac{1}{2}(v, v) - \left(\frac{dv}{dx}, w\right) + (f, w), \qquad v \in V, \ w \in W \ .$$

Then it can be checked (see the end of this section) that problem (3.4) is equivalent to the *saddle point problem*: Find $u \in V$ and $p \in W$ such that

$$F(u, w) \le F(u, p) \le F(v, p) \qquad \forall v \in V, \ w \in W \ . \tag{3.5}$$

For this reason, problem (3.4) is also referred to as a *saddle point problem.*

To construct the mixed finite element method for solving (3.1), for a positive integer M let $0 = x_1 < x_2 < \ldots < x_M = 1$ be a partition of I into a set of subintervals $I_{i-1} = (x_{i-1}, x_i)$, with length $h_i = x_i - x_{i-1}$, $i = 2, 3, \ldots, M$. Set $h = \max\{h_i, 2 \le i \le M\}$. Define the *mixed finite element spaces*

$$\begin{aligned} V_h = \{v : \ & v \text{ is a continuous function on } [0, 1] \\ & \text{and is linear on each subinterval } I_i\} \ , \\ W_h = \{w : \ & w \text{ is constant on each subinterval } I_i\} \ . \end{aligned}$$

Note that $V_h \subset V$ and $W_h \subset W$. Now, the *mixed finite element method* for (3.1) is defined as follows:

$$\text{Find } u_h \in V_h \text{ and } p_h \in W_h \text{ such that}$$

$$\begin{aligned} (u_h, v) - \left(\frac{dv}{dx}, p_h\right) &= 0, \qquad v \in V_h \ , \\ \left(\frac{du_h}{dx}, w\right) &= (f, w), \qquad w \in W_h \ . \end{aligned} \tag{3.6}$$

It can be shown that (3.6) has a unique solution. In fact, let $f = 0$; take $v = u_h$ and $w = p_h$ in (3.6) and add the resulting equations to give

$$(u_h, u_h) = 0 \ ,$$

so that $u_h = 0$. Consequently, it follows from (3.6) that

$$\left(\frac{dv}{dx}, p_h\right) = 0, \qquad v \in V_h \ .$$

Choose $v \in V_h$ such that $dv/dx = p_h$ (thanks to the definition of V_h and W_h) in this equation to see that $p_h = 0$. Hence the solution of (3.6) is unique. Uniqueness also yields existence since (3.6) is a finite-dimensional linear system.

In the same argument as for the equivalence between (3.4) and (3.5), problem (3.6) is equivalent to the saddle point problem: Find $u_h \in V_h$ and $p_h \in W_h$ such that

$$F(u_h, w) \le F(u_h, p_h) \le F(v, p_h) \qquad \forall v \in V_h, \ w \in W_h \ . \tag{3.7}$$

We introduce the *basis functions* $\varphi_i \in V_h$, $i = 1, 2, \ldots, M$,

$$\varphi_i(x_j) = \begin{cases} 1 & \text{if } i = j \ , \\ 0 & \text{if } i \ne j \ ; \end{cases}$$

see Fig. 1.3. Also, the basis functions $\psi_i \in W_h$, $i = 1, 2, \ldots, M-1$, are defined by

$$\psi_i(x) = \begin{cases} 1 & \text{if } x \in I_i \ , \\ 0 & \text{otherwise} \, . \end{cases}$$

These functions ψ_i are *characteristic* functions. Now, functions $v \in V_h$ and $w \in W_h$ have the unique representations

$$v(x) = \sum_{i=1}^{M} v_i \varphi_i(x), \quad w(x) = \sum_{i=1}^{M-1} w_i \psi_i(x), \quad 0 \le x \le 1 \ ,$$

where $v_i = v(x_i)$ and $w_i = w|_{I_i}$. Take v and w in (3.6) to be these basis functions to see that

$$\begin{aligned} &(u_h, \varphi_j) - \left(\frac{d\varphi_j}{dx}, p_h\right) = 0, \qquad && j = 1, 2, \ldots, M \ , \\ &\left(\frac{du_h}{dx}, \psi_j\right) = (f, \psi_j), && j = 1, 2, \ldots, M-1 \ . \end{aligned} \tag{3.8}$$

Set

$$u_h(x) = \sum_{i=1}^{M} u_i \varphi_i(x), \quad u_i = u_h(x_i) \ ,$$

and

$$p_h(x) = \sum_{k=1}^{M-1} p_k \psi_k(x), \quad p_k = p_h|_{I_k} \ .$$

Substitute these two expressions into (3.8) to give

$$\begin{aligned} &\sum_{i=1}^{M} (\varphi_i, \varphi_j)\, u_i - \sum_{k=1}^{M-1} \left(\frac{d\varphi_j}{dx}, \psi_k \right) p_k = 0, \quad j = 1, \ldots, M \ , \\ &\sum_{i=1}^{M} \left(\frac{d\varphi_i}{dx}, \psi_j \right) u_i = (f, \psi_j), \qquad\qquad j = 1, \ldots, M-1 \ . \end{aligned} \tag{3.9}$$

We introduce the matrices and vectors

$$\begin{aligned} &\mathbf{A} = (a_{ij})_{i,j=1,2,\ldots,M} \ , \quad \mathbf{B} = (b_{jk})_{j=1,2,\ldots,M,\ k=1,2,\ldots,M-1} \ , \\ &\mathbf{U} = (u_i)_{i=1,2,\ldots,M} \ , \qquad \mathbf{p} = (p_k)_{k=1,2,\ldots,M-1} \ , \quad \mathbf{f} = (f_j)_{j=1,2,\ldots,M-1} \ , \end{aligned}$$

where

$$a_{ij} = (\varphi_i, \varphi_j), \quad b_{jk} = -\left(\frac{d\varphi_j}{dx}, \psi_k \right), \quad f_j = (f, \psi_j) \ .$$

With these, system (3.9) can be written in matrix form

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{U} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ -\mathbf{f} \end{pmatrix} \ , \tag{3.10}$$

where $\mathbf{B}^T$ is the transpose of $\mathbf{B}$. Note that (3.10) is symmetric, but *indefinite*. It can be shown that the matrix $\mathbf{M}$ defined by

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{0} \end{pmatrix}$$

has both positive and negative eigenvalues (cf. Exercise 3.6).

We remark that the matrix $\mathbf{A}$ is symmetric and positive definite; refer to Sect. 1.1.2. It is also sparse. In the one-dimensional case, it is tridiagonal. In fact, it follows from the definition of the basis functions that

$$a_{ij} = (\varphi_i, \varphi_j) = 0 \quad \text{if } |i-j| \geq 2 \ ,$$

so that

$$a_{11} = \frac{h_2}{3}, \quad a_{MM} = \frac{h_M}{3} \ ,$$

and, for $i = 2, 3, \ldots, M-1$,

$$a_{i-1,i} = \frac{h_i}{6}, \quad a_{ii} = \frac{h_i}{3} + \frac{h_{i+1}}{3}, \quad a_{i,i+1} = \frac{h_{i+1}}{6} \ .$$

It can be also seen that

$$b_{jj} = 1, \quad b_{j+1,j} = -1, \quad j = 1, 2, \ldots, M-1 \, ;$$

all other entries of $\mathbf{B}$ are zero. That is, the $M \times (M-1)$ matrix $\mathbf{B}$ has the form

$$\mathbf{B} = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 & 0 \\ -1 & 1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & 0 \\ 0 & 0 & 0 & \ldots & -1 & 1 \\ 0 & 0 & 0 & \ldots & 0 & -1 \end{pmatrix} .$$

In the case where the partition is uniform, i.e., $h = h_i$, the matrix $\mathbf{A}$ is given by

$$\mathbf{A} = \frac{h}{6} \begin{pmatrix} 2 & 1 & 0 & \ldots & 0 & 0 \\ 1 & 4 & 1 & \ldots & 0 & 0 \\ 0 & 1 & 4 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 4 & 1 \\ 0 & 0 & 0 & \ldots & 1 & 2 \end{pmatrix} .$$

We end this section with two remarks. First, even for the one-dimensional problem, an error analysis for the mixed finite element method (3.6) is delicate. General error estimates for this method will be described in Sect. 3.8. We just point out that an error estimate of the following type can be obtained for (3.6):

$$\|p - p_h\| + \|u - u_h\| \leq Ch \, , \tag{3.11}$$

where u, p and u_h, p_h are the respective solutions of (3.4) and (3.6), C depends on the size of the second derivative of p, and we recall the norm (cf. Sect. 1.2)

$$\|v\| = \|v\|_{L^2(I)} = \left(\int_0^1 v^2 \, dx \right)^{1/2} .$$

When u is sufficiently smooth (e.g., $u \in H^2(I)$), we can show the error estimate

$$\|u - u_h\| \leq Ch^2 \, . \tag{3.12}$$

Error bounds (3.11) and (3.12) are optimal for p and u.

Second, we establish the equivalence between (3.4) and (3.5). Suppose that (u, p) is a solution of (3.4). With any $v \in V$, set $\tau = v - u \in V$; we see that

$$\begin{aligned} F(v,p) = F(u+\tau,p) &= \frac{1}{2}\,(u+\tau,u+\tau) - \left(\frac{du}{dx}+\frac{d\tau}{dx},p\right) + (f,p) \\ &= \frac{1}{2}\,(u,u) - \left(\frac{du}{dx},p\right) + (f,p) + (u,\tau) - \left(\frac{d\tau}{dx},p\right) + \frac{1}{2}\,(\tau,\tau) \\ &= F(u,p) + \frac{1}{2}\,(\tau,\tau) \geq F(u,p)\ . \end{aligned}$$

Thus the second inequality in (3.5) is shown. The first inequality can be proven similarly.

Conversely, let (u,p) be a solution of (3.5). Then, for any $v \in V$ and any $\epsilon \in \mathbb{R}$, it follows from the second inequality in (3.5) that

$$F(u,p) \leq F(u+\epsilon v,p)\ .$$

We define the function

$$\begin{aligned} G(\epsilon) &= F(u+\epsilon v,p) \\ &= \frac{1}{2}\,(u,u) + \epsilon(u,v) + \frac{\epsilon^2}{2}(v,v) - \left(\frac{du}{dx},p\right) - \epsilon\left(\frac{dv}{dx},p\right) + (f,p)\ . \end{aligned}$$

Then we see that G has a minimum at $\epsilon = 0$, so $\dfrac{dG}{d\epsilon}(0) = 0$. Note that

$$\frac{dG}{d\epsilon}(0) = (u,v) - \left(\frac{dv}{dx},p\right)\ ,$$

so (u,p) satisfies the first equation in (3.4). The second equation in (3.4) follows from the first inequality in (3.5) in the same fashion.

3.2 A Two-Dimensional Model Problem

We now extend the mixed finite element method in the previous section to a stationary problem in two dimensions

$$\begin{aligned} -\Delta p &= f \qquad \text{in } \Omega\ , \\ p &= 0 \qquad \text{on } \Gamma\ , \end{aligned} \tag{3.13}$$

where Ω is a bounded domain in the plane with boundary Γ and $f \in L^2(\Omega)$ is a given function. We recall that

$$L^2(\Omega) = \left\{ v :\ v \text{ is defined on } \Omega \text{ and } \int_\Omega v^2\ d\mathbf{x} < \infty \right\}\ .$$

We also use the space

$$\mathbf{H}(\mathrm{div},\Omega) = \left\{ \mathbf{v} = (v_1,v_2) \in (L^2(\Omega))^2 : \nabla\cdot\mathbf{v} \in L^2(\Omega) \right\}\ ,$$

where

$$\nabla \cdot \mathbf{v} = \frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} \ .$$

It can be checked (cf. Exercise 3.7) that for any decomposition of Ω into subdomains such that the interiors of these subdomains are pairwise disjoint, the space $\mathbf{H}(\mathrm{div}, \Omega)$ consists of those functions whose normal components are continuous across the interior edges in this decomposition. Define

$$\mathbf{V} = \mathbf{H}(\mathrm{div}, \Omega), \quad W = L^2(\Omega) \ .$$

Set

$$\mathbf{u} = -\nabla p \ . \tag{3.14}$$

Equation (3.13) is then given by

$$\nabla \cdot \mathbf{u} = f \ . \tag{3.15}$$

Multiply (3.14) by $\mathbf{v} \in \mathbf{V}$ and integrate over Ω to see that

$$(\mathbf{u}, \mathbf{v}) = -(\mathbf{v}, \nabla p) \ .$$

Applying Green's formula (1.19) to the right-hand side of this equation, we have

$$(\mathbf{u}, \mathbf{v}) = (\nabla \cdot \mathbf{v}, p) \ ,$$

where we use the boundary condition in (3.13). Also, multiplying (3.15) by any $w \in W$, we get

$$(\nabla \cdot \mathbf{u}, w) = (f, w) \ .$$

Thus we have the system for $\mathbf{u}$ and p

$$\begin{aligned} (\mathbf{u}, \mathbf{v}) - (\nabla \cdot \mathbf{v}, p) &= 0, \qquad && \mathbf{v} \in \mathbf{V} \ , \\ (\nabla \cdot \mathbf{u}, w) &= (f, w), && w \in W \ . \end{aligned} \tag{3.16}$$

This is the mixed variational form of (3.13). If $\mathbf{u}$ and p satisfy (3.14) and (3.15), they also satisfy (3.16). The converse also holds if p is sufficiently smooth (e.g., if $p \in H^2(\Omega)$); see Exercise 3.8. In a similar fashion as for (3.4) and (3.5), (3.16) can be written as a saddle point problem.

For a polygonal domain Ω, let K_h be a partition of Ω into non-overlapping (open) triangles such that no vertex of one triangle lies in the interior of an edge of another triangle. Define the mixed finite element spaces

$$\begin{aligned} \mathbf{V}_h &= \{\mathbf{v} \in \mathbf{V} : \ \mathbf{v}\big|_K = (b_K x_1 + a_K, b_K x_2 + c_K) \ , \\ &\qquad\qquad a_K, b_K, c_K \in \mathbb{R}, \ K \in K_h\} \ , \\ W_h &= \{w : \ w \text{ is constant on each triangle in } K_h\} \ . \end{aligned}$$

As noted, $\mathbf{V}_h$ can be also described as follows:

$$\mathbf{V}_h = \{\mathbf{v} : \mathbf{v}\big|_K = (b_K x_1 + a_K, b_K x_2 + c_K),\ K \in K_h\ , \\ a_K, b_K, c_K \in \mathbb{R},\ \text{and the normal components of } \mathbf{v} \\ \text{are continuous across the interior edges in } K_h\}\ .$$

Note that $\mathbf{V}_h \subset \mathbf{V}$ and $W_h \subset W$. The mixed finite element method for (3.13) is defined as follows:

$$\begin{aligned} &\text{Find } \mathbf{u}_h \in \mathbf{V}_h \text{ and } p_h \in W_h \text{ such that} \\ &(\mathbf{u}_h, \mathbf{v}) - (\nabla \cdot \mathbf{v}, p_h) = 0, \qquad \mathbf{v} \in \mathbf{V}_h\ , \\ &(\nabla \cdot \mathbf{u}_h, w) = (f, w), \qquad w \in W_h\ . \end{aligned} \tag{3.17}$$

It can be proven as for (3.6) that (3.17) has a unique solution.

Let $\{\mathbf{x}_i\}$ be the set of the midpoints of edges in K_h, $i = 1, 2, \ldots, M$. With each point $\mathbf{x}_i$, we associate a unit normal vector $\boldsymbol{\nu}_i$. For $\mathbf{x}_i \in \Gamma$, $\boldsymbol{\nu}_i$ is just the outward unit normal to Γ; for $\mathbf{x}_i \in e = \bar{K}_1 \cap \bar{K}_2$, $K_1, K_2 \in K_h$, let $\boldsymbol{\nu}_i$ be any unit vector orthogonal to e (cf. Fig. 3.1). We now define the basis functions of $\mathbf{V}_h$, $i = 1, 2, \ldots, M$, by

$$(\boldsymbol{\varphi}_i \cdot \boldsymbol{\nu}_i)(\mathbf{x}_j) = \begin{cases} 1 & \text{if } i = j\ , \\ 0 & \text{if } i \neq j\ . \end{cases}$$

Any function $\mathbf{v} \in \mathbf{V}_h$ has the unique representation

$$\mathbf{v}(\mathbf{x}) = \sum_{i=1}^{M} v_i \boldsymbol{\varphi}_i(\mathbf{x}), \qquad \mathbf{x} \in \Omega\ ,$$

where $v_i = (\mathbf{v} \cdot \boldsymbol{\nu}_i)(\mathbf{x}_i)$. Also, the basis functions $\psi_i \in W_h$, $i = 1, 2, \ldots, N$, can be defined as in the previous section; i.e.,

$$\psi_i(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in K_i\ , \\ 0 & \text{otherwise}\ , \end{cases}$$

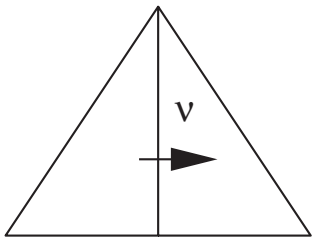

Fig. 3.1. An illustration of the unit normal $\boldsymbol{\nu}$

where $\bar{\Omega} = \bigcup_{i=1}^{N} \bar{K}_i$ and N is the number of triangles in K_h. Any function $w \in W_h$ also has the representation

$$w(\mathbf{x}) = \sum_{i=1}^{N} w_i \psi_i(\mathbf{x}), \qquad \mathbf{x} \in \Omega,\ w_i = w\big|_{K_i} .$$

In the same manner as in the previous section, system (3.17) can be recast in matrix form (cf. Exercise 3.9):

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{U} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ -\mathbf{f} \end{pmatrix} , \tag{3.18}$$

where

$$\mathbf{A} = (a_{ij})_{i,j=1,2,\ldots,M} , \quad \mathbf{B} = (b_{jk})_{j=1,2,\ldots,M,\ k=1,2,\ldots,N} ,$$
$$\mathbf{U} = (u_i)_{i=1,2,\ldots,M} , \quad \mathbf{p} = (p_k)_{k=1,2,\ldots,N} , \quad \mathbf{f} = (f_j)_{j=1,2,\ldots,N} ,$$

with

$$a_{ij} = \big(\boldsymbol{\varphi}_i, \boldsymbol{\varphi}_j\big), \quad b_{jk} = -\big(\nabla \cdot \boldsymbol{\varphi}_j, \psi_k\big), \quad f_j = (f, \psi_j) .$$

Again, the matrix $\mathbf{M}$ defined by

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{0} \end{pmatrix}$$

has both positive and negative eigenvalues. The matrix $\mathbf{A}$ is symmetric, positive definite, and sparse. In fact, it has at most five nonzero entries in each row in the present case (cf. Exercise 3.9). The matrix $\mathbf{B}$ is also sparse, with two nonzero entries in each row in the present case.

Let $\mathbf{u},\ p$ and $\mathbf{u}_h,\ p_h$ be the respective solutions of (3.16) and (3.17). Then the following error estimate holds (cf. Sect. 3.8):

$$\|p - p_h\| + \|\mathbf{u} - \mathbf{u}_h\| \le Ch , \tag{3.19}$$

where C depends on the size of the second partial derivatives of p. This estimate is optimal for the present pair of mixed finite element spaces.

3.3 Extension to Boundary Conditions of Other Types

3.3.1 A Neumann Boundary Condition

In the previous section, we considered the Dirichlet boundary condition in (3.13). We now extend the mixed finite element method to the stationary problem with the *homogeneous Neumann boundary condition*:

$$\begin{aligned} -\Delta p &= f \qquad && \text{in } \Omega \ , \\ \frac{\partial p}{\partial \boldsymbol{\nu}} &= 0 \qquad && \text{on } \Gamma \ , \end{aligned} \tag{3.20}$$

where $\partial p/\partial \boldsymbol{\nu}$ is the derivative of p normal to boundary Γ.

Application of Green's formula (1.19) to (3.20) yields

$$\int_\Omega f \ d\mathbf{x} = 0 \ .$$

This is a *compatibility condition*. In this case, p is unique up to an additive constant.

We define the spaces

$$\begin{aligned} \mathbf{V} &= \{\mathbf{v} = (v_1, v_2) \in \mathbf{H}(\mathrm{div}, \Omega) : \ \mathbf{v} \cdot \boldsymbol{\nu} = 0 \text{ on } \Gamma\} \ , \\ W &= \left\{ w \in L^2(\Omega) : \ \int_\Omega w \ d\mathbf{x} = 0 \right\} \ . \end{aligned}$$

With the choice of these two spaces, the mixed variational form of (3.20) is

$$\begin{aligned} &\text{Find } \mathbf{u} \in \mathbf{V} \text{ and } p \in W \text{ such that} \\ &(\mathbf{u}, \mathbf{v}) - (\nabla \cdot \mathbf{v}, p) = 0, \qquad && \mathbf{v} \in \mathbf{V} \ , \\ &(\nabla \cdot \mathbf{u}, w) = (f, w), && w \in W \ . \end{aligned} \tag{3.21}$$

Note that the Neumann boundary condition becomes the *essential* condition that must be incorporated into the definition of the space $\mathbf{V}$. In contrast, the Dirichlet boundary condition is the essential condition in the finite element method (cf. Sect. 1.1.3).

Let K_h be a partition of Ω into non-overlapping triangles, as defined in the previous section. We define the mixed finite element spaces

$$\begin{aligned} \mathbf{V}_h = \{&\mathbf{v} \in \mathbf{H}(\mathrm{div}, \Omega) : \ \mathbf{v}\big|_K = (b_K x_1 + a_K, b_K x_2 + c_K) \ , \\ &a_K, b_K, c_K \in \mathbb{R}, \ K \in K_h, \text{ and } \mathbf{v} \cdot \boldsymbol{\nu} = 0 \text{ on } \Gamma\} \ , \\ W_h = \Big\{ &w : \ w\big|_K \text{ is constant on each } K \in K_h \text{ and } \int_\Omega w \ d\mathbf{x} = 0 \Big\} \ . \end{aligned}$$

Again, $\mathbf{V}_h \subset \mathbf{V}$ and $W_h \subset W$. The mixed finite element method for (3.20) reads as follows:

$$\begin{aligned} &\text{Find } \mathbf{u}_h \in \mathbf{V}_h \text{ and } p_h \in W_h \text{ such that} \\ &(\mathbf{u}_h, \mathbf{v}) - (\nabla \cdot \mathbf{v}, p_h) = 0, \qquad && \mathbf{v} \in \mathbf{V}_h \ , \\ &(\nabla \cdot \mathbf{u}_h, w) = (f, w), && w \in W_h \ . \end{aligned} \tag{3.22}$$

This system can be rewritten in matrix form as in (3.18), and the error estimate (3.19) also holds.

3.3.2 A Boundary Condition of Third Type

We now consider a boundary condition of *third type*:

$$\begin{aligned} -\Delta p &= f && \text{in } \Omega\ , \\ \gamma p + \frac{\partial p}{\partial \boldsymbol{\nu}} &= g && \text{on } \Gamma\ , \end{aligned} \tag{3.23}$$

where γ is a strictly positive function on Γ and g is a given function. This boundary condition is also called a *mixed*, *Robin*, or *Dankwerts* boundary condition.

With the linear spaces $\mathbf{V}$ and W defined as in Sect. 3.2, the mixed variational form of (3.23) is

$$\begin{aligned} &\text{Find } \mathbf{u} \in \mathbf{V} \text{ and } p \in W \text{ such that} \\ &(\mathbf{u}, \mathbf{v}) + \int_\Gamma \gamma^{-1} \mathbf{u} \cdot \boldsymbol{\nu}\ \mathbf{v} \cdot \boldsymbol{\nu}\ d\ell - (\nabla \cdot \mathbf{v}, p) \\ &\qquad = -\int_\Gamma \gamma^{-1} g \mathbf{v} \cdot \boldsymbol{\nu}\ d\ell, \qquad \mathbf{v} \in \mathbf{V}\ , \\ &(\nabla \cdot \mathbf{u}, w) = (f, w), \qquad w \in W\ . \end{aligned} \tag{3.24}$$

Similarly, with the mixed finite element spaces in Sect. 3.2, the mixed finite element method for (3.23) is given by

$$\begin{aligned} &\text{Find } \mathbf{u}_h \in \mathbf{V}_h \text{ and } p_h \in W_h \text{ such that} \\ &(\mathbf{u}_h, \mathbf{v}) + \int_\Gamma \gamma^{-1} \mathbf{u}_h \cdot \boldsymbol{\nu}\ \mathbf{v} \cdot \boldsymbol{\nu}\ d\ell - (\nabla \cdot \mathbf{v}, p_h) \\ &\qquad = -\int_\Gamma \gamma^{-1} g \mathbf{v} \cdot \boldsymbol{\nu}\ d\ell, \qquad \mathbf{v} \in \mathbf{V}_h\ , \\ &(\nabla \cdot \mathbf{u}_h, w) = (f, w), \qquad w \in W_h\ . \end{aligned} \tag{3.25}$$

The matrix form and error estimate of (3.25) can be obtained in the same fashion as in Sect. 3.2 (cf. Exercise 3.13).

The two-dimensional *Poisson equation* has been considered so far in this chapter. The mixed finite element method for more general partial differential equations will be treated in later sections and chapters.

3.4 Mixed Finite Element Spaces

We consider a stationary problem for the unknown p:

$$\begin{aligned} -\nabla \cdot (\mathbf{a} \nabla p) &= f && \text{in } \Omega\ , \\ p &= g && \text{on } \Gamma\ , \end{aligned} \tag{3.26}$$

where $\Omega \subset \mathbb{R}^d$ ($d = 2$ or 3) is a bounded two- or three-dimensional domain with boundary Γ, the diffusion tensor $\mathbf{a}$ is assumed to be bounded, symmetric, and uniformly positive-definite in $\mathbf{x}$:

$$0 < a_* \leq |\boldsymbol{\eta}|^2 \sum_{i,j=1}^{d} a_{ij}(\mathbf{x})\eta_i\eta_j \leq a^* < \infty,\ \mathbf{x} \in \Omega,\ \boldsymbol{\eta} \neq \mathbf{0} \in \mathbb{R}^d\ , \qquad (3.27)$$

$\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_d)$, and f and g are given real-valued piecewise continuous bounded functions in Ω and Γ, respectively. This problem was considered in the preceding two chapters. To write (3.26) in a mixed variational form, the Sobolev spaces introduced in Sect. 3.2 will be exploited. The norms of the two spaces $W = L^2(\Omega)$ and $\mathbf{V} = \mathbf{H}(\mathrm{div}, \Omega)$ are, respectively, defined by

$$\|w\| \equiv \|w\|_{L^2(\Omega)} = \left(\int_\Omega w^2\ d\mathbf{x} \right)^{1/2}, \qquad w \in W\ ,$$

and

$$\|\mathbf{v}\|_{\mathbf{V}} \equiv \|\mathbf{v}\|_{\mathbf{H}(\mathrm{div},\Omega)} = \left\{ \|\mathbf{v}\|^2 + \|\nabla \cdot \mathbf{v}\|^2 \right\}^{1/2}, \qquad \mathbf{v} \in \mathbf{V}\ .$$

The definition of $\mathbf{H}(\mathrm{div}, \Omega)$ for $\Omega \subset \mathbb{R}^3$ is similar to that in Sect. 3.2; in this case, recall that

$$\nabla \cdot \mathbf{v} = \frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} + \frac{\partial v_3}{\partial x_3}, \qquad \mathbf{v} = (v_1, v_2, v_3)\ .$$

Let

$$\mathbf{u} = -\mathbf{a}\nabla p\ . \qquad (3.28)$$

In the same way as in the derivation of (3.16), problem (3.26) is written in the mixed variational form:

$$\begin{aligned} &\text{Find } \mathbf{u} \in \mathbf{V} \text{ and } p \in W \text{ such that} \\ &(\mathbf{a}^{-1}\mathbf{u}, \mathbf{v}) - (\nabla \cdot \mathbf{v}, p) = -\int_\Gamma g\mathbf{v} \cdot \boldsymbol{\nu}\ d\ell, \quad \mathbf{v} \in \mathbf{V}\ , \\ &(\nabla \cdot \mathbf{u}, w) = (f, w), \qquad\qquad\qquad\qquad\quad w \in W\ . \end{aligned} \qquad (3.29)$$

There is a constant $C_1 > 0$ such that the *inf-sup* condition holds (cf. Sect. 3.8)

$$\sup_{\mathbf{0} \neq \mathbf{v} \in \mathbf{V}} \frac{|(\nabla \cdot \mathbf{v}, w)|}{\|\mathbf{v}\|_{\mathbf{V}}} \geq C_1 \|w\| \qquad \forall w \in W\ . \qquad (3.30)$$

Because of (3.27) and (3.30), problem (3.29) has a unique solution $\mathbf{u} \in \mathbf{V}$ and $p \in W$, with $\mathbf{u}$ given by (3.28).

Let $\mathbf{V}_h \subset \mathbf{V}$ and $W_h \subset W$ be certain finite dimensional subspaces. The discrete version of (3.29) is

$$\begin{aligned} &\text{Find } \mathbf{u}_h \in \mathbf{V}_h \text{ and } p_h \in W_h \text{ such that} \\ &(\mathbf{a}^{-1}\mathbf{u}_h, \mathbf{v}) - (\nabla \cdot \mathbf{v}, p_h) = -\int_\Gamma g\mathbf{v} \cdot \boldsymbol{\nu}\ d\ell, \quad \mathbf{v} \in \mathbf{V}_h\ , \\ &(\nabla \cdot \mathbf{u}_h, w) = (f, w), \qquad\qquad\qquad\qquad\quad w \in W_h\ . \end{aligned} \qquad (3.31)$$

For this problem to have a unique solution, it is natural to impose a discrete *inf-sup* condition similar to (3.30):

$$\sup_{\mathbf{0}\neq\mathbf{v}\in\mathbf{V}_h} \frac{|(\nabla\cdot\mathbf{v},w)|}{\|\mathbf{v}\|_{\mathbf{V}}} \geq C_2\|w\| \qquad \forall w\in W_h\ , \tag{3.32}$$

where $C_2>0$ is a constant independent of h.

In the previous two sections, we have considered the mixed finite element spaces $\mathbf{V}_h$ and W_h over triangles. These spaces are the lowest-order triangular spaces introduced by Raviart-Thomas (1977), and they satisfy condition (3.32) (cf. Sect. 3.8). In this section, we describe other mixed finite element spaces that satisfy this *stability condition*. These spaces are RTN (Raviart-Thomas, 1977; Néd'elec, 1980), BDM (Brezzi et al., 1985), BDDF (Brezzi et al., 1987A), BDFM (Brezzi et al., 1987B), and CD (Chen-Douglas, 1989) spaces.

Condition (3.32) is also called the *Babuška-Brezzi condition* or sometimes the *Ladyshenskaja-Babuška-Brezzi condition*.

For simplicity, let Ω be a polygonal domain in this section. For a curved domain, the definition of the mixed finite element spaces under consideration is the same, but the degrees of freedom for $\mathbf{V}_h$ need to be modified (Brezzi-Fortin, 1991).

3.4.1 Mixed Finite Element Spaces on Triangles

For $\Omega\subset\mathbb{R}^2$, let K_h be a partition of Ω into triangles such that adjacent elements completely share their common edge. For a triangle $K\in K_h$, let

$$P_r(K)=\{v:\ v \text{ is a polynomial of degree at most } r \text{ on } K\}\ ,$$

where $r\geq 0$ is an integer. Mixed finite element spaces $\mathbf{V}_h\times W_h$ are defined locally on each element $K\in K_h$, so let $\mathbf{V}_h(K)=\mathbf{V}_h|_K$ (the restriction of $\mathbf{V}_h$ to K) and $W_h(K)=W_h|_K$.

3.4.1.1 The RT Spaces on Triangles

As noted, these spaces are the first mixed finite element spaces introduced by Raviart-Thomas (1977). They are defined for each $r\geq 0$ by

$$\mathbf{V}_h(K)=\big(P_r(K)\big)^2\oplus\big((x_1,x_2)P_r(K)\big),\quad W_h(K)=P_r(K)\ ,$$

where the notation $\oplus$ indicates a direct sum and $(x_1,x_2)P_r(K)=\big(x_1P_r(K),$ $x_2P_r(K)\big)$. The case $r=0$ was used in the previous sections. In this case, we observe that $\mathbf{V}_h(K)$ has the form

$$\mathbf{V}_h(K)=\{v: v=(a_K+b_Kx_1,c_K+b_Kx_2),\quad a_K,\ b_K,\ c_K\in\mathbb{R}\}\ ,$$

and its dimension is three. As discussed in Sect. 3.2, as parameters, or *the degrees of freedom*, to describe the functions in $\mathbf{V}_h$, we use the values of normal components of the functions at the midpoints of edges in K_h (cf. Fig. 3.2). Also, in the case $r = 0$, the degrees of freedom for W_h can be the averages of functions over K, as in Sect. 3.2.

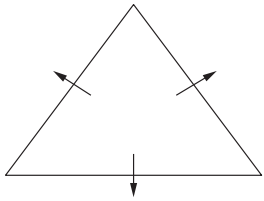

Fig. 3.2. The triangular RT

In general, for $r \geq 0$ the dimensions of $\mathbf{V}_h(K)$ and $W_h(K)$ are

$$\dim\big(\mathbf{V}_h(K)\big) = (r+1)(r+3), \quad \dim\big(W_h(K)\big) = \frac{(r+1)(r+2)}{2} .$$

They will be useful in the definition of certain projection operators into $\mathbf{V}_h$ (cf. Sect. 3.8.4). The degrees of freedom for the space $\mathbf{V}_h(K)$, with $r \geq 0$, are given by (Raviart-Thomas, 1977)

$$\begin{aligned} &(\mathbf{v}\cdot\boldsymbol{\nu}, w)_e \qquad \forall w \in P_r(e),\ e \in \partial K\ , \\ &(\mathbf{v}, \mathbf{w})_K \qquad \forall \mathbf{w} \in \big(P_{r-1}(K)\big)^2\ , \end{aligned}$$

where $\boldsymbol{\nu}$ is the outward unit normal to $e \in \partial K$. We claim that this is a legitimate choice; i.e., a function in $\mathbf{V}_h$ is uniquely determined by these degrees of freedom. Because $\dim\big(\mathbf{V}_h(K)\big)$ equals the number of degrees of freedom (i.e., $(r+1)(r+3)$), it suffices to show that if these degrees of freedom vanish

$$\begin{aligned} &(\mathbf{v}\cdot\boldsymbol{\nu}, w)_e = 0 \qquad \forall w \in P_r(e),\ e \in \partial K\ , \\ &(\mathbf{v}, \mathbf{w})_K = 0 \qquad \forall \mathbf{w} \in \big(P_{r-1}(K)\big)^2\ , \end{aligned} \tag{3.33}$$

then $\mathbf{v} \equiv \mathbf{0}$ on K. Since $\mathbf{v}\cdot\boldsymbol{\nu} \in P_r(e)$ on each $e \in \partial K$, the first equation of (3.33) yields $\mathbf{v}\cdot\boldsymbol{\nu} = 0$ on e. For $w \in P_r(K)$, Green's formula (1.19) implies

$$\int_K \nabla\cdot\mathbf{v}\ w\ d\mathbf{x} = -\int_K \mathbf{v}\cdot\nabla w\ d\mathbf{x} + \int_{\partial K} \mathbf{v}\cdot\boldsymbol{\nu} w\ d\ell\ .$$

The second term in the right-hand side of this equation vanishes. Since $\nabla w \in \big(P_{r-1}(K)\big)^2$, the first term also vanishes by the second equation of (3.33). Consequently,

$$\int_K \nabla\cdot\mathbf{v}\ w\ d\mathbf{x} = 0\ ,$$

which implies $\nabla \cdot \mathbf{v} = 0$ because $\nabla \cdot \mathbf{v} \in P_r(K)$. Therefore, there exists a *stream function* $\phi \in H^1(K)$ such that $\mathbf{v} = \mathbf{curl}\,\phi$, where $\mathbf{curl}\,\phi = (-\partial\phi/\partial x_2, \partial\phi/\partial x_1)$.

Set $\mathbf{v} = \mathbf{q} + (x_1, x_2)v$, where $\mathbf{q} \in (P_r(K))^2$ and $v \in P_r(K)$. Without loss of generality, let v be a homogeneous polynomial of degree r. Then

$$\nabla \cdot ((x_1, x_2)v) = 2v + x_1 \frac{\partial v}{\partial x_1} + x_2 \frac{\partial v}{\partial x_2} = (r+2)v \in P_r(K)\ .$$

Since $\nabla \cdot \mathbf{v} = 0$, $\nabla \cdot \mathbf{q} + (r+2)v = 0$. As a result, $v = 0$ because $\nabla \cdot \mathbf{q} \in P_{r-1}(K)$. Hence $\mathbf{v} = \mathbf{q} \in (P_r(K))^2$, and the stream function $\phi \in P_{r+1}(K)$.

Due to the fact that $\mathbf{v} \cdot \boldsymbol{\nu} = 0$ on ∂K, $\partial\phi/\partial\mathbf{t} = 0$ on ∂K, where $\mathbf{t}$ is a tangential direction. Thus ϕ is a constant on ∂K. We may assume that $\phi = 0$ on ∂K since ϕ is unique up to a constant. This implies $\phi = \lambda_1\lambda_2\lambda_3\psi$, where $\psi \in P_{r-2}(K)$ and λ_1, λ_2, and λ_3 are the barycentric coordinates of the triangle K (cf. Sect. 1.4).

Finally, for $\mathbf{w} \in (P_{r-1}(K))^2$, it follows from the second equation of (3.33) and integration by parts that

$$\begin{aligned} 0 &= \int_K \mathbf{v} \cdot \mathbf{w}\ d\mathbf{x} = \int_K \mathbf{curl}\,\phi \cdot \mathbf{w}\ d\mathbf{x} \\ &= \int_K \phi\ \mathrm{curl}\,\mathbf{w}\ d\mathbf{x} = \int_K \lambda_1\lambda_2\lambda_3\psi\ \mathrm{curl}\,\mathbf{w}\ d\mathbf{x}\ , \end{aligned}$$

where $\mathrm{curl}\,\mathbf{w} = \frac{\partial w_1}{\partial x_2} - \frac{\partial w_2}{\partial x_1}$, with $\mathbf{w} = (w_1, w_2)$. Letting $\psi = \mathrm{curl}\,\mathbf{w}$ in this equation yields

$$\int_K \lambda_1\lambda_2\lambda_3\psi^2\ d\mathbf{x} = 0\ .$$

Since $\lambda_1\lambda_2\lambda_3 \geq 0$ on K, $\psi = 0$; thus $\phi = 0$ and $\mathbf{v} = \mathbf{0}$. This proves *unisolvance*.

3.4.1.2 The BDM Spaces on Triangles

The BDM spaces on triangles (Brezzi et al., 1985) lie between corresponding RT spaces, are of smaller dimension than the RT space of the same index, and provide asymptotic error estimates for the vector variable of the same order as the corresponding RT space. They are defined for each $r \geq 1$ by

$$\mathbf{V}_h(K) = \big(P_r(K)\big)^2, \quad W_h(K) = P_{r-1}(K)\ .$$

The simplest BDM spaces on triangles are those with $r = 1$. In this case, $\mathbf{V}_h(K)$ is

$$\begin{aligned} \mathbf{V}_h(K) = \{v : v = (a_K^1 + a_K^2 x_1 + a_K^3 x_2, a_K^4 + a_K^5 x_1 + a_K^6 x_2)\ , \\ a_K^i \in \mathbb{R},\ i = 1, 2, \ldots, 6\}\ , \end{aligned}$$

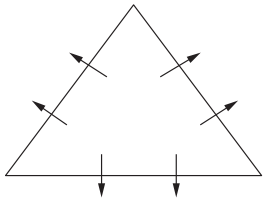

Fig. 3.3. The triangular BDM

so its dimension is six. The degrees of freedom for $\mathbf{V}_h$ are the values of normal components of functions at the two quadratic Gauss points on each edge in K_h (cf. Fig. 3.3). The space $W_h(K)$ with $r = 1$ consists of constants.

In general, for $r \geq 1$ the dimensions of $\mathbf{V}_h(K)$ and $W_h(K)$ are

$$\dim\big(\mathbf{V}_h(K)\big) = (r+1)(r+2), \quad \dim\big(W_h(K)\big) = \frac{r(r+1)}{2} \ .$$

Let

$$B_{r+1}(K) = \{v \in P_{r+1}(K) : v|_{\partial K} = 0\} = \lambda_1\lambda_2\lambda_3 P_{r-2}(K) \ .$$

The degrees of freedom for $\mathbf{V}_h(K)$ are (Brezzi et al., 1985)

$$\begin{aligned} &(\mathbf{v}\cdot\boldsymbol{\nu}, w)_e && \forall w \in P_r(e),\ e \in \partial K \ , \\ &(\mathbf{v}, \nabla w)_K && \forall w \in P_{r-1}(K) \ , \\ &(\mathbf{v},\ \mathbf{curl}\ w)_K && \forall w \in B_{r+1}(K) \ . \end{aligned}$$

They are a legitimate choice; see Exercise 3.14.

3.4.2 Mixed Finite Element Spaces on Rectangles

We now consider the case where Ω is a rectangular domain and K_h is a partition of Ω into rectangles such that the horizontal and vertical edges of rectangles are parallel to the x_1- and x_2-coordinate axes, respectively, and adjacent elements completely share their common edge. Define

$$Q_{l,r}(K) = \left\{ v : \ v(\mathbf{x}) = \sum_{i=0}^{l}\sum_{j=0}^{r} v_{ij} x_1^i x_2^j,\ \mathbf{x} = (x_1, x_2) \in K,\ v_{ij} \in \mathbb{R} \right\};$$

i.e., $Q_{l,r}(K)$ is the space of polynomials of degree at most l in x_1 and r in x_2, $l, r \geq 0$.

3.4.2.1 The RT Spaces on Rectangles

These spaces are an extension of the RT spaces on triangles to rectangles (Raviart-Thomas, 1977), and for each $r \geq 0$ are defined by

$$\mathbf{V}_h(K) = Q_{r+1,r}(K) \times Q_{r,r+1}(K), \quad W_h(K) = Q_{r,r}(K) \ .$$

In the case $r = 0$, $\mathbf{V}_h(K)$ takes the form

$$\mathbf{V}_h(K) = \{v : v = (a_K^1 + a_K^2 x_1, a_K^3 + a_K^4 x_2), \quad a_K^i \in \mathbb{R},\ i = 1, 2, 3, 4\} \ ,$$

and its dimension is four. The degrees of freedom for $\mathbf{V}_h$ are the values of normal components of functions at the midpoint on each edge in K_h (cf. Fig. 3.4). In this case, $Q_{0,0}(K) = P_0(K)$.

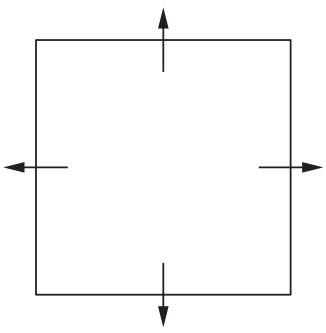

Fig. 3.4. The rectangular RT

For a general $r \geq 0$, the dimensions of $\mathbf{V}_h(K)$ and $W_h(K)$ are

$$\dim\big(\mathbf{V}_h(K)\big) = 2(r+1)(r+2), \quad \dim\big(W_h(K)\big) = (r+1)^2 \ .$$

The degrees of freedom for $\mathbf{V}_h(K)$ are given by (cf. Exercise 3.15)

$$\begin{aligned} &(\mathbf{v} \cdot \boldsymbol{\nu}, w)_e \qquad \forall w \in P_r(e),\ e \in \partial K \ , \\ &(\mathbf{v}, \mathbf{w})_K \qquad \forall \mathbf{w} = (w_1, w_2),\ w_1 \in Q_{r-1,r}(K),\ w_2 \in Q_{r,r-1}(K) \ . \end{aligned}$$

3.4.2.2 The BDM Spaces on Rectangles

The BDM spaces (Brezzi et al., 1985) on rectangles differ considerably from the RT spaces on rectangles in that the vector elements are based on augmenting the space of vector polynomials of total degree r by exactly two additional vectors in place of augmenting the space of vector tensor-products of polynomials of degree r by $2r + 2$ polynomials of higher degree. A lower dimensional space for the scalar variable is also used. These spaces, for any $r \geq 1$ are given by

$$\begin{aligned} &\mathbf{V}_h(K) = \big(P_r(K)\big)^2 \oplus \operatorname{span}\left\{\mathbf{curl}\ \left(x_1^{r+1} x_2\right), \mathbf{curl}\ \left(x_1 x_2^{r+1}\right)\right\} \ , \\ &W_h(K) = P_{r-1}(K) \ . \end{aligned}$$

In the case $r = 1$, $\mathbf{V}_h(K)$ is

$$\mathbf{V}_h(K) = \{v : v = \big(a_K^1 + a_K^2 x_1 + a_K^3 x_2 - a_K^4 x_1^2 - 2a_K^5 x_1 x_2 \ , \\ a_K^6 + a_K^7 x_1 + a_K^8 x_2 + 2a_K^4 x_1 x_2 + a_K^5 x_2^2\big) \ , \\ a_K^i \in \mathbb{R}, \ i = 1, 2, \ldots, 8\} \ ,$$

and its dimension is eight. The degrees of freedom for $\mathbf{V}_h$ are the values of normal components of functions at the two quadratic Gauss points on each edge in K_h (cf. Fig. 3.5).

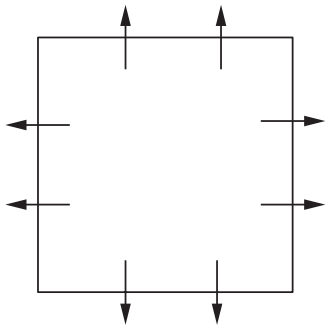

Fig. 3.5. The rectangular BDM

For any $r \geq 1$, the dimensions of $\mathbf{V}_h(K)$ and $W_h(K)$ are

$$\dim\big(\mathbf{V}_h(K)\big) = (r+1)(r+2) + 2, \quad \dim\big(W_h(K)\big) = \frac{r(r+1)}{2} \ .$$

The degrees of freedom for $\mathbf{V}_h(K)$ are (cf. Exercise 3.16)

$$(\mathbf{v} \cdot \boldsymbol{\nu}, w)_e \qquad \forall w \in P_r(e), \ e \in \partial K \ , \\ (\mathbf{v}, \mathbf{w})_K \qquad \forall \mathbf{w} \in \big(P_{r-2}(K)\big)^2 \ .$$

3.4.2.3 The BDFM Spaces on Rectangles

These spaces (Brezzi et al., 1987B) are related to the BDM spaces on rectangles and are also called reduced BDM spaces. They give the same rates of convergence as the corresponding RT spaces with fewer parameters per rectangle except for the lowest degree space. For each $r \geq 0$, they are defined by

$$\mathbf{V}_h(K) = \{w \in P_{r+1}(K) : \text{ the coefficient of } x_2^{r+1} \text{ vanishes}\} \\ \times \{w \in P_{r+1}(K) : \text{ the coefficient of } x_1^{r+1} \text{ vanishes}\}, \\ W_h(K) = P_r(K) \ .$$

In the case $r = 0$, the BDFM spaces are just the RT spaces on rectangles. For a general $r \geq 0$, the dimensions of $\mathbf{V}_h(K)$ and $W_h(K)$ are

$$\dim\big(\mathbf{V}_h(K)\big) = (r+2)(r+3) - 2, \quad \dim\big(W_h(K)\big) = \frac{(r+1)(r+2)}{2} .$$

The degrees of freedom for $\mathbf{V}_h(K)$ are defined by (cf. Exercise 3.17)

$$\begin{aligned} &(\mathbf{v}\cdot\boldsymbol{\nu}, w)_e && \forall w \in P_r(e),\ e \in \partial K , \\ &(\mathbf{v}, \mathbf{w})_K && \forall \mathbf{w} \in \big(P_{r-1}(K)\big)^2 . \end{aligned}$$

While rectangular elements are presented, an extension to general quadrilaterals can be made through change of variables from a reference rectangular element to quadrilaterals (Wang-Mathew, 1994); refer to Sect. 1.5.

3.4.3 Mixed Finite Element Spaces on Tetrahedra

Let K_h be a partition of $\Omega \subset \mathbb{R}^3$ into tetrahedra such that adjacent elements completely share their common face. In three dimensions, P_r is now the space of polynomials of degree r in three variables x_1, x_2, and x_3.

3.4.3.1 The RTN Spaces on Tetrahedra

These spaces (Néd'elec, 1980) are the three dimensional analogues of the RT spaces on triangles, and they are defined for each $r \geq 0$ by

$$\mathbf{V}_h(K) = \big(P_r(K)\big)^3 \oplus \big((x_1, x_2, x_3)P_r(K)\big), \quad W_h(K) = P_r(K) ,$$

where $(x_1, x_2, x_3)P_r(K) = (x_1 P_r(K), x_2 P_r(K), x_3 P_r(K))$. As in two dimensions, for $r = 0$, $\mathbf{V}_h$ is

$$\begin{aligned} \mathbf{V}_h(K) = \{v : v = (&a_K + b_K x_1, c_K + b_K x_2, d_K + b_K x_3), \\ &a_K,\ b_K,\ c_K \in \mathbb{R}\} , \end{aligned}$$

and its dimension is four. The degrees of freedom are the values of normal components of functions at the centroid of each face in K (cf. Fig. 3.6).

In general, for $r \geq 0$ the dimensions of $\mathbf{V}_h(K)$ and $W_h(K)$ are

$$\begin{aligned} \dim\big(\mathbf{V}_h(K)\big) &= \frac{(r+1)(r+2)(r+4)}{2} , \\ \dim\big(W_h(K)\big) &= \frac{(r+1)(r+2)(r+3)}{6} . \end{aligned}$$

The degrees of freedom for $\mathbf{V}_h(K)$ are

$$\begin{aligned} &(\mathbf{v}\cdot\boldsymbol{\nu}, w)_e && \forall w \in P_r(e),\ e \in \partial K , \\ &(\mathbf{v}, \mathbf{w})_K && \forall \mathbf{w} \in \big(P_{r-1}(K)\big)^3 . \end{aligned}$$

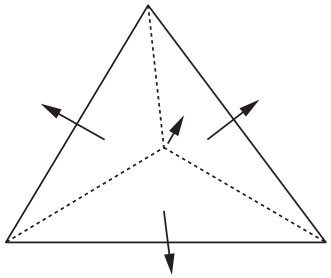

Fig. 3.6. The RTN on a tetrahedron

3.4.3.2 The BDDF Spaces on Tetrahedra

The BDDF spaces (Brezzi et al., 1987A) are an extension of the BDM spaces on triangles to tetrahedra, and they are given for each $r \geq 1$ by

$$\mathbf{V}_h(K) = \big(P_r(K)\big)^3, \quad W_h(K) = P_{r-1}(K) \; .$$

The dimensions of $\mathbf{V}_h(K)$ and $W_h(K)$ are

$$\dim\big(\mathbf{V}_h(K)\big) = \frac{(r+1)(r+2)(r+3)}{2} \; ,$$

$$\dim\big(W_h(K)\big) = \frac{r(r+1)(r+2)}{6} \; .$$

The degrees of freedom for $\mathbf{V}_h(K)$ are

$$\begin{aligned}
&(\mathbf{v}\cdot\boldsymbol{\nu}, w)_e \qquad \forall w \in P_r(e),\ e \in \partial K \; , \\
&(\mathbf{v}, \nabla w)_K \qquad \forall w \in P_{r-1}(K) \; , \\
&(\mathbf{v}, \mathbf{w})_K \qquad \forall \mathbf{w} \in \{\mathbf{z} \in (P_r(K))^3 : \mathbf{z}\cdot\boldsymbol{\nu} = 0 \text{ on } \partial K \\
&\qquad\qquad\qquad \text{and } (\mathbf{z}, \nabla w)_K = 0,\ w \in P_{r-1}(K)\} \; .
\end{aligned}$$

3.4.4 Mixed Finite Element Spaces on Parallelepipeds

Let $\Omega \subset \mathbb{R}^3$ be a rectangular domain and K_h be a partition of Ω into rectangular parallelepipeds such that their faces are parallel to the coordinate axes and adjacent elements completely share their common face. Define, with $\mathbf{x} = (x_1, x_2, x_3)$,

$$Q_{l,m,r}(K) = \left\{ v : \; v(\mathbf{x}) = \sum_{i=0}^{l}\sum_{j=0}^{m}\sum_{k=0}^{r} v_{ijk} x_1^i x_2^j x_3^k,\ \mathbf{x} \in K,\ v_{ijk} \in \mathbb{R} \right\} ;$$

i.e., $Q_{l,m,r}(K)$ is the space of polynomials of degree at most l in x_1, m in x_2, and r in x_3 on K, respectively, $l, m, r \geq 0$.

3.4.4.1 The RTN Spaces on Rectangular Parallelepipeds

These spaces (Néd'elec, 1980) are the three dimensional analogues of the RT spaces on rectangles and for each $r \geq 0$ are defined by

$$\begin{aligned}\mathbf{V}_h(K) &= Q_{r+1,r,r}(K) \times Q_{r,r+1,r}(K) \times Q_{r,r,r+1}(K)\ ,\\ W_h(K) &= Q_{r,r,r}(K)\ .\end{aligned}$$

For $r = 0$, $\mathbf{V}_h$ is

$$\begin{aligned}\mathbf{V}_h(K) = \{v : v = (a_K^1 + a_K^2 x_1, a_K^3 + a_K^4 x_2, a_K^5 + a_K^6 x_3)\ ,\\ a_K^i \in \mathbb{R},\ i = 1, 2, \ldots, 6\}\ ,\end{aligned}$$

and its dimension is six. The degrees of freedom are the values of normal components of functions at the centroid of each face in K (cf. Fig. 3.7).

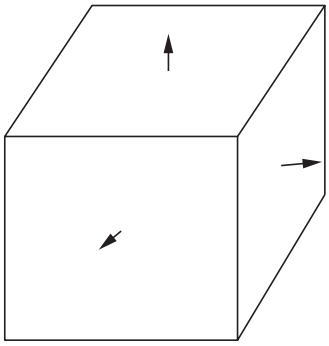

Fig. 3.7. The RTN on a cube

For $r \geq 0$, the dimensions of $\mathbf{V}_h(K)$ and $W_h(K)$ are

$$\dim\big(\mathbf{V}_h(K)\big) = 3(r+1)^2(r+2), \quad \dim\big(W_h(K)\big) = (r+1)^3\ ,$$

and the degrees of freedom for $\mathbf{V}_h(K)$ are

$$\begin{aligned}&(\mathbf{v}\cdot\boldsymbol{\nu}, w)_e \qquad \forall w \in Q_{r,r}(e),\ e \in \partial K\ ,\\ &(\mathbf{v}, \mathbf{w})_K \qquad \forall \mathbf{w} = (w_1, w_2, w_3),\ w_1 \in Q_{r-1,r,r}(K)\ ,\\ &\qquad\qquad w_2 \in Q_{r,r-1,r}(K),\ w_3 \in Q_{r,r,r-1}(K)\ .\end{aligned}$$

3.4.4.2 The BDDF Spaces on Rectangular Parallelepipeds

These spaces (Brezzi et al., 1987A) are the three dimensional analogues of the BDM spaces on rectangles. They are defined for $r \geq 1$ by

$$\mathbf{V}_h(K) = \big(P_r(K)\big)^3 \oplus \operatorname{span}\big\{\mathbf{curl}(0,0,x_1^{r+1}x_2),\ \mathbf{curl}(0,x_1x_3^{r+1},0)\ ,$$
$$\mathbf{curl}(x_2^{r+1}x_3,0,0),\ \mathbf{curl}(0,0,x_1x_2^{i+1}x_3^{r-i})\ ,$$
$$\mathbf{curl}(0,x_1^{i+1}x_2^{r-i}x_3,0),\ \mathbf{curl}(x_1^{r-i}x_2x_3^{i+1},0,0)\big\}\ ,$$
$$W_h(K) = P_{r-1}(K)\ ,$$

where $i = 1, 2, \ldots, r$ and, with $\mathbf{v} = (v_1, v_2, v_3)$,

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

The dimensions of $\mathbf{V}_h(K)$ and $W_h(K)$ are

$$\dim\big(\mathbf{V}_h(K)\big) = \frac{(r+1)(r+2)(r+3)}{2} + 3(r+1)\ ,$$

$$\dim\big(W_h(K)\big) = \frac{r(r+1)(r+2)}{6}\ .$$

The degrees of freedom for $\mathbf{V}_h(K)$ are

$$(\mathbf{v}\cdot\boldsymbol{\nu}, w)_e \qquad \forall w \in P_r(e),\ e \in \partial K\ ,$$
$$(\mathbf{v}, \mathbf{w})_K \qquad \forall \mathbf{w} \in (P_{r-2}(K))^3\ .$$

3.4.4.3 The BDFM Spaces on Rectangular Parallelepipeds

These spaces (Brezzi et al., 1987B) are related to the BDDF spaces on rectangular parallelepipeds and are also called the reduced BDDF spaces. They are defined for each $r \geq 0$ as

$$\mathbf{V}_h(K) = \left\{w \in P_{r+1}(K):\ \text{the coefficient of } \sum_{i=0}^{r+1} x_2^{r+1-i}x_3^i \text{ vanishes}\right\}$$
$$\times\left\{w \in P_{r+1}(K):\ \text{the coefficient of } \sum_{i=0}^{r+1} x_3^{r+1-i}x_1^i \text{ vanishes}\right\}$$
$$\times\left\{w \in P_{r+1}(K):\ \text{the coefficient of } \sum_{i=0}^{r+1} x_1^{r+1-i}x_2^i \text{ vanishes}\right\}\ ,$$

$$W_h(K) = P_r(K)\ .$$

The dimensions of $\mathbf{V}_h(K)$ and $W_h(K)$ are

$$\dim\big(\mathbf{V}_h(K)\big) = \frac{(r+2)(r+3)(r+4)}{2} - 3(r+2)\ ,$$

$$\dim\big(W_h(K)\big) = \frac{(r+1)(r+2)(r+3)}{6}\ .$$

The degrees of freedom for $\mathbf{V}_h(K)$ are

$$\begin{aligned} &(\mathbf{v}\cdot\boldsymbol{\nu}, w)_e && \forall w \in P_r(e),\ e \in \partial K\ , \\ &(\mathbf{v}, \mathbf{w})_K && \forall \mathbf{w} \in (P_{r-1}(K))^3\ . \end{aligned}$$

3.4.5 Mixed Finite Element Spaces on Prisms

Let $\Omega \subset \mathbb{R}^3$ be a domain of the form $\Omega = G \times (l_1, l_2)$, where $G \subset \mathbb{R}^2$ and l_1 and l_2 are real numbers. Let K_h be a partition of Ω into prisms such that their bases are triangles in the (x_1, x_2)-plane with three vertical edges parallel to the x_3-axis and adjacent prisms completely share their common face. $P_{l,r}$ denotes the space of polynomials of degree l in the two variables x_1 and x_2 and of degree r in the variable x_3.

3.4.5.1 The RTN Spaces on Prisms

These spaces (Néd'elec, 1986) are an extension of the RTN spaces on rectangular parallelepipeds to prisms and are defined for each $r \geq 0$ by

$$\mathbf{V}_h(K) = \big\{\mathbf{v} = (v_1, v_2, v_3) :\ v_3 \in P_{r,r+1}(K)\big\}, \quad W_h(K) = P_{r,r}(K)\ ,$$

where (v_1, v_2) satisfies that, for x_3 fixed,

$$(v_1, v_2) \in \big(P_r(K)\big)^2 \oplus \big((x_1, x_2)P_r(K)\big)\ ,$$

and v_1 and v_2 are of degree r in x_3. For $r = 0$, $\mathbf{V}_h$ has the form

$$\begin{aligned} \mathbf{V}_h(K) = \{v :\ v = (&a_K^1 + a_K^2 x_1, a_K^3 + a_K^2 x_2, a_K^4 + a_K^5 x_3)\ , \\ &a_K^i \in \mathbb{R},\ i = 1, 2, \ldots, 5\}\ , \end{aligned}$$

and its dimension is five. The degrees of freedom are the values of normal components of functions at the centroid of each face in K (cf. Fig. 3.8).

For $r \geq 0$, the dimensions of $\mathbf{V}_h(K)$ and $W_h(K)$ are

$$\dim\big(\mathbf{V}_h(K)\big) = (r+1)^2(r+3) + \frac{(r+1)(r+2)^2}{2}\ ,$$

$$\dim\big(W_h(K)\big) = \frac{(r+1)^2(r+2)}{2}\ .$$

The degrees of freedom for $\mathbf{V}_h(K)$ are

$$\begin{aligned} &(\mathbf{v}\cdot\boldsymbol{\nu}, w)_e && \forall w \in P_r(e) \text{ for the two horizontal faces}\ , \\ &(\mathbf{v}\cdot\boldsymbol{\nu}, w)_e && \forall w \in Q_{r,r}(e) \text{ for the three vertical faces}\ , \\ &\big((v_1, v_2), (w_1, w_2)\big)_K && \forall (w_1, w_2) \in (P_{r-1,r}(K))^2\ , \\ &(v_3, w_3)_K && \forall w_3 \in P_{r,r-1}(K)\ . \end{aligned}$$

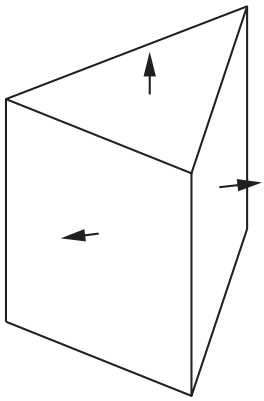

Fig. 3.8. The RTN on a prism

3.4.5.2 The First CD Spaces on Prisms

The first CD spaces (Chen-Douglas, 1989) are an analogue of the RTN spaces on prisms, but different degrees of freedom are used and the number of these degrees is smaller than required by the RNT spaces. They are defined for each $r \geq 0$ by

$$\mathbf{V}_h(K) = \{\mathbf{v} = (v_1, v_2, v_3) : \ (v_1, v_2) \in (P_{r+1,r}(K))^2, v_3 \in P_{r,r+1}(K)\} ,$$
$$W_h(K) = P_{r,r}(K) ,$$

where the dimensions of $\mathbf{V}_h(K)$ and $W_h(K)$ are

$$\dim\big(\mathbf{V}_h(K)\big) = (r+1)(r+2)(r+3) + \frac{(r+1)(r+2)^2}{2} ,$$

$$\dim\big(W_h(K)\big) = \frac{(r+1)^2(r+2)}{2} .$$

Let

$$B_{r+2,r}(K) = \{v \in P_{r+2,r}(K) : v|_e = 0 \text{ on the three vertical faces}\} .$$

The degrees of freedom for $\mathbf{V}_h(K)$ are

$$\begin{aligned}
&(\mathbf{v} \cdot \boldsymbol{\nu}, w)_e && \forall w \in P_r(e) \text{ for the two horizontal faces} ,\\
&(\mathbf{v} \cdot \boldsymbol{\nu}, w)_e && \forall w \in Q_{r+1,r}(e) \text{ for the three vertical faces} ,\\
&\big((v_1, v_2), \nabla_{(x_1,x_2)} w\big)_K && \forall w \in P_{r,r}(K) ,\\
&\big((v_1, v_2), \mathbf{curl}_{(x_1,x_2)} w\big)_K && \forall w \in B_{r+2,r}(K) ,\\
&(v_3, w_3)_K && \forall w_3 \in P_{r,r-1}(K) ,
\end{aligned}$$

where $\nabla_{(x_1,x_2)}$ and $\mathbf{curl}_{(x_1,x_2)}$ indicate the corresponding operators with respect to x_1 and x_2.

3.4.5.3 The Second CD Spaces on Prisms

The second CD spaces (Chen-Douglas, 1989) are based on the BDDF spaces on rectangular parallelepipeds and use a much smaller number of degrees of freedom than the RTN and first CD spaces on prisms. They are defined for each $r \geq 1$ by

$$\begin{aligned}\mathbf{V}_h(K) = \big(P_r(K)\big)^3 \oplus \operatorname{span}\big\{&\mathbf{curl}(x_2^{r+1}x_3, 0, 0)\ ,\\ &\mathbf{curl}(x_2x_3^{r+1}, -x_1x_3^{r+1}, 0)\ ,\\ &\mathbf{curl}(0, x_1^{i+1}x_2^{r-i}x_3, 0),\ i = 1, 2, \ldots, r\big\}\ ,\\ W_h(K) = P_{r-1}(K)\ .\end{aligned}$$

The dimensions of $\mathbf{V}_h(K)$ and $W_h(K)$ are

$$\dim\big(\mathbf{V}_h(K)\big) = \frac{(r+1)(r+2)(r+3)}{2} + r + 2\ ,$$

$$\dim\big(W_h(K)\big) = \frac{r(r+1)(r+2)}{6}\ .$$

Let

$$B_{r+1}(K) = \{v \in P_{r+1}(K) : v|_e = 0 \text{ on the three vertical faces of } K\}\ .$$

The degrees of freedom for $\mathbf{V}_h(K)$ are

$$\begin{aligned}&(\mathbf{v}\cdot\boldsymbol{\nu}, w)_e && \forall w \in P_r(e),\ e \in \partial K\ ,\\ &\big((v_1, v_2), \nabla_{(x_1,x_2)}w\big)_K && \forall w \in P_{r-1}(K)\ ,\\ &\big((v_1, v_2),\ \mathbf{curl}_{(x_1,x_2)}w\big)_K && \forall w \in B_{r+1}(K)\ ,\\ &(v_3, w_3)_K && \forall w_3 \in P_{r-2}(K)\ .\end{aligned}$$

3.4.5.4 The Third CD Spaces on Prisms

The third CD spaces (Chen-Douglas, 1989) are based on the BDFM spaces on rectangular parallelepipeds and also use a much smaller number of degrees of freedom than the RTN and first CD spaces on prisms. They are defined for each $r \geq 0$ by

$$\begin{aligned}\mathbf{V}_h(K) = \big\{&w \in P_{r+1}(K) :\ \text{the coefficient of } x_3^{r+1} \text{ vanishes}\big\}\\ &\times \big\{w \in P_{r+1}(K) :\ \text{the coefficient of } x_3^{r+1} \text{ vanishes}\big\}\\ &\times \left\{w \in P_{r+1}(K) :\ \text{the coefficient of } \sum_{i=0}^{r+1} x_1^{r+1-i}x_2^i \text{ vanishes}\right\}\ ,\\ W_h(K) = P_r(K)\ .\end{aligned}$$

The dimensions of $\mathbf{V}_h(K)$ and $W_h(K)$ are

$$\dim\big(\mathbf{V}_h(K)\big) = \frac{(r+2)(r+3)(r+4)}{2} - r - 4 \,,$$
$$\dim\big(W_h(K)\big) = \frac{(r+1)(r+2)(r+3)}{6} \,.$$

The degrees of freedom for $\mathbf{V}_h(K)$ are

$$\begin{aligned}
&(\mathbf{v}\cdot\boldsymbol{\nu}, w)_e && \forall w \in P_r(e) \text{ for the two horizontal faces} \,, \\
&(\mathbf{v}\cdot\boldsymbol{\nu}, w)_e && \forall w \in P_{r+1} \setminus \{x_3^{r+1}\}|_e \\
& && \text{for the three vertical faces} \,, \\
&\big((v_1, v_2), \nabla_{(x_1,x_2)} w\big)_K && \forall w \in P_{r-1}(K) \,, \\
&\big((v_1, v_2), \mathbf{curl}_{(x_1,x_2)} w\big)_K && \forall w \in B_{r+2}(K) \,, \\
&(v_3, w_3)_K && \forall w_3 \in P_{r-1}(K) \,.
\end{aligned}$$

The mixed finite element spaces presented in this section satisfy the *inf-sup* condition (3.32) (cf. Sect. 3.8) and lead to optimal approximation properties (see the next section). In this section, we have considered only a polygonal domain Ω. For a more general domain, the partition T_h can have curved edges or faces on the boundary Γ, and the mixed spaces are constructed in a similar fashion (Raviart-Thomas, 1977; Néd'elec, 1980; Brezzi et al., 1985, 1987A, 1987B; Chen-Douglas, 1989).

3.5 Approximation Properties

The RTN, BDM, BDFM, BDDF, and CD mixed finite element spaces have the approximation properties

$$\begin{aligned}
&\inf_{\mathbf{v}_h \in \mathbf{V}_h} \|\mathbf{v} - \mathbf{v}_h\| \le C h^l \|\mathbf{v}\|_{\mathbf{H}^l(\Omega)}, && 1 \le l \le r+1 \,, \\
&\inf_{\mathbf{v}_h \in \mathbf{V}_h} \|\nabla\cdot(\mathbf{v} - \mathbf{v}_h)\| \le C h^l \|\nabla\cdot\mathbf{v}\|_{H^l(\Omega)}, && 0 \le l \le r^* \,, \\
&\inf_{w_h \in W_h} \|w - w_h\| \le C h^l \|w\|_{H^l(\Omega)}, && 0 \le l \le r^* \,,
\end{aligned} \tag{3.34}$$

where $r^* = r+1$ for the RTN, BDFM, and first and third CD spaces and $r^* = r$ for the BDM, BDDF, and second CD spaces. Using (3.34), we can establish the corresponding error estimates for the mixed finite element method (3.17) when $\mathbf{V}_h$ and W_h are these mixed spaces; refer to Sect. 3.8.

3.6 Mixed Methods for Nonlinear Problems

The mixed finite element method was considered by Johnson-Thomée (1981) for a linear parabolic problem and by Chen-Douglas (1991) for the following nonlinear parabolic problem:

$$\begin{aligned} &c(p)\frac{\partial p}{\partial t} - \nabla \cdot \big(a(p)\nabla p\big) = f(p) && \text{in } \Omega \times J\ , \\ &p = 0 && \text{on } \Gamma \times J\ , \\ &p(\cdot, 0) = p_0 && \text{in } \Omega\ , \end{aligned} \tag{3.35}$$

where $c(p) = c(\mathbf{x}, t, p)$, $a(p) = a(\mathbf{x}, t, p)$, $f(p) = f(\mathbf{x}, t, p)$, $J = (0, T]$ $(T > 0)$, and $\Omega \subset \mathbb{R}^d$, $d = 2$ or 3. This problem has been studied in the preceding two chapters for the conforming and nonconforming finite element methods. Here we very briefly describe an application of the mixed method. We assume that (3.35) admits a unique solution. Furthermore, we assume that the coefficients $c(p)$, $a(p)$, and $f(p)$ are *globally Lipschitz continuous* in p; i.e., for some constants C_ξ, they satisfy

$$|\xi(p_1) - \xi(p_2)| \le C_\xi |p_1 - p_2|, \qquad p_1,\, p_2 \in \mathbb{R},\ \xi = c,\ a,\ f\ . \tag{3.36}$$

Set

$$\mathbf{u} = -a(p)\nabla p, \quad \mathbf{V} = \mathbf{H}(\mathrm{div}, \Omega), \quad W = L^2(\Omega)\ .$$

Then (3.35) can be recast in the mixed formulation:

$$\begin{aligned} &\text{Find } \mathbf{u} : J \to \mathbf{V} \text{ and } p : J \to W \text{ such that} \\ &\big(a^{-1}(p)\mathbf{u}, \mathbf{v}\big) - (\nabla \cdot \mathbf{v}, p) = 0, && \mathbf{v} \in \mathbf{V},\ t \in J\ , \\ &\left(c(p)\frac{\partial p}{\partial t}, w\right) + (\nabla \cdot \mathbf{u}, w) = \big(f(p), w\big), && w \in W,\ t \in J\ , \end{aligned} \tag{3.37}$$

with $p(\cdot, 0) = p_0$.

Let $\mathbf{V}_h \times W_h \subset \mathbf{V} \times W$ be any of the mixed finite element spaces introduced in Sect. 3.4. The mixed finite element method for (3.35) is

$$\begin{aligned} &\text{Find } \mathbf{u}_h : J \to \mathbf{V}_h \text{ and } p_h : J \to W_h \text{ such that} \\ &\big(a^{-1}(p_h)\mathbf{u}_h, \mathbf{v}\big) - (\nabla \cdot \mathbf{v}, p_h) = 0, && \mathbf{v} \in \mathbf{V}_h\ , \\ &\left(c(p_h)\frac{\partial p_h}{\partial t}, w\right) + (\nabla \cdot \mathbf{u}_h, w) = \big(f(p_h), w\big), && w \in W_h\ , \end{aligned} \tag{3.38}$$

where $p_h(\cdot, 0)$ can be any appropriate projection of p_0 in W_h, e.g., its L^2-projection in W_h:

$$(p_h(\cdot, 0) - p_0, w) = 0, \qquad w \in W_h\ .$$

After the introduction of basis functions in $\mathbf{V}_h$ and W_h, as in Sect. 3.2, (3.38) can be written in the matrix form

$$\begin{aligned} &\mathbf{A}(\mathbf{p})\mathbf{U} + \mathbf{B}\mathbf{p} = \mathbf{0}, && t \in J\ , \\ &\mathbf{C}(\mathbf{p})\frac{d\mathbf{p}}{dt} - \mathbf{B}^T\mathbf{U} = \mathbf{f}(\mathbf{p}), && t \in J\ . \end{aligned} \tag{3.39}$$

Under the assumption that the coefficient $c(p)$ is bounded below by a positive constant, this nonlinear system of ODEs locally has a unique solution for h small enough. In fact, because of assumption (3.36) about c, a, and f, the solution $\mathbf{U}(t)$, $\mathbf{p}(t)$ exists for all t for h small enough (Chen-Douglas, 1991). The various solution approaches (e.g., linearization, implicit time approximation, and explicit time approximation) developed in Sect. 1.8 for the finite element method can be applied to (3.39) in the same fashion. We conclude with a remark that the mixed finite element method has been also studied for stationary nonlinear problems (Milner, 1985; Chen, 1989).

3.7 Linear System Solution Techniques

3.7.1 Introduction

As discussed in the previous sections, the system arising from the mixed finite element method is of the form

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{U} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{g} \\ -\mathbf{f} \end{pmatrix} . \tag{3.40}$$

As noted, while the matrix

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{0} \end{pmatrix}$$

is nonsingular under the *inf-sup* condition (3.32), it is not positive definite. This limits application of many iterative algorithms to (3.40).

Since $\mathbf{A}$ is positive definite, $\mathbf{U}$ can be eliminated from the first equation of (3.40):

$$\mathbf{U} = \mathbf{A}^{-1}\mathbf{g} - \mathbf{A}^{-1}\mathbf{B}\mathbf{p} .$$

Substitute this equation into the second equation of (3.40) to see that

$$\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\mathbf{p} = \mathbf{B}^T\mathbf{A}^{-1}\mathbf{g} + \mathbf{f} , \tag{3.41}$$

so we have a single system for $\mathbf{p}$. By (3.32), the matrix $\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}$ is symmetric and positive definite. Hence system (3.41) is easier to solve than system (3.40). The *Uzawa algorithm* is a particular implementation of an iterative algorithm for solving (3.41); see Sect. 3.7.2. A common problem with such an algorithm is that the action of the matrix $\mathbf{A}^{-1}$ in each step of the iteration needs to be computed, and this computation is generally expensive.

There exist iterative algorithms for solving (3.40) without the inversion of $\mathbf{A}$. The *minimal residual algorithm* can be applied to a more direct preconditioned reformulation of (3.40), for example; refer to Sect. 3.7.3. There also exist a variety of specific algorithms that strongly depend on the underlying

differential equation problem and the choice of mixed finite element spaces. These include alternating direction iterative algorithms (cf. Sect. 3.7.4) and mixed-hybrid algorithms (cf. Sect. 3.7.5). The mixed finite element method is related to the nonconforming method studied in the preceding chapter, and can be implemented by the latter (cf. Sect. 3.7.6).

3.7.2 The Uzawa Algorithm

The Uzawa algorithm (Arrow et al., 1958) is a classical iterative algorithm for saddle point problems. It is defined as follows: Given an initial guess $\mathbf{p}^0 \in \mathbb{R}^N$, find $(\mathbf{U}^k, \mathbf{p}^k) \in \mathbb{R}^M \times \mathbb{R}^N$ such that, for $k = 1, 2, \ldots$,

$$\begin{aligned} \mathbf{A}\mathbf{U}^k &= \mathbf{g} - \mathbf{B}\mathbf{p}^{k-1} \ , \\ \mathbf{p}^k &= \mathbf{p}^{k-1} + \alpha \left(\mathbf{B}^T \mathbf{U}^k + \mathbf{f} \right) \ , \end{aligned} \tag{3.42}$$

where α is a given real number. To see convergence of this algorithm, we define the residual

$$\mathbf{e}^k = -\mathbf{B}^T \mathbf{U}^k - \mathbf{f} \ .$$

Using (3.41) and (3.42), we see that

$$\mathbf{e}^k = -\mathbf{B}^T \mathbf{A}^{-1} \left(\mathbf{g} - \mathbf{B}\mathbf{p}^{k-1} \right) - \mathbf{f} = -\mathbf{B}^T \mathbf{A}^{-1} \mathbf{B} \left(\mathbf{p} - \mathbf{p}^{k-1} \right) \ ,$$

so

$$\mathbf{p}^k - \mathbf{p}^{k-1} = -\alpha \mathbf{e}^k = \alpha \mathbf{B}^T \mathbf{A}^{-1} \mathbf{B} \left(\mathbf{p} - \mathbf{p}^{k-1} \right) \ .$$

Hence the Uzawa algorithm is equivalent to applying a gradient algorithm to (3.41) using a fixed step size α. From the analysis of the gradient algorithm (Axelsson, 1994; Golub-van Loan, 1996), the iteration converges if

$$\alpha < 2 \left\| \mathbf{B}^T \mathbf{A}^{-1} \mathbf{B} \right\|^{-1} \ ,$$

where $\| \cdot \|$ is the matrix norm induced from the usual real Euclidean norm (i.e., the ℓ_2-norm).

The step size can be varied and at each step can be chosen, for example:

$$\alpha_k = \frac{\mathbf{e}^k \cdot \mathbf{e}^k}{(\mathbf{B}\mathbf{e}^k) \cdot (\mathbf{A}^{-1}\mathbf{B}\mathbf{e}^k)} \ .$$

Note that if we use this choice, we would invert $\mathbf{A}$ in every step of the iteration. That can be avoided by storing an anuxiliary vector. This approach leads to a two-level iteration: an inner iteration for solving a system with the stiffness matrix $\mathbf{A}$ and the outer Uzawa iteration (3.42).

As in Sect. 1.10.2, due to a large condition number of $\mathbf{B}^T \mathbf{A}^{-1} \mathbf{B}$, it is more effective to use a conjugate gradient method for solving (3.41). This leads to a modified Uzawa algorithm:

- Given an initial guess $\mathbf{p}^0 \in \mathbb{R}^N$, solve $\mathbf{A}\mathbf{U}^1 = \mathbf{g} - \mathbf{B}\mathbf{p}^0$;

- Set $\mathbf{d}^1 = -\mathbf{e}^1 = \mathbf{B}^T\mathbf{U}^1 + \mathbf{f}$;
- For $k = 1, 2, \ldots$, find $\mathbf{U}^{k+1} \in \mathbb{R}^M$ and $\mathbf{p}^k \in \mathbb{R}^N$ such that

$$\begin{aligned}
&\mathbf{r}^k = \mathbf{B}\mathbf{d}^k \ , \\
&\mathbf{s}^k = \mathbf{A}^{-1}\mathbf{r}^k \ , \\
&\mathbf{p}^k = \mathbf{p}^{k-1} + \alpha_k \mathbf{d}^k \quad \text{where } \alpha_k = \frac{\mathbf{e}^k \cdot \mathbf{e}^k}{\mathbf{r}^k \cdot \mathbf{s}^k} \ , \\
&\mathbf{U}^{k+1} = \mathbf{U}^k - \alpha_k \mathbf{s}^k \ , \\
&\mathbf{e}^{k+1} = -\mathbf{B}^T\mathbf{U}^{k+1} - \mathbf{f} \ , \\
&\mathbf{d}^{k+1} = -\mathbf{e}^{k+1} + \beta_k \mathbf{d}^k \quad \text{where } \beta_k = \frac{\mathbf{e}^{k+1} \cdot \mathbf{e}^{k+1}}{\mathbf{e}^k \cdot \mathbf{e}^k} \ .
\end{aligned} \tag{3.43}$$

As discussed, algorithms (3.42) and (3.43) require the evaluation of the action of the matrix $\mathbf{A}^{-1}$ at each step of the iteration. There are so-called *inexact Uzawa algorithms* that replace the exact inverse by an approximate evaluation of $\mathbf{A}^{-1}$ (Elman-Golub, 1994; Bramble et al., 1997). Also, since the Uzawa algorithms converge slowly, one can introduce their preconditioned versions, as discussed in Sect. 1.10.2 (also see the next subsection).

3.7.3 The Minimum Residual Iterative Algorithm

The *minimum residual iterative algorithm* (Paige-Saunders, 1975) can be used to solve system (3.40). As previously, let the dimensions of $\mathbf{V}_h$ and W_h be M and N, respectively. Because $\mathbf{M} = (m_{ij})$ is symmetric and nonsingular, we can define the "energy" inner product

$$\langle \mathbf{v}, \mathbf{w} \rangle_{\mathbf{M}} = \langle \mathbf{M}\mathbf{v}, \mathbf{M}\mathbf{w} \rangle = \sum_{i,j,k=1}^{MN} v_j m_{ij} m_{ik} w_k, \qquad \mathbf{v},\, \mathbf{w} \in \mathbb{R}^{MN} \ ,$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product in $\mathbb{R}^{MN}$. Set

$$\mathbf{F} = \begin{pmatrix} \mathbf{g} \\ -\mathbf{f} \end{pmatrix} \ .$$

The minimum residual iterative algorithm for finding the approximations $\mathbf{P}^k \in \mathbb{R}^{MN}$ $(k = 1, 2, \ldots)$ to the solution of (3.40) is defined as follows:

- Set $\mathbf{P}^0 = \mathbf{x}^0 = 0$ and $\beta_1 \mathbf{P}^1 = \mathbf{x}^1 = \mathbf{F}$;
- For $k = 1, 2, \ldots$, find $\mathbf{P}^{k+1} \in \mathbb{R}^{MN}$ and $\mathbf{x}^{k+1} \in \mathbb{R}^{MN}$ by

$$\begin{aligned}
&\mathbf{r}^k = \mathbf{F} - \mathbf{M}\mathbf{P}^k \ , \\
&\beta_{k+1}\mathbf{x}^{k+1} = \mathbf{M}\mathbf{x}^k - \left\langle \mathbf{M}\mathbf{x}^k, \mathbf{x}^k \right\rangle_{\mathbf{M}} \mathbf{x}^k - \left\langle \mathbf{M}\mathbf{x}^k, \mathbf{x}^{k-1} \right\rangle_{\mathbf{M}} \mathbf{x}^{k-1} \ , \\
&\mathbf{P}^{k+1} = \mathbf{P}^k + \left\langle \mathbf{r}^k, \mathbf{M}\mathbf{x}^{k+1} \right\rangle \mathbf{x}^{k+1} \ ,
\end{aligned}$$

where the constants $\beta_k > 0$ are chosen such that $\langle \mathbf{x}^k, \mathbf{x}^k \rangle_{\mathbf{M}} = 1$. This is possible: When $\mathbf{r}^k \neq \mathbf{0}$, we can show that $\beta_{k+1}\mathbf{x}^{k+1} \neq \mathbf{0}$ (Rusten-Winther, 1992); when $\mathbf{r}^k = \mathbf{0}$, the above iteration stops.

The convergence rate of this algorithm depends on the location of eigenvalues of $\mathbf{M}$. It can be shown (Rusten-Winther, 1992) that this rate can be estimated by the condition numbers of $\mathbf{A}$ and $\mathbf{B}$. Since their condition numbers increase as the discretization is refined and the convergence is thus slow, a direct application of the minimum residual algorithm is usually not practical. Therefore, to speed up the convergence, preconditioned versions of this algorithm have been suggested (Ewing et al., 1990; Rusten-Winther, 1992). For completeness, we briefly mention this technique.

Let $\mathbf{L} \in \mathbb{R}^{M\times M}$ and $\mathbf{S} \in \mathbb{R}^{N\times N}$ be two nonsingular matrices. Then system (3.40) is equivalent to the system

$$\begin{aligned} &\mathbf{L}^{-1}\mathbf{A}\mathbf{L}^{-T}\mathbf{v} + \mathbf{L}^{-1}\mathbf{B}\mathbf{S}^{-1}\mathbf{q} = \mathbf{L}^{-1}\mathbf{g}\,, \\ &\left(\mathbf{L}^{-1}\mathbf{B}\mathbf{S}^{-1}\right)^T \mathbf{v} = -\mathbf{S}^{-T}\mathbf{f}\,, \end{aligned} \tag{3.44}$$

where $\mathbf{v} = \mathbf{L}^T\mathbf{U}$ and $\mathbf{q} = \mathbf{S}\mathbf{p}$. System (3.44) has the same structure as (3.40). The minimum residual algorithm applied to (3.44) converges faster if $\mathbf{L}$ and $\mathbf{S}$ are appropriately chosen. The matrices $\mathbf{L}$ and $\mathbf{S}$ should have the property that linear systems with coefficient matrices given by $\mathbf{L}\mathbf{L}^T$ or $\mathbf{S}^T\mathbf{S}$ can be solved by a fast solver. This requirement is necessary since such linear systems have to be solved once in each iteration of the preconditioned minimum residual algorithm. One example of the choices for $\mathbf{L}$ and $\mathbf{S}$ is that $\mathbf{L} = \mathbf{I}$, the identity matrix, and $\mathbf{S}$ should be chosen such that $\mathbf{S}^T\mathbf{S}$ is a preconditioner for $\mathbf{B}^T\mathbf{B}$. $\mathbf{S}^T\mathbf{S}$ can be obtained from the incomplete Cholesky factorization of $\mathbf{B}^T\mathbf{B}$ (Rusten-Winther, 1992), for example.

3.7.4 Alternating Direction Iterative Algorithms

The Uzawa and Arrow-Hurwitz alternating-direction iterative algorithms have been developed for solving the system of algebraic equations arising from the mixed finite element method considered in this chapter (Brezzi et al., 1987A,B; Douglas et al., 1987). We now describe these iterative algorithms for solving (3.40). As an example, we limit ourselves to the Uzawa-type algorithms for the Raviart-Thomas spaces on rectangles; the Arrow-Hurwitz-type algorithms and other mixed finite element families can be treated as well (Douglas et al., 1987).

The Uzawa alternating-direction algorithms are based on a *virtual parabolic problem* introduced by adding a virtual time derivative of $\mathbf{p}$ to the second equation of (3.40) and initiating the resulting evolution by an initial guess for $\mathbf{p}$. Thus we consider the system

$$\begin{aligned} &\mathbf{A}\mathbf{U} + \mathbf{B}\mathbf{p} = \mathbf{g}, && t \geq 0\ , \\ &\mathbf{D}\frac{d\mathbf{p}}{dt} - \mathbf{B}^T\mathbf{U} = \mathbf{f}, && t \geq 0\ , \\ &\mathbf{p}(0) = \mathbf{p}^0, \end{aligned} \tag{3.45}$$

where the choice of $\mathbf{D}$ is somewhat arbitrary, though it should be symmetric and positive definite. System (3.45) corresponds to a mixed finite element method for an initial value problem:

$$\tilde{d}\frac{\partial p}{\partial t} - \nabla \cdot (a \nabla p) = f\ , \tag{3.46}$$

for some coefficient $\tilde{d}$ and with an appropriate boundary condition. Let now the domain Ω be a rectangle and K_h be a partition of Ω into subrectangles. Then, if the Raviart-Thomas spaces on rectangles in Sect. 3.4.2.1 are used, it follows from the construction of these spaces that (3.45) splits into equations of the form

$$\begin{aligned} &\mathbf{A}_i\mathbf{U}_i + \mathbf{B}_i\mathbf{p} = \mathbf{g}_i, && i = 1, 2\ , \\ &\mathbf{D}\frac{d\mathbf{p}}{dt} - \mathbf{B}_1^T\mathbf{U}_1 - \mathbf{B}_2^T\mathbf{U}_2 = \mathbf{f}\ , \\ &\mathbf{p}(0) = \mathbf{p}^0\ , \end{aligned} \tag{3.47}$$

where the $\mathbf{U}_1$-parameters and $\mathbf{U}_2$-parameters are ordered in an x_1-orientation and an x_2-orientation, respectively, and the matrices $\mathbf{A}_i$ are block tridiagonal as well as symmetric and positive definite.

The Uzawa iterative algorithm is described as follows: Let $\mathbf{p}^0$ be given arbitrarily and determine $\mathbf{U}^0$ (only $\mathbf{U}_2^0$ needs to be computed to initiate the iteration) by the system

$$\mathbf{A}_i\mathbf{U}_i^0 + \mathbf{B}_i\mathbf{p}^0 = \mathbf{g}_i, \qquad i = 1, 2\ .$$

The general step splits into the following x_1-sweep and x_2-sweep:

$$\begin{aligned} &\mathbf{A}_1\mathbf{U}_1^{n+1/2} + \mathbf{B}_1\mathbf{p}^{n+1/2} = \mathbf{g}_1\ , \\ &\mathbf{D}\frac{\mathbf{p}^{n+1/2} - \mathbf{p}^n}{\Delta t^n} - \mathbf{B}_1^T\mathbf{U}_1^{n+1/2} - \mathbf{B}_2^T\mathbf{U}_2^n = \mathbf{f}\ , \\ &\mathbf{A}_2\mathbf{U}_2^{n+1/2} + \mathbf{B}_2\mathbf{p}^{n+1/2} = \mathbf{g}_2\ , \end{aligned} \tag{3.48}$$

and

$$\begin{aligned} &\mathbf{A}_2\mathbf{U}_2^{n+1} + \mathbf{B}_2\mathbf{p}^{n+1} = \mathbf{g}_2\ , \\ &\mathbf{D}\frac{\mathbf{p}^{n+1} - \mathbf{p}^{n+1/2}}{\Delta t^n} - \mathbf{B}_1^T\mathbf{U}_1^{n+1/2} - \mathbf{B}_2^T\mathbf{U}_2^{n+1} = \mathbf{f}\ , \\ &\mathbf{A}_1\mathbf{U}_1^{n+1} + \mathbf{B}_1\mathbf{p}^{n+1} = \mathbf{g}_1\ , \end{aligned} \tag{3.49}$$

where Δt^n is a sequence of parameters. Note that $\mathbf{U}_2^{n+1/2}$and $\mathbf{U}_1^{n+1}$ do not enter into the evolution; they need not be calculated at all, though it is

probably a good idea to compute them to be consistent with the final $\mathbf{p}$ upon termination of the iteration.

A spectral analysis for the iteration in (3.48) and (3.49) was given by Brown (1982). The analytical result shows that this iteration converges for any symmetric positive definite matrix $\mathbf{D}$ and constant sequence $\Delta t^n = \Delta t > 0$. Moreover, for $0 < C_1 < C_2$, there is n such that the error estimate holds (Brown, 1982):

$$\|\mathbf{u}_h^n - \mathbf{u}_h\| + \|p_h^n - p_h\| \leq \frac{1}{2}\left(\|\mathbf{u}_h^0 - \mathbf{u}_h\| + \|p_h^0 - p_h\|\right) , \tag{3.50}$$

for any virtual time steps such that $C_1 \leq \Delta t^1 \leq \Delta t^2 \leq \ldots \leq \Delta t^n \leq C_2$, where $\|\cdot\|$ indicates the L^2-norm. When Ω is a rectangle and the coefficient a in (3.46) is constant, then a time step cycle can be chosen as a geometric sequence with $n = \mathcal{O}(\log h^{-1})$ such that (3.50) holds (Douglas-Pietra, 1985). Then it follows that at most $\mathcal{O}((\log \epsilon^{-1})(\log h^{-1}))$ iterations are required to reduce the initial error (in the form measured by (3.50)) by a factor ϵ. For a variable coefficient a in (3.46), an alternating-direction iterator for a constant coefficient problem can be utilized as a preconditioner for the conjugate gradient algorithm. The same complexity bound can be obtained for the present iteration. Namely, no more than $\mathcal{O}((\log \epsilon^{-1})(\log h^{-1}))$ iterations are needed to reduce the error by a factor ϵ.

3.7.5 Mixed-Hybrid Algorithms

As mentioned, the constraint $\mathbf{V}_h \subset \mathbf{V}$ implies that the normal components of the functions in $\mathbf{V}_h$ are continuous across the interior boundaries in K_h (cf. Exercise 3.7). Following Arnold-Brezzi (1985), we relax this constraint on $\mathbf{V}_h$ by defining

$$\tilde{\mathbf{V}}_h = \left\{\mathbf{v} \in \left(L^2(\Omega)\right)^d : \mathbf{v}|_K \in \mathbf{V}_h(K) \text{ for each } K \in K_h\right\}, \quad d = 2 \text{ or } 3 .$$

We then need to introduce *Lagrange multipliers* to enforce the required continuity on $\tilde{\mathbf{V}}_h$, so we define

$$L_h = \left\{\mu \in L^2\left(\bigcup_{e \in \mathcal{E}_h} e\right) : \mu|_e \in \mathbf{V}_h \cdot \boldsymbol{\nu}|_e \text{ for each } e \in \mathcal{E}_h\right\} ,$$

where $\mathcal{E}_h$ indicates the set of all edges or faces in K_h. Now, the hybrid form of the mixed method (3.17) is

Find $(\mathbf{u}_h, p_h, \lambda_h) \in \tilde{\mathbf{V}}_h \times W_h \times L_h$ such that

$$\begin{aligned}
&(\mathbf{u}_h, \mathbf{v}) - \sum_{K \in K_h} \{(\nabla \cdot \mathbf{v}, p_h)_K - (\mathbf{v} \cdot \boldsymbol{\nu}_K, \lambda_h)_{\partial K \setminus \Gamma}\} = 0, \qquad && \mathbf{v} \in \tilde{\mathbf{V}}_h \ , \\
&\sum_{K \in K_h} (\nabla \cdot \mathbf{u}_h, w)_K = (f, w), && w \in W_h \ , \\
&\sum_{K \in K_h} (\mathbf{u}_h \cdot \boldsymbol{\nu}_K, \mu)_{\partial K \setminus \Gamma} = 0, && \mu \in L_h \ ,
\end{aligned} \tag{3.51}$$

where $\boldsymbol{\nu}_K$ denotes the outward unit normal to K. Note that the third equation of (3.51) enforces the continuity requirement on $\mathbf{u}_h$, so in fact $\mathbf{u}_h \in \mathbf{V}_h$.

As in Sect. 3.2, after the introduction of basis functions in $\tilde{\mathbf{V}}_h$, W_h, and L_h, (3.51) can be expressed in the matrix form

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} & \mathbf{C} \\ \mathbf{B}^T & \mathbf{0} & \mathbf{0} \\ \mathbf{C}^T & \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{U} \\ \mathbf{p} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ -\mathbf{f} \\ \mathbf{0} \end{pmatrix} \ , \tag{3.52}$$

where $\boldsymbol{\lambda}$ is the degrees of freedom of λ_h. The advantage of system (3.52) is that the matrix $\mathbf{A}$ is block-diagonal, with each block corresponding to a single element. Hence $\mathbf{A}$ is easily inverted at the element level. This, together with the first equation in (3.52), leads to

$$\mathbf{U} = -\mathbf{A}^{-1}\mathbf{B}\mathbf{p} - \mathbf{A}^{-1}\mathbf{C}\boldsymbol{\lambda} \ . \tag{3.53}$$

Substituting it into the second and third equations in (3.52), we see that

$$\begin{aligned}
\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\mathbf{p} + \mathbf{B}^T\mathbf{A}^{-1}\mathbf{C}\boldsymbol{\lambda} &= \mathbf{f} \ , \\
\mathbf{C}^T\mathbf{A}^{-1}\mathbf{B}\mathbf{p} + \mathbf{C}^T\mathbf{A}^{-1}\mathbf{C}\boldsymbol{\lambda} &= \mathbf{0} \ .
\end{aligned} \tag{3.54}$$

By (3.32), $\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}$ is symmetric and positive definite, so the first equation of (3.54) yields

$$\mathbf{p} = \left(\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{f} - \left(\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{A}^{-1}\mathbf{C}\boldsymbol{\lambda} \ . \tag{3.55}$$

Substituting this equation into the second equation of (3.54) implies the linear system for $\boldsymbol{\lambda}$

$$\begin{aligned}
\Big(\mathbf{C}^T\mathbf{A}^{-1}\mathbf{C} - \big(\mathbf{C}^T\mathbf{A}^{-1}\mathbf{B}\big)\big(\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\big)^{-1}\big(\mathbf{B}^T\mathbf{A}^{-1}\mathbf{C}\big)\Big)\boldsymbol{\lambda} \\
= -\big(\mathbf{C}^T\mathbf{A}^{-1}\mathbf{B}\big)\big(\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\big)^{-1}\mathbf{f} \ .
\end{aligned} \tag{3.56}$$

This system for $\boldsymbol{\lambda}$ is symmetric, positive definite, and sparse. Therefore, we can solve (3.56) for $\boldsymbol{\lambda}$, and then recover $\mathbf{p}$ via (3.55) and $\mathbf{U}$ via (3.53). System (3.56) can be solved via the iterative algorithms developed in Sect. 1.10.2, for example.

3.7.6 An Equivalence Relationship

When the lowest-order Raviart-Thomas mixed space on triangles (respectively, rectangles) is applied to the discretization of problem (3.13), it is interesting to see that system (3.56) is the same as that generated by the triangular P_1 nonconforming finite element method (respectively, rotated Q_1 nonconforming finite element method) introduced in the preceding chapter (Chen, 1996). As an example, we examine the lowest-order Raviart-Thomas mixed space on triangles (cf. Sect. 3.4.1.1) for the solution of the model problem in two dimensions:

$$\begin{aligned} -\nabla \cdot (\mathbf{a}\nabla p) &= f \qquad \text{in } \Omega\ , \\ p &= 0 \qquad \text{on } \Gamma\ , \end{aligned} \tag{3.57}$$

where $\mathbf{a}$ and f are given as in (3.26). For this mixed space, the spaces $\tilde{\mathbf{V}}_h$, W_h, and L_h are specifically given by

$$\begin{aligned} \tilde{\mathbf{V}}_h &= \left\{ \mathbf{v} \in \left(L^2(\Omega)\right)^2 : \mathbf{v}|_K \in \left(P_0(K)\right)^2 \oplus \left((x_1, x_2)P_0(K)\right),\ K \in K_h \right\}\ , \\ W_h &= \left\{ w \in L^2(\Omega) : w|_K \in P_0(K),\ K \in K_h \right\}\ , \\ L_h &= \left\{ \mu \in L^2(\mathcal{E}_h) : \mu|_e \in P_0(e),\ e \in \mathcal{E}_h \right\}\ , \end{aligned}$$

where $P_0(K)$ is the space of constants defined on K and $\mathcal{E}_h$ is the set of all edges in K_h.

Let P_h be the L^2-projection onto W_h: For $v \in L^2(\Omega)$, $P_h v \in W_h$ satisfies

$$(P_h v - v, w) = 0 \qquad \forall w \in W_h\ .$$

Set $\mathcal{K}_h = P_h \mathbf{a}^{-1}$ (componentwise). Then a modified hybrid form of the mixed method for (3.57) (cf. (3.51)) is

$$\begin{aligned} &\text{Find } (\mathbf{u}_h, p_h, \lambda_h) \in \tilde{\mathbf{V}}_h \times W_h \times L_h \text{ such that} \\ &(\mathcal{K}_h \mathbf{u}_h, \mathbf{v}) - \sum_{K \in K_h} \{(\nabla \cdot \mathbf{v}, p_h)_K - (\mathbf{v} \cdot \boldsymbol{\nu}_K, \lambda_h)_{\partial K \setminus \Gamma}\} = 0, \qquad \mathbf{v} \in \tilde{\mathbf{V}}_h\ , \\ &\sum_{K \in K_h} (\nabla \cdot \mathbf{u}_h, w)_K = (f, w), \qquad w \in W_h\ , \\ &\sum_{K \in K_h} (\mathbf{u}_h \cdot \boldsymbol{\nu}_K, \mu)_{\partial K \setminus \Gamma} = 0, \qquad \mu \in L_h\ . \end{aligned} \tag{3.58}$$

For each K in K_h, set

$$\bar{f}_K = \frac{1}{|K|}(f, 1)_K = \frac{1}{|K|} \int_K f \, d\mathbf{x}\ ,$$

where $|K|$ denotes the area of K. Also, set $\mathcal{K}_h = (\alpha_{ij})$ and $\mathbf{u}_h|_K = (u_{K1}, u_{K2}) = (a_K^1 + b_K x_1, a_K^2 + b_K x_2)$. Then it follows from the second equation of (3.58) that

$$b_K = \frac{\bar{f}_K}{2}. \tag{3.59}$$

Next, take $\mathbf{v} = (1,0)$ in K and $\mathbf{v} = \mathbf{0}$ elsewhere, and take $\mathbf{v} = (0,1)$ in K and $\mathbf{v} = \mathbf{0}$ elsewhere, respectively, in the first equation of (3.58) to obtain

$$\sum_{i=1}^{2} (\alpha_{ji} u_{Ki}, 1)_K + \sum_{i=1}^{3} |e_K^i| \nu_{Kj}^i \lambda_h|_{e_K^i} = 0, \quad j = 1,2\,, \tag{3.60}$$

where $|e_K^i|$ is the length of the edge e_K^i of K and $\boldsymbol{\nu}_K^i = (\nu_{K1}^i, \nu_{K2}^i)$ is the outward unit normal to e_K^i, $i = 1,2$. Letting $\boldsymbol{\beta}_K = (\beta_{ij}^K) = ((\alpha_{ij},1)_K)^{-1}$, (3.60) can be then used to solve for the coefficients a_K^1 and a_K^2:

$$\begin{aligned} a_K^j = &- \sum_{i=1}^{3} |e_K^i| \left(\beta_{j1}^K \nu_{K1}^i + \beta_{j2}^K \nu_{K2}^i\right) \lambda_h|_{e_K^i} \\ &- \frac{\bar{f}_K}{2} \sum_{i=1}^{2} \left(\beta_{ji}^K, \alpha_{i1} x_1 + \alpha_{i2} x_2\right)_K, \quad j = 1,2\,. \end{aligned} \tag{3.61}$$

Let the basis in L_h be chosen as usual. Namely, take $\mu = 1$ on one edge and $\mu = 0$ elsewhere in the third equation of (3.58). Then, apply (3.59) and (3.61) to see that the contributions of the triangle K to the stiffness matrix $\mathbf{A}$ and the right-hand side $\mathbf{f}$ are

$$a_{ij}^K = \bar{\boldsymbol{\nu}}_K^i \boldsymbol{\beta}_K \bar{\boldsymbol{\nu}}_K^j, \qquad f_{Ki} = -\frac{(\mathbf{J}_K^f, \bar{\boldsymbol{\nu}}_K^i)_K}{|K|} + (\mathbf{J}_K^f, \boldsymbol{\nu}_K^i)_{e_K^i}\,,$$

where $\bar{\boldsymbol{\nu}}_K^i = |e_E^i| \boldsymbol{\nu}_E^i$ and $\mathbf{J}_K^f = \bar{f}_K(x_1, x_2)/2$. Hence we obtain the following system for λ_h by the mixed-hybrid algorithm:

$$\mathbf{A}\boldsymbol{\lambda} = \mathbf{f}\,, \tag{3.62}$$

where $\mathbf{A} = (a_{ij})$, $\boldsymbol{\lambda}$ is the degrees of freedom of λ_h, and $\mathbf{f} = (f_i)$.

We now consider the nonconforming finite element method (2.3) for (3.57). Let V_h be the nonconforming P_1 finite element space as defined in Sect. 2.1.1:

$$\begin{aligned} V_h = \{ v \in L^2(\Omega) : {} & v|_K \text{ is linear, } K \in K_h; v \text{ is continuous} \\ & \text{at the midpoints of interior edges and} \\ & \text{is zero at the midpoints of edges on } \Gamma \}\,. \end{aligned}$$

We modify (2.3) as follows: Find $p_h \in V_h$ such that

$$a_h(p_h, v) = (P_h f, v) \qquad \forall v \in V_h\,, \tag{3.63}$$

where

$$a_h(p_h, v) = \sum_{K \in K_h} \left(\mathcal{K}_h^{-1} \nabla p_h, \nabla v\right)_K \ .$$

That is, the L^2-projection is used in the discrete bilinear form $a_h(\cdot, \cdot)$ and the right-hand side of (3.63). The quantity $\mathcal{K}_h^{-1}$ is called the *harmonical average* of $\mathbf{a}$. Let $\{\varphi_i\}$ be the basis of V_h as defined in Sect. 2.1.1. Associated with an edge $e_K^i \in \partial K$, we have

$$\varphi_i|_K = \frac{1}{|K|} \bar{\boldsymbol{\nu}}_K^i \cdot ((x_1, x_2) - \mathbf{m}_l), \qquad i \neq l \ ,$$

for some midpoint $\mathbf{m}_l$. It can be checked that (Chen, 1996)

$$\left(\mathcal{K}_h^{-1} \nabla \varphi_i, \nabla \varphi_j\right)_K = \bar{\boldsymbol{\nu}}_K^i \boldsymbol{\beta}_K \bar{\boldsymbol{\nu}}_K^j \ ,$$

which is a_{ij}^K. Also, it can be shown that

$$f_{Ki} = \bar{f}_K (1, \varphi_i)_K = (P_h f, \varphi_i)_K \ .$$

Therefore, method (3.63) also leads to the system of algebraic equations (3.62). Namely, methods (3.58) and (3.63) produce the same system.

For a differential problem more general than (3.57) and other mixed finite element spaces, there is also an equivalence relationship between these spaces and certain nonconforming finite element spaces (Arnold-Brezzi, 1985; Chen, 1993A; Arbogast-Chen, 1995). This equivalence relationship is useful in the development of iterative algorithms for solving linear systems arising from the mixed method (Chen, 1996; Chen et al., 1996).

We end with mentioning that when $\mathbf{V}_h \times W_h$ are the lowest-order RTN spaces over rectangular parallelepipeds (cf. Sects. 3.4.2.1 and 3.4.4.1), it can be shown that the linear system arising from the mixed method can be written as a system generated by a *cell-centered* (or *block-centered*) finite difference scheme using certain quadrature rules (Russell-Wheeler, 1983).

3.8 Theoretical Considerations

In this section, we give an abstract formulation of the mixed finite element method for second-order partial differential equations. The reader who is not interested in the theory may bypass this section.

3.8.1 An Abstract Formulation

Suppose that V and W are two Hilbert spaces, and $a(\cdot, \cdot) : V \times V \to \mathbb{R}$ and $b(\cdot, \cdot) : V \times W \to \mathbb{R}$ are two *bilinear forms* (cf. Sect. 1.3.1). Also, let $L : V \to \mathbb{R}$ and $Q : W \to \mathbb{R}$ be two *linear functionals*. We consider the problem:

$$\begin{aligned} &\text{Find } u \in V \text{ and } p \in W \text{ such that} \\ &a(u,v) + b(v,p) = L(v) \qquad \forall v \in V\ , \\ &b(u,w) = Q(w) \qquad \forall w \in W\ . \end{aligned} \tag{3.64}$$

Introduce the functional $F : V \times W \to \mathbb{R}$

$$F(v,w) = \frac{1}{2}a(v,v) + b(v,w) - L(v) - Q(w), \qquad v \in V,\ w \in W\ .$$

Then, using the same argument as for (3.4) and (3.5), it can be seen that if, for example, $a(\cdot,\cdot)$ is symmetric and V-elliptic (cf. (1.39) and (1.41)), (3.64) is equivalent to the *saddle point problem*:

$$\begin{aligned} &\text{Find } u \in V \text{ and } p \in W \text{ such that} \\ &F(u,w) \le F(u,p) \le F(v,p) \qquad \forall v \in V,\ w \in W\ . \end{aligned}$$

To study (3.64), we need some assumptions on the bilinear forms a and b. It is natural to assume that they are continuous:

$$\begin{aligned} &|a(v_1,v_2)| \le a^* \|v_1\|_V \|v_2\|_V \qquad \forall v_1,\ v_2 \in V\ , \\ &|b(v,w)| \le b^* \|v\|_V \|w\|_W \qquad \forall v \in V,\ w \in W\ . \end{aligned} \tag{3.65}$$

Also, we define the linear spaces

$$\begin{aligned} Z(Q) &= \{v \in V : b(v,w) = Q(w) \quad \forall w \in W\}\ , \\ Z &= Z(0) = \{v \in V : b(v,w) = 0 \quad \forall w \in W\}\ . \end{aligned}$$

Because b is continuous, Z is a closed subspace of V. We now write (3.64) in terms of proper operators. Denote by $\langle \cdot,\cdot \rangle_{V' \times V}$ the duality pairing between V' and V, where V' is the *dual space* of V (i.e., the set of bounded linear functionals on V; cf. Sect. 1.2.5). With a, we associate the operator $A : V \to V'$ defined by

$$\langle Au, v \rangle_{V' \times V} = a(u,v) \qquad \forall v \in V\ .$$

Next, we define the operators $B : W \to V'$ and $B' : V \to W'$ by

$$\begin{aligned} &\langle Bp, v \rangle_{V' \times V} = b(v,p) \qquad \forall v \in V\ , \\ &\langle B'u, w \rangle_{W' \times W} = b(u,w) \qquad \forall w \in W\ . \end{aligned}$$

With these operators, (3.64) is equivalent to

$$\begin{aligned} &Au + Bp = L\ , \\ &B'u = Q\ . \end{aligned} \tag{3.66}$$

Let $Z^\perp$ indicate the *orthogonal complement* of Z in V; i.e.,

$$Z^\perp = \{v \in V : (v, z)_V = 0 \quad \forall z \in Z\} ,$$

where $(\cdot, \cdot)_V$ is the inner product of V. Next, let Z^0 be the *polar set* of Z:

$$Z^0 = \{l \in V' : \langle l, z\rangle_{V' \times V} = 0 \quad \forall z \in Z\} .$$

Theorem (Closed Range Theorem). *Let $\mathcal{X}$ and $\mathcal{Y}$ be two Banach spaces with their respective dual spaces $\mathcal{X}'$ and $\mathcal{Y}'$, and $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$ be a bounded linear operator. Then the following two statements are equivalent*:

(i) *The range $\mathcal{F}(\mathcal{X})$ is closed in $\mathcal{Y}$.*
(ii) $\mathcal{F}(\mathcal{X}) = (\ker(\mathcal{F}'))^0$, *where* $\ker(\mathcal{F}')$ *indicates the kernel of $\mathcal{F}'$ (the adjoint of $\mathcal{F}$); i.e.,*

$$\ker(\mathcal{F}') = \{\mathrm{y} \in \mathcal{Y}' : \mathcal{F}'(\mathrm{y}) = 0\} .$$

This theorem can be found in Yosida (1971), for example. A linear mapping between two normed linear spaces is an *isomorphism* if it is bijective and, together with its inverse, is bounded.

Lemma 3.1. *The following three statements are equivalent*:

(i) *There is a constant $b_* > 0$ such that*

$$\inf_{w \in W} \sup_{v \in V} \frac{b(v, w)}{\|v\|_V \|w\|_W} \geq b_* . \tag{3.67}$$

(ii) *The operator $B : W \to Z^0 \subset V'$ is an isomorphism. Moreover,*

$$\|Bw\|_{V'} \geq b_* \|w\|_W \qquad \forall w \in W . \tag{3.68}$$

(iii) *The operator $B' : Z^\perp \to W'$ is an isomorphism. Furthermore,*

$$\|B'v\|_{W'} \geq b_* \|v\|_V \qquad \forall v \in Z^\perp . \tag{3.69}$$

Proof. As an example, we only prove the equivalence between (i) and (ii). The others can be shown similarly.

Under (i), we see that B is one-to-one. Let $l \in B(W)$, the range of B; then there exists $w \in W$ such that $w = B^{-1}l$. It follows from (i) that

$$b_* \|w\|_W \leq \sup_{v \in V} \frac{b(v, w)}{\|v\|_V} = \sup_{v \in V} \frac{\langle l, v\rangle_{V' \times V}}{\|v\|_V} = \|l\|_{V'} . \tag{3.70}$$

Namely, (3.68) holds. Moreover, B^{-1} is continuous on $B(W)$. By the continuity of B and B^{-1}, we see that $B(W)$ is closed. Then it follows from the Closed Range Theorem that $B(W) = Z^0$. Hence (ii) is proven.

Now, suppose that (ii) is true. Then, for any $w \in W$, there exists $l \in Z^0 \subset V'$ such that $Bw = l$. Consequently, (3.67) follows from (3.68) and (3.70). □

Inequality (3.67) is termed the *inf-sup condition*. Note that (3.64) induces a linear operator $\mathcal{L} : V \times W \to V' \times W'$ through

$$(u, p) \longmapsto (L, Q) \ . \tag{3.71}$$

Theorem (Hahn-Banach Theorem). *If $\mathcal{M}$ is a subspace of a real normed space $\mathcal{X}$ and $\mathcal{F}_0$ is a continuous linear functional defined on $\mathcal{M}$ with norm $\|\mathcal{F}_0\|_{\mathcal{M}'}$, then there is a continuous linear extension $\mathcal{F}$ of $\mathcal{F}_0$ to $\mathcal{X}$ such that $\|\mathcal{F}\|_{\mathcal{X}'} = \|\mathcal{F}_0\|_{\mathcal{M}'}$.*

This theorem can be found in Conway (1985), for example.

Theorem 3.2. *For problem* (3.64), *the operator $\mathcal{L} : V \times W \to V' \times W'$ is an isomorphism if and only if the following two conditions hold:*

(i) *the bilinear form a is Z-elliptic; i.e., there exists $a_* > 0$ such that*

$$a(v, v) \geq a_* \|v\|_V^2 \qquad \forall v \in Z \ , \tag{3.72}$$

(ii) *and the bilinear form b satisfies* (3.67).

Proof. Let $\mathcal{L}$ be an isomorphism. Especially, $\mathcal{L}^{-1}$ is bounded. It follows from the Hahn-Banach Theorem that every functional $L \in Z'$ has an extension $\check{L} \in V'$ such that $\|L\|_{Z'} = \|\check{L}\|_{V'}$. Define $(u, p) = \mathcal{L}^{-1}(\check{L}, 0)$. Then $u \in Z$ is a minimum of $a(v, v)/2 - L(v), v \in Z$. The operator $L \longmapsto u \in Z$ is bounded, so the bilinear form a is Z-elliptic.

Also, let $Q \in W'$, and define $(u, p) = \mathcal{L}^{-1}(0, Q)$ such that $\|u\|_V \leq C\|Q\|_{W'}$ for some positive constant C. Let $u_0 \in Z^\perp$ be the projection of u. Because $\|u_0\|_V \leq \|u\|_V$, the operator $Q \longmapsto u \longmapsto u_0$ is bounded. Moreover, $B'u_0 = Q$. Consequently, $B' : Z^\perp \to W'$ is an isomorphism. Thus, by (iii) in Lemma 3.1, we see that the bilinear form b satisfies (3.67).

Conversely, suppose that a and b satisfy (3.72) and (3.67), respectively. Let $(L, Q) \in V' \times W'$. First, it follows from (iii) in Lemma 3.1 that there is $u_1 \in Z^\perp$ such that $B'u_1 = Q$ and $\|u_1\|_V \leq \|Q\|_{W'}/b_*$.

Next, set $\lambda = u - u_1$. Then (3.64) is equivalent to

$$\begin{aligned} &a(\lambda, v) + b(v, p) = L(v) - a(u_1, v) && \forall v \in V \ , \\ &b(\lambda, w) = 0 && \forall w \in W \ . \end{aligned} \tag{3.73}$$

Thus it suffices to prove (3.73). First, using (3.72), the functional

$$\frac{1}{2} a(v, v) - L(v) + a(u_1, v)$$

attains its minimum for some $\lambda \in Z$ such that

$$\|\lambda\|_V \leq \frac{1}{a_*} \left(\|L\|_{V'} + C\|u_1\|_V \right) \ ,$$

for some positive constant C. Namely, $\lambda \in Z$ satisfies

$$a(\lambda, v) = L(v) - a(u_1, v) \qquad \forall v \in Z \ . \tag{3.74}$$

Second, if there exists $p \in W$ such that

$$b(v, p) = L(v) - a(u_1 + \lambda, v) \qquad \forall v \in V \ , \tag{3.75}$$

then (3.73) will follow. Note that the right-hand side of (3.75) defines a functional in V', and this functional is in Z^0 by (3.74). Hence, applying (ii) in Lemma 3.1, this functional can be expressed as Bp with

$$\|p\|_W \le \frac{1}{b_*} \left(\|L\|_{V'} + C\|u\|_V \right) \ .$$

Thus the solvability of (3.73) is shown. Uniqueness follows from the two conditions (i) and (ii) on a and b. Therefore, $\mathcal{L}$ is surjective and injective. Furthermore, applying the above bounds on u_1, λ, and p, we see that

$$\begin{aligned} \|u\|_V &\le \frac{\|L\|_{V'}}{a_*} + \left(1 + \frac{C}{a_*}\right) \frac{\|Q\|_{W'}}{b_*} \ , \\ \|p\|_W &\le \left(1 + \frac{C}{a_*}\right) \left(\frac{\|L\|_{V'}}{b_*} + \frac{C\|Q\|_{W'}}{b_*^2} \right) \ , \end{aligned} \tag{3.76}$$

which implies that the inverse operator $\mathcal{L}^{-1}$ is continuous. Hence $\mathcal{L}$ is an isomorphism. □

We remark that (3.72) is required to hold in the space Z instead of V. This is the usual case in most applications.

3.8.2 The Mixed Finite Element Method

Suppose that V_h and W_h are the respective finite dimensional subspaces of V and W. Then the discrete counterpart of (3.64) is

$$\begin{aligned} &\text{Find } u_h \in V_h \text{ and } p_h \in W_h \text{ such that} \\ &a(u_h, v) + b(v, p_h) = L(v) && \forall v \in V_h \ , \\ &b(u_h, w) = Q(w) && \forall w \in W_h \ . \end{aligned} \tag{3.77}$$

In view of $Z(Q)$ and Z, we also define their discrete counterparts

$$\begin{aligned} Z_h(Q) &= \{ v \in V_h : b(v, w) = Q(w) \quad \forall w \in W_h \} \ , \\ Z_h &= Z_h(0) = \{ v \in V_h : b(v, w) = 0 \quad \forall w \in W_h \} \ . \end{aligned}$$

Lemma 3.3. *Let the bilinear form b satisfy*

$$\sup_{v \in V_h} \frac{b(v, w)}{\|v\|_V} \ge b_* \|w\|_W \qquad \forall w \in W_h \ , \tag{3.78}$$

where the constant $b_* > 0$ is independent of h. Then there is a constant C, independent of h, such that for every $p \in Z(Q)$,

$$\inf_{v \in Z_h(Q)} \|p - v\|_V \le C \inf_{w \in V_h} \|p - v\|_V \ .$$

Proof. The minimization of $\|p - v\|_V$ subject to the constraint $v \in Z_h(Q)$ implies

$$\begin{aligned} (v, y)_V + b(y, q) &= (p, y)_V \qquad && \forall y \in V_h \ , \\ b(v, z) &= Q(z) && \forall z \in W_h \ , \end{aligned}$$

for some $q \in W_h$. Then, for any $w \in V_h$,

$$\begin{aligned} (v - w, y)_V + b(y, q) &= (p - w, y)_V \qquad && \forall y \in V_h \ , \\ b(v - w, z) &= b(p - w, z) && \forall z \in W_h \ . \end{aligned}$$

The functionals on the right-hand side of this system are bounded by $C\|p - w\|_V$. Thus, in the same way as for (3.76), we see that $\|v - w\|_V \le C\|p - w\|_V$, which, together with the triangle inequality, implies the desired result. □

Theorem 3.4. *In addition to the assumptions of Theorem* 3.2, *if there are constants $a_* > 0$ and $b_* > 0$, independent of h, such that*

(i) *the bilinear form a is Z_h-elliptic:*

$$a(v, v) \ge a_* \|v\|_V^2 \qquad \forall v \in Z_h \ , \tag{3.79}$$

(ii) *and the bilinear form b satisfies* (3.78), *then* (3.77) *has a unique solution $u_h \in V_h$ and $p_h \in W_h$. Moreover,*

$$\begin{aligned} &\|u - u_h\|_V + \|p - p_h\|_W \\ &\qquad \le C \left(\inf_{v \in V_h} \|u - v\|_V + \inf_{w \in W_h} \|p - w\|_W \right) \ . \end{aligned} \tag{3.80}$$

Proof. Existence and uniqueness of (3.77) can be shown as in Theorem 3.2. It suffices to prove (3.80). Subtracting (3.77) from (3.64), we see that

$$\begin{aligned} a(u - u_h, v) + b(v, p - p_h) &= 0 \qquad && \forall v \in V_h \ , \\ b(u - u_h, w) &= 0 && \forall w \in W_h \ . \end{aligned} \tag{3.81}$$

Let $v \in Z_h(Q)$. Since $u_h - v \in Z_h$, it follows from (3.65), (3.79), and (3.81) that, with $w \in W_h$,

$$\begin{aligned} a_* \|u_h - v\|_V^2 &\le a(u_h - v, u_h - v) \\ &= a(u_h - u, u_h - v) + a(u - v, u_h - v) \\ &= b(u_h - v, p - p_h) + a(u - v, u_h - v) \\ &= b(u_h - v, p - w) + a(u - v, u_h - v) \\ &\le C \left(\|p - w\|_W + \|u - v\|_V \right) \|u_h - v\|_V \ . \end{aligned} \tag{3.82}$$

Next, for $w \in W_h$, by (3.81) we get

$$b(v, w - p_h) = -a(u - u_h, v) - b(v, p - w) \qquad \forall v \in V_h \ ,$$

so that, using (3.65) and (3.78),

$$\begin{aligned} b_* \|w - p_h\|_W &\le \sup_{v \in V_h} \frac{b(v, w - p_h)}{\|v\|_V} \\ &= \sup_{v \in V_h} \frac{-a(u - u_h, v) - b(v, p - w)}{\|v\|_V} \\ &\le C \left(\|u - u_h\|_V + \|p - w\|_W \right) \ . \end{aligned} \tag{3.83}$$

Combine Lemma 3.3, (3.82), and (3.83) to obtain (3.80). □

In general, $Z_h \not\subset Z$. If $Z_h \subset Z$, a better estimate can be obtained, as shown below.

Theorem 3.5. *In addition to the assumptions of Theorem* 3.2, *if* $Z_h \subset Z$, *then*

$$\|u - u_h\|_V \le C \inf_{v \in V_h} \|u - v\|_V \ . \tag{3.84}$$

Proof. Let $\lambda \in Z_h(Q)$. Then, for $v \in Z_h$ we see that, by $Z_h \subset Z$ and (3.81),

$$\begin{aligned} a(u_h - \lambda, v) &= a(u_h - u, v) + a(u - \lambda, v) \\ &= b(v, p - p_h) + a(u - \lambda, v) \\ &= a(u - \lambda, v) \le C \|u - \lambda\|_V \|v\|_V \ . \end{aligned}$$

Taking $v = u_h - \lambda$ leads to the desired result. □

Inequalities (3.79) and (3.78) are referred to as the *Babuška-Brezzi condition* or sometimes the *Ladyshenskaja-Babuška-Brezzi condition*. Often condition (3.78) alone is termed the Ladyshenskaja-Babuška-Brezzi condition, as mentioned earlier. It is sometimes called the discrete *inf-sup condition*. The following result is useful in the verification of this condition (Fortin, 1977).

Theorem 3.6. *Assume that the bilinear form* b *satisfies* (3.67). *If there exists a bounded projection operator* $\Pi_h : V \to V_h$ *such that*

$$b(v - \Pi_h v, w) = 0 \qquad \forall w \in W_h \ ,$$

and the bound is independent of h, *then the discrete inf-sup condition* (3.78) *holds.*

Proof. From (3.67), it follows that, for any $w \in W_h$,

$$\begin{aligned} b_* \|w\|_W &\le \sup_{v \in V} \frac{b(v, w)}{\|v\|_V} = \sup_{v \in V} \frac{b(\Pi_h v, w)}{\|v\|_V} \\ &\le C \sup_{v \in V} \frac{b(\Pi_h v, w)}{\|\Pi_h v\|_V} \le C \sup_{v \in V_h} \frac{b(v, w)}{\|v\|_V} \ . \end{aligned}$$

This implies the desired result. □

3.8.3 Examples

As noted, in this chapter we concentrate on applications of the mixed finite element method to second-order partial differential equations. Other applications will be presented in Chaps. 7–10.

We consider the model problem

$$\begin{aligned} -\nabla \cdot (\mathbf{a}\nabla p) &= f \qquad \text{in } \Omega\,, \\ p &= 0 \qquad \text{on } \Gamma\,, \end{aligned} \tag{3.85}$$

where $\mathbf{a}$ is a $d \times d$ ($d = 2$ or 3) matrix and $f \in L^2(\Omega)$ is given. Assume that $\mathbf{a}$ satisfies (3.27). This problem was considered in Sect. 3.4. The following spaces have been introduced in Sect. 3.2:

$$\mathbf{V} = \mathbf{H}(\mathrm{div}, \Omega), \quad W = L^2(\Omega)\,.$$

We recall the inner product of $\mathbf{V}$:

$$(\mathbf{v}_1, \mathbf{v}_2) = \int_\Omega \mathbf{v}_1 \cdot \mathbf{v}_2 \; d\mathbf{x} + \int_\Omega \nabla \cdot \mathbf{v}_1 \nabla \cdot \mathbf{v}_2 \; d\mathbf{x}, \qquad \mathbf{v}_1,\ \mathbf{v}_2 \in \mathbf{V}\,.$$

Set

$$\mathbf{u} = -\mathbf{a}\nabla p\,. \tag{3.86}$$

Equation (3.85) is then written as

$$\nabla \cdot \mathbf{u} = f\,. \tag{3.87}$$

As for (3.29), problem (3.85) can be recast as follows:

$$\begin{aligned} &\text{Find } \mathbf{u} \in \mathbf{V} \text{ and } p \in W \text{ such that} \\ &(\mathbf{a}^{-1}\mathbf{u}, \mathbf{v}) - (\nabla \cdot \mathbf{v}, p) = 0 \qquad \forall \mathbf{v} \in \mathbf{V}\,, \\ &(\nabla \cdot \mathbf{u}, w) = (f, w) \qquad \forall w \in W\,. \end{aligned} \tag{3.88}$$

Define

$$\begin{aligned} a(\mathbf{u}, \mathbf{v}) &= (\mathbf{a}^{-1}\mathbf{u}, \mathbf{v}), \qquad \mathbf{u},\ \mathbf{v} \in \mathbf{V}\,, \\ b(\mathbf{v}, w) &= -(\nabla \cdot \mathbf{v}, w), \qquad \mathbf{v} \in \mathbf{V},\ w \in W\,. \end{aligned}$$

Then (3.88) is of form (3.64) with

$$L(\mathbf{v}) = 0, \quad \mathbf{v} \in \mathbf{V}, \quad Q(w) = -(f, w), \quad w \in W\,.$$

Obviously, the bilinear forms a and b satisfy the continuity condition (3.65). Also, for any $\mathbf{v} \in Z$ we see that $\nabla \cdot \mathbf{v} = 0$, so

$$a(\mathbf{v}, \mathbf{v}) = \|\mathbf{a}^{-1/2}\mathbf{v}\|^2_{\mathbf{L}^2(\Omega)}$$

is Z-elliptic. Next, for $w \in L^2(\Omega)$ there is $v \in C_0^\infty(\Omega)$ such that

$$\|w - v\|_{L^2(\Omega)} \leq \frac{1}{2}\|w\|_{L^2(\Omega)} \ .$$

Define $y = \inf\{x_1 : \mathbf{x} = (x_1, x_2, \ldots, x_d) \in \Omega\}$ and

$$v_1(\mathbf{x}) = \int_y^{x_1} v(\tau, x_2, \ldots, x_d)\ d\tau, \quad v_i = 0, \quad i = 2, \ldots, d \ .$$

It is clear that $\nabla \cdot \mathbf{v} = v$, where $\mathbf{v} = (v_1, v_2, \ldots, v_d)$, and that

$$\|\mathbf{v}\|_{\mathbf{L}^2(\Omega)} \leq C\|v\|_{L^2(\Omega)} \ .$$

Consequently, we see that

$$\frac{b(\mathbf{v}, w)}{\|\mathbf{v}\|_{\mathbf{V}}} \geq \frac{(v, w)}{(1 + C)\|v\|_{L^2(\Omega)}} \geq \frac{1}{2(1 + C)}\|w\|_{L^2(\Omega)} \ ;$$

i.e., b satisfies (3.67). Thus the conditions in Theorem 3.2 are satisfied.

Let $\mathbf{V}_h \subset \mathbf{V}$ and $W_h \subset W$ be the RTN, BDM, BDDF, BDFM, or CD spaces introduced in Sect. 3.4. All these spaces possess the property

$$\nabla \cdot \mathbf{V}_h = W_h \ . \tag{3.89}$$

The discrete version of (3.88) reads as follows:

$$\begin{array}{ll} \text{Find } \mathbf{u}_h \in \mathbf{V}_h \text{ and } p_h \in W_h \text{ such that} & \\ (\mathbf{a}^{-1}\mathbf{u}_h, \mathbf{v}) - (\nabla \cdot \mathbf{v}, p_h) = 0 & \forall \mathbf{v} \in \mathbf{V}_h \ , \\ (\nabla \cdot \mathbf{u}_h, w) = (f, w) & \forall w \in W_h \ . \end{array} \tag{3.90}$$

As in the continuous case, using (3.89), it can be shown that a is Z_h-elliptic; i.e., (3.79) is satisfied. As for (3.78), we note that each of the mixed finite element spaces possesses a projection operator $\mathbf{\Pi}_h : (H^1(\Omega))^d \to \mathbf{V}_h$ which satisfies the conditions in Theorem 3.6 (see the next subsection). Thus Theorems 3.4 and 3.6 can be applied.

3.8.4 Construction of Projection Operators

Each of the RTN, BDM, BDDF, BDFM, and CD spaces possesses the property that there are projection operators $\mathbf{\Pi}_h : (H^1(\Omega))^d \to \mathbf{V}_h$ and $P_h : W \to W_h$ such that

$$\begin{array}{ll} \big(\nabla \cdot (\mathbf{v} - \mathbf{\Pi}_h \mathbf{v}), w\big) = 0 & \forall w \in W_h \ , \\ (\nabla \cdot \mathbf{y}, z - P_h z) = 0 & \forall \mathbf{y} \in \mathbf{V}_h \ . \end{array} \tag{3.91}$$

That is, on $(H^1(\Omega))^d \cap \mathbf{V}_h$ and with $\mathrm{div} = \nabla\cdot$,

$$\mathrm{div}\mathbf{\Pi}_h = P_h \mathrm{div} \ . \tag{3.92}$$

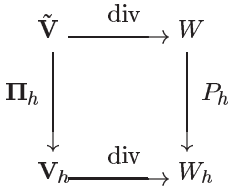

Fig. 3.9. The commuting diagram

Relation (3.92) means that the diagram in Fig. 3.9 commutes where $\tilde{\mathbf{V}} = (H^1(\Omega))^d$.

These two operators satisfy the approximation properties

$$\begin{aligned}
&\|\mathbf{v} - \mathbf{\Pi_h v}\|_{\mathbf{L}^2(\Omega)} \le Ch^l \|\mathbf{v}\|_{\mathbf{H}^l(\Omega)}, && 1 \le l \le r+1\ , \\
&\|\nabla \cdot (\mathbf{v} - \mathbf{\Pi}_h \mathbf{v})\|_{L^2(\Omega)} \le Ch^l \|\nabla \cdot \mathbf{v}\|_{H^l(\Omega)}, && 0 \le l \le r^*\ , \\
&\|w - P_h w\|_{L^2(\Omega)} \le Ch^l \|w\|_{H^l(\Omega)}, && 0 \le l \le r^*\ ,
\end{aligned} \tag{3.93}$$

where r and r^* are given as in (3.34).

The operator $P_h : W \to W_h$ is just the standard L^2-projection:

$$(P_h z - z, w) = 0, \qquad w \in W_h\ ,$$

while the operator $\mathbf{\Pi}_h : \tilde{\mathbf{V}} \to \mathbf{V}_h$ needs to be defined for each individual mixed space. As an example, we define $\mathbf{\Pi}_h$ for the RTN spaces on triangles and rectangles in detail. The operator $\mathbf{\Pi}_h$ is defined in terms of the degrees of freedom of $\mathbf{V}_h$ (cf. Sect. 3.4).

Example 3.1. The RTN space on triangles is defined in Sect. 3.4.1.1. Let $K \in K_h$ be a triangle with edges e_i, $i = 1, 2, 3$. Then we define $\mathbf{\Pi}_h|_K : H^1(K) \to \mathbf{V}_h(K)$ by

$$\begin{aligned}
&\int_{e_i} (\mathbf{v} - \mathbf{\Pi}_h \mathbf{v}) \cdot \boldsymbol{\nu} w \ d\ell = 0 && \forall w \in P_r(e_i),\ i = 1, 2, 3\ , \\
&\int_K (\mathbf{v} - \mathbf{\Pi}_h \mathbf{v}) \cdot \mathbf{w} \ d\mathbf{x} = 0 && \forall \mathbf{w} \in \left(P_{r-1}(K)\right)^2\ ,
\end{aligned} \tag{3.94}$$

for $r \ge 0$. When $r = 0$, only the first equation is needed.

Observe that the number of equations (the degrees of freedom) in the first and second equations in (3.94) is, respectively, $3(r+1)$ and $r(r+1)$, so the total number is $(r+1)(r+3)$, which is equal to the number of dimensions of $\mathbf{V}_h(K)$ (cf. Sect. 3.4.1.1). Hence, to show existence of $\mathbf{\Pi}_h$, it suffices to prove that a vector $\mathbf{v}$ in $\mathbf{V}_h(K)$ having vanishing degrees of freedom must itself vanish on K, which was shown in Sect. 3.4.1.1.

We simply point out that since $\mathbf{\Pi}_h$ reproduces $(P_k(K))^2$, it follows (Dupont-Scott, 1980; also see Sect. 1.9) that

$$\|\mathbf{v} - \boldsymbol{\Pi}_h \mathbf{v}\|_{\mathbf{L}^2(K)} \leq C h_K^l \|\mathbf{v}\|_{\mathbf{H}^l(K)}, \qquad 1 \leq l \leq r+1 \,.$$

This implies the first inequality in (3.93).

Example 3.2. The RTN space on rectangles is defined in Sect. 3.4.2.1. For a rectangle $K \in K_h$ with edges $(e_i,\ i = 1,2,3,4)$ parallel to the coordinate axes, we define $\boldsymbol{\Pi}_h|_K : H^1(K) \to \mathbf{V}_h(K)$ by

$$\begin{aligned} &\int_{e_i} (\mathbf{v} - \boldsymbol{\Pi}_h \mathbf{v}) \cdot \boldsymbol{\nu} w \; d\ell = 0 \quad \forall w \in P_r(e_i),\ i = 1,2,3,4 \,, \\ &\int_K (\mathbf{v} - \boldsymbol{\Pi}_h \mathbf{v}) \cdot \mathbf{w} \; d\mathbf{x} = 0 \quad \forall \mathbf{w} \in Q_{r-1,r}(K) \times Q_{r,r-1}(K) \,, \end{aligned} \tag{3.95}$$

for $r \geq 0$. The number of equations in (3.95) is $4(r+1) + 2r(r+1) = 2(r+1)(r+2)$, which is the number of dimensions of $\mathbf{V}_h(K)$ (refer to Sect. 3.4.2.1).

Unisolvance of $\boldsymbol{\Pi}_h \mathbf{v}$ can be established as in Example 3.1. Let $K = (0,1) \times (0,1)$ be the reference element, and let $\mathbf{v} = (v_1, v_2) \in \mathbf{V}_h(K)$ satisfy

$$\begin{aligned} &\int_{e_i} \mathbf{v} \cdot \boldsymbol{\nu} w \; d\ell = 0 \qquad \forall w \in P_r(e_i),\ i = 1,2,3,4 \,, \\ &\int_K \mathbf{v} \cdot \mathbf{w} \; d\mathbf{x} = 0 \qquad \forall \mathbf{w} \in Q_{r-1,r}(K) \times Q_{r,r-1}(K) \,. \end{aligned} \tag{3.96}$$

From the first equation of (3.96), we conclude that $\mathbf{v} \cdot \boldsymbol{\nu} = 0$ on ∂K. This implies, as in Sect. 3.4.1.1, that $v_1 = x_1(1-x_1)z$, where $z \in Q_{r-1,r}(K)$. Then, from the second equation of (3.96), we see that $z = 0$ and thus, $v_1 = 0$. Similarly, $v_2 = 0$.

3.8.5 Error Estimates

Theorem 3.4 can be utilized to obtain error estimates for (3.90). However, thanks to some special features of the mixed finite element spaces under consideration such as those in (3.91) or (3.92), better estimates can be derived.

In this subsection, we assume that Ω is a smooth domain (or a convex polygonal domain).

Lemma 3.7. *Given $w \in W_h$, there exists $\mathbf{v} \in \mathbf{V}_h$ such that $\nabla \cdot \mathbf{v} = w$ and*

$$\|\mathbf{v}\|_{\mathbf{V}} \leq C \|w\|_{L^2(\Omega)} \,.$$

Proof. For $w \in W_h$, let ψ be the solution (unique up to an additive constant) of the problem

$$\begin{aligned} &\Delta \psi = w \qquad &&\text{in } \Omega \,, \\ &\nabla \psi \cdot \boldsymbol{\nu} = 0 \qquad &&\text{on } \Gamma \,. \end{aligned} \tag{3.97}$$

Then an *elliptic regularity* result (cf. (1.121)) implies

$$\|\psi\|_{H^2(\Omega)} \le C\|w\|_{L^2(\Omega)} \,. \tag{3.98}$$

Now, take $\mathbf{v} = \mathbf{\Pi}_h \nabla\psi \in \mathbf{V}_h$. It follows from (3.92) that

$$\nabla \cdot \mathbf{v} = \nabla \cdot (\mathbf{\Pi}_h \nabla\psi) = P_h(\Delta\psi) = w \,,$$

and (3.93) and (3.98) that

$$\begin{aligned} \|\mathbf{v}\|_{\mathbf{L}^2(\Omega)} &\le \|\mathbf{\Pi}_h \nabla\psi - \nabla\psi\|_{\mathbf{L}^2(\Omega)} + \|\nabla\psi\|_{\mathbf{L}^2(\Omega)} \\ &\le Ch\|\psi\|_{H^2(\Omega)} + \|\nabla\psi\|_{\mathbf{L}^2(\Omega)} \le C\|w\|_{L^2(\Omega)} \,, \end{aligned}$$

so that the desired result follows. □

Theorem 3.8. *Let* $(\mathbf{u}, p) \in \mathbf{V} \times W$ *and* $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times W_h$ *be the respective solution of* (3.88) *and* (3.90). *Then*

$$\begin{aligned} &\|\nabla \cdot (\mathbf{u} - \mathbf{u}_h)\|_{L^2(\Omega)} \le C\|\nabla \cdot (\mathbf{u} - \mathbf{\Pi}_h \mathbf{u})\|_{L^2(\Omega)} \,, \\ &\|\mathbf{u} - \mathbf{u}_h\|_{\mathbf{L}^2(\Omega)} \le C\|\mathbf{u} - \mathbf{\Pi}_h \mathbf{u}\|_{\mathbf{L}^2(\Omega)} \,, \\ &\|p - p_h\|_{L^2(\Omega)} \le C\left(\|\mathbf{u} - \mathbf{u}_h\|_{\mathbf{L}^2(\Omega)} + \|p - P_h p\|_{L^2(\Omega)}\right) \,. \end{aligned} \tag{3.99}$$

Proof. Subtract (3.90) from (3.88) to give the error equations

$$\begin{aligned} &\left(\mathbf{a}^{-1}(\mathbf{u} - \mathbf{u}_h), \mathbf{v}\right) - (\nabla \cdot \mathbf{v}, p - p_h) = 0 \qquad &&\forall \mathbf{v} \in \mathbf{V}_h \,, \\ &\left(\nabla \cdot (\mathbf{u} - \mathbf{u}_h), w\right) = 0 &&\forall w \in W_h \,. \end{aligned} \tag{3.100}$$

First, take $w = \nabla \cdot (\mathbf{\Pi}_h \mathbf{u} - \mathbf{u}_h)$ in the second equation of (3.100) to see that

$$\begin{aligned} \left(\nabla \cdot (\mathbf{u} - \mathbf{u}_h), \nabla \cdot (\mathbf{u} - \mathbf{u}_h)\right) &= \left(\nabla \cdot (\mathbf{u} - \mathbf{u}_h), \nabla \cdot (\mathbf{u} - \mathbf{\Pi}_h \mathbf{u})\right) \\ &\quad + \left(\nabla \cdot (\mathbf{u} - \mathbf{u}_h), \nabla \cdot (\mathbf{\Pi}_h \mathbf{u} - \mathbf{u}_h)\right) \\ &= \left(\nabla \cdot (\mathbf{u} - \mathbf{u}_h), \nabla \cdot (\mathbf{u} - \mathbf{\Pi}_h \mathbf{u})\right) \,, \end{aligned}$$

which, together with Cauchy's inequality (1.10), yields the first equation in (3.99).

Next, choose $\mathbf{v} = \mathbf{\Pi}_h(\mathbf{u} - \mathbf{u}_h)$ in the first equation and $w = P_h(p - p_h)$ in the second equation of (3.100) and add the resulting equations to give

$$\begin{aligned} \left(\mathbf{a}^{-1}(\mathbf{u} - \mathbf{u}_h), \mathbf{\Pi}_h(\mathbf{u} - \mathbf{u}_h)\right) &+ \left(\nabla \cdot (\mathbf{u} - \mathbf{u}_h), P_h(p - p_h)\right) \\ &- \left(\nabla \cdot \mathbf{\Pi}_h(\mathbf{u} - \mathbf{u}_h), p - p_h\right) = 0 \,. \end{aligned}$$

It follows from (3.92) that the last two terms in the left-hand side of the above equation cancel, so that

$$\left(\mathbf{a}^{-1}(\mathbf{u} - \mathbf{u}_h), \mathbf{\Pi}_h(\mathbf{u} - \mathbf{u}_h)\right) = 0 \,.$$

Hence we have

$$\left(\mathbf{a}^{-1}(\mathbf{u}-\mathbf{u}_h),\mathbf{u}-\mathbf{u}_h\right)=\left(\mathbf{a}^{-1}(\mathbf{u}-\mathbf{u}_h),\mathbf{u}-\boldsymbol{\Pi}_h\mathbf{u}\right)\,,$$

which implies the second equation in (3.99).

Finally, take $\mathbf{v}$ in the first equation of (3.100) associated with $P_h(p-p_h)$ according to Lemma 3.7:

$$\begin{aligned}\left(P_h(p-p_h),p-p_h\right)&=\left(\nabla\cdot\mathbf{v},p-p_h\right)\\&=\left(\mathbf{a}^{-1}(\mathbf{u}-\mathbf{u}_h),\mathbf{v}\right)\\&\le C\|\mathbf{u}-\mathbf{u}_h\|^2_{\mathbf{L}^2(\Omega)}+\frac{1}{4}\|P_h(p-p_h)\|^2_{L^2(\Omega)}\,,\end{aligned}$$

which, together with the equation

$$\begin{aligned}(P_h(p-p_h),P_h(p-p_h))&=(P_h(p-p_h),p-p_h)\\&\quad-(P_h(p-p_h),p-P_hp)\,,\end{aligned}$$

leads to the third result in (3.99). □

Corollary 3.9. *Let* $(\mathbf{u},p)\in\mathbf{V}\times W$ *and* $(\mathbf{u}_h,p_h)\in\mathbf{V}_h\times W_h$ *be the respective solution of* (3.88) *and* (3.90). *Then*

$$\begin{aligned}&\|\nabla\cdot(\mathbf{u}-\mathbf{u}_h)\|_{L^2(\Omega)}\le Ch^l\|\nabla\cdot\mathbf{u}\|_{H^l(\Omega)}, && 0\le l\le r^*\,,\\&\|\mathbf{u}-\mathbf{u}_h\|_{\mathbf{L}^2(\Omega)}\le Ch^l\|\mathbf{u}\|_{\mathbf{H}^l(\Omega)}, && 1\le l\le r+1\,,\\&\|p-p_h\|_{L^2(\Omega)}\le C\left(h^{l_1}\|\mathbf{u}\|_{\mathbf{H}^{l_1}(\Omega)}+h^{l_2}\|p\|_{H^{l_2}(\Omega)}\right), && 1\le l_1\le r+1\,,\\&&& 0\le l_2\le r^*\,,\end{aligned}$$

where r *and* r^* *are defined as in* (3.34).

The proof of this corollary follows from (3.93) and Theorem 3.8 immediately.

We remark that we have obtained the error bounds only in the L^2-norm. These errors can be also bounded in other norms such as in the H^{-l}-norm ($l\ge 1$), where $H^{-l}(\Omega)$ is the dual space to $H^l(\Omega)$ (Douglas-Roberts, 1985). The application of the mixed finite element method to other partial differential problems will be presented in Chaps. 7–10.

3.9 Bibliographical Remarks

For more details on the mixed finite element method, the reader should refer to the book by Brezzi-Fortin (1991) or to a book chapter by Roberts-Thomas (1989). For more information on the analysis of each of the mixed RTN (Raviart-Thomas, 1977; Néd'elec, 1980), BDM (Brezzi et al., 1985), BDDF (Brezzi et al., 1987A), BDFM (Brezzi et al., 1987B), and CD (Chen-Douglas, 1989) finite element spaces, the reader may see the respective paper.

3.10 Exercises

3.1. Show that if $u \in V = H^1(I)$ and $p \in W = L^2(I)$ satisfy (3.4) and if p is twice continuously differentiable, then p satisfies (3.1).

3.2. Write a code to solve problem (3.1) approximately using the mixed finite element method introduced in Sect. 3.1. Use $f(x) = 4\pi^2 \sin(2\pi x)$ and a uniform partition of $(0, 1)$ with $h = 0.1$. Also, compute the errors

$$\|p - p_h\| = \left(\int_0^1 (p - p_h)^2 \, dx \right)^{1/2} ,$$

$$\|u - u_h\| = \left(\int_0^1 (u - u_h)^2 \, dx \right)^{1/2} ,$$

with $h = 0.1$, 0.01, and 0.001, and compare them. Here p, u and p_h, u_h are the solutions to (3.4) and (3.6), respectively (cf. Sect. 3.1). (If necessary, refer to Sect. 3.7 for a linear solver.)

3.3. Consider the problem with an inhomogeneous boundary condition:

$$-\frac{d^2p}{dx^2} = f(x), \qquad 0 < x < 1 ,$$
$$p(0) = p_{D0}, \quad p(1) = p_{D1} ,$$

where f is a given real-valued piecewise continuous bounded function in $(0, 1)$, and p_{D0} and p_{D1} are real numbers. Write this problem in a mixed variational formulation, and construct a mixed finite element method using the finite element spaces described in Sect. 3.1. Determine the corresponding linear system of algebraic equations for a uniform partition.

3.4. Consider the problem with a Neumann boundary condition at $x = 1$:

$$-\frac{d^2p}{dx^2} = f(x), \qquad 0 < x < 1 ,$$
$$p(0) = \frac{dp}{dx}(1) = 0 .$$

Express this problem in a mixed variational formulation, formulate a mixed finite element method using the finite element spaces considered in Sect. 3.1, and determine the corresponding linear system of algebraic equations for a uniform partition.

3.5. Construct finite element subspaces $V_h \times W_h$ of $H^1(I) \times L^2(I)$ that, respectively, consist of piecewise quadratic and linear functions on a partition of $I = (0, 1)$. How can the parameters (degrees of freedom) be chosen to describe such functions in V_h and W_h? Find the corresponding basis functions. Then define a mixed finite element method for (3.1) using these spaces $V_h \times W_h$ and express the corresponding linear system of algebraic equations for a uniform partition of I.

3.6. Show that the matrix $\mathbf{M}$ defined in Sect. 3.1 has both positive and negative eigenvalues.

3.7. Define the space

$$\mathbf{H}(\mathrm{div},\Omega) = \left\{\mathbf{v} = (v_1, v_2) \in (L^2(\Omega))^2 : \nabla \cdot \mathbf{v} \in L^2(\Omega)\right\} .$$

Show that for any decomposition of $\Omega \subset \mathbb{R}^2$ into subdomains such that the interiors of these subdomains are pairwise disjoint, $\mathbf{v} \in \mathbf{H}(\mathrm{div},\Omega)$ if and only if its normal components are continuous across the interior edges in this decomposition.

3.8. Prove that if $\mathbf{u} \in \mathbf{V} = \mathbf{H}(\mathrm{div},\Omega)$ and $p \in W = L^2(\Omega)$ satisfy (3.16) and if $p \in H^2(\Omega)$, then p satisfies (3.13).

3.9. Let the basis functions $\{\boldsymbol{\varphi}_i\}$ and $\{\psi_i\}$ of $\mathbf{V}_h$ and W_h be defined as in Sect. 3.2. For a uniform partition of $\Omega = (0,1) \times (0,1)$ given as in Fig. 1.7, determine the matrices $\mathbf{A}$ and $\mathbf{B}$ in system (3.18).

3.10. Write a code to solve problem (3.13) approximately using the mixed finite element method developed in Sect. 3.2. Use $f(x_1, x_2) = 8\pi^2 \sin(2\pi x_1)\sin(2\pi x_2)$ and a uniform partition of $\Omega = (0,1) \times (0,1)$ as given in Fig. 1.7. Also, compute the errors

$$\|p - p_h\| = \left(\int_\Omega (p - p_h)^2 \; d\mathbf{x}\right)^{1/2} ,$$
$$\|\mathbf{u} - \mathbf{u}_h\| = \left(\int_\Omega |\mathbf{u} - \mathbf{u}_h|^2 \; d\mathbf{x}\right)^{1/2} ,$$

with $h = 0.1$, 0.01, and 0.001, and compare them. Here p, $\mathbf{u}$ and p_h, $\mathbf{u}_h$ are the solutions to (3.16) and (3.17), respectively, and h is the mesh size in the x_1- and x_2-directions. (If necessary, refer to Sect. 3.7 for a linear solver.)

3.11. Consider problem (3.13) with an inhomogeneous boundary condition, i.e.,

$$\begin{aligned} -\Delta p &= f && \text{in } \Omega , \\ p &= g && \text{on } \Gamma , \end{aligned}$$

where Ω is a bounded domain in the plane with boundary Γ, and f and g are given. Express this problem in a mixed variational formulation, formulate a mixed finite element method using the finite element spaces given in Sect. 3.2, and determine the corresponding linear system of algebraic equations for a uniform partition of $\Omega = (0,1) \times (0,1)$ as displayed in Fig. 1.7.

3.12. Consider the problem

$$\begin{aligned} -\Delta p &= f && \text{in } \Omega , \\ p &= g_D && \text{on } \Gamma_D , \\ \frac{\partial p}{\partial \boldsymbol{\nu}} &= g_N && \text{on } \Gamma_N , \end{aligned}$$

where Ω is a bounded domain in the plane with boundary Γ, $\bar{\Gamma} = \bar{\Gamma}_D \cup \bar{\Gamma}_N$, $\Gamma_D \cap \Gamma_N = \emptyset$, and f, g_D, and g_N are given functions. Write down a mixed variational formulation for this problem and formulate a mixed finite element method using the finite element spaces given in Sect. 3.2.

3.13. Let $\{\boldsymbol{\varphi}_i\}$ and $\{\psi_i\}$ be the basis functions of $\mathbf{V}_h$ and W_h in system (3.25), respectively. Write (3.25) in matrix form.

3.14. Let $\mathbf{V}_h(K)$, with $r \geq 1$, be the BDM space on the triangle K (cf. Sect. 3.4.1.2). Show that a function $\mathbf{v} \in \mathbf{V}_h(K)$ is uniquely defined by the degrees of freedom

$$\begin{aligned} &(\mathbf{v} \cdot \boldsymbol{\nu}, w)_e && \forall w \in P_r(e),\ e \in \partial K\ , \\ &(\mathbf{v}, \nabla w)_K && \forall w \in P_{r-1}(K)\ , \\ &(\mathbf{v},\ \mathbf{curl}\ w)_K && \forall w \in B_{r+1}(K)\ . \end{aligned}$$

(If necessary, see Brezzi et al. (1985).)

3.15. Let $\mathbf{V}_h(K)$, with $r \geq 0$, be the RT space on the rectangle K (cf. Sect. 3.4.2.1). Show that a function $\mathbf{v} \in \mathbf{V}_h(K)$ is uniquely defined by the degrees of freedom

$$\begin{aligned} &(\mathbf{v} \cdot \boldsymbol{\nu}, w)_e && \forall w \in P_r(e),\ e \in \partial K\ , \\ &(\mathbf{v}, \mathbf{w})_K && \forall \mathbf{w} = (w_1, w_2),\ w_1 \in Q_{r-1,r}(K),\ w_2 \in Q_{r,r-1}(K)\ . \end{aligned}$$

(If necessary, refer to Raviart-Thomas (1977).)

3.16. Let $\mathbf{V}_h(K)$, with $r \geq 1$, be the BDM space on the rectangle K (cf. Sect. 3.4.2.2). Show that a function $\mathbf{v} \in \mathbf{V}_h(K)$ is uniquely defined by the degrees of freedom

$$\begin{aligned} &(\mathbf{v} \cdot \boldsymbol{\nu}, w)_e && \forall w \in P_r(e),\ e \in \partial K\ , \\ &(\mathbf{v}, \mathbf{w})_K && \forall \mathbf{w} \in (P_{r-2}(K))^2\ . \end{aligned}$$

(If necessary, see Brezzi et al. (1985).)

3.17. Let $\mathbf{V}_h(K)$, with $r \geq 0$, be the BDFM space on the rectangle K (cf. Sect. 3.4.2.3). Show that a function $\mathbf{v} \in \mathbf{V}_h(K)$ is uniquely defined by the degrees of freedom

$$\begin{aligned} &(\mathbf{v} \cdot \boldsymbol{\nu}, w)_e && \forall w \in P_r(e),\ e \in \partial K\ , \\ &(\mathbf{v}, \mathbf{w})_K && \forall \mathbf{w} \in (P_{r-1}(K))^2\ . \end{aligned}$$

(If necessary, consult Brezzi et al. (1987B).)

3.18. After introducing basis functions in $\tilde{\mathbf{V}}_h$, W_h, and L_h, prove that system (3.51) can be expressed in the matrix form (3.52).

3.19. Consider the time-dependent problem

$$\begin{aligned} &\frac{\partial p}{\partial t} - \nabla \cdot (\mathbf{a} \nabla p) = f && \text{in } \Omega \times J\ , \\ &p = 0 && \text{on } \Gamma \times J\ , \\ &p(\cdot, 0) = p_0 && \text{in } \Omega\ , \end{aligned}$$

where $\mathbf{a}$ is a $d \times d$ matrix ($d = 2$ or 3), and f and p_0 are given functions. Write down a mixed variational formulation for this problem and formulate a mixed finite element method using any pair of mixed spaces $\mathbf{V}_h \times W_h$ defined in Sect. 3.4 and the backward Euler method (cf. (1.80)) or Crank-Nicolson method (cf. (1.83)) for the time derivative. Show a stability result similar to (1.82) for the resulting method in the case $f = 0$.

3.20. Consider the problem (cf. Sect. 3.3.1)

$$\begin{aligned} -\Delta p &= f && \text{in } \Omega \,, \\ \frac{\partial p}{\partial \boldsymbol{\nu}} &= 0 && \text{on } \Gamma \,. \end{aligned}$$

Write down a mixed variational formulation for this problem and prove that the conditions in Theorem 3.2 are satisfied.

3.21. Consider the problem (cf. Sect. 3.3.2)

$$\begin{aligned} -\Delta p &= f && \text{in } \Omega \,, \\ \gamma p + \frac{\partial p}{\partial \boldsymbol{\nu}} &= g && \text{on } \Gamma \,, \end{aligned}$$

where γ is a positive constant. Formulate a mixed variational formulation for this problem and show that the conditions in Theorem 3.2 hold.

3.22. Give a mixed variational formulation for the problem

$$\begin{aligned} -\nabla \cdot (\mathbf{a} \nabla p) &= f && \text{in } \Omega \,, \\ p &= g_D && \text{on } \Gamma_D \,, \\ \gamma p + \mathbf{a} \nabla p \cdot \boldsymbol{\nu} &= g_N && \text{on } \Gamma_N \,, \end{aligned}$$

where $\mathbf{a}$ is a $d \times d$ matrix ($d = 2$ or 3), f, g_D, and g_N are given functions of $\mathbf{x}$, and γ is a constant. Under what conditions on $\mathbf{a}$ and γ are the conditions in Theorem 3.2 satisfied?

3.23. Prove that the *inf-sup* condition (3.67) is equivalent to the property: For every $v \in V$, there is a decomposition $v = v_1 + v_2$ such that $v_1 \in Z$, $v_2 \in Z^\perp$, and

$$\|v_2\|_V \leq b_*^{-1} \|B' v\|_{W'} \,,$$

where the constant $b_* > 0$ is independent of v, and Z and $Z^\perp$ are defined as in Sect. 3.8.1.

3.24. Assume that the bilinear form b satisfies condition (3.67). Show that if the discrete *inf-sup* condition (3.78) holds, then there exists a projection operator $\Pi_h : V \to V_h$ such that

$$b(v - \Pi_h v, w) = 0 \qquad \forall w \in W_h \,,$$

and Π_h is uniformly bounded. (Compare with Theorem 3.6.)

3.25. Consider the biharmonic problem (cf. Example 1.5)

$$\begin{aligned} \Delta^2 p &= f && \text{in } \Omega \,, \\ p = \frac{\partial p}{\partial \boldsymbol{\nu}} &= 0 && \text{on } \Gamma \,. \end{aligned}$$

Prove that $u = \Delta p \in H^1(\Omega)$ and $p \in H_0^1(\Omega)$ satisfy the mixed weak formulation

$$\begin{aligned} (u, v) + (\nabla v, \nabla p) &= 0 && \forall v \in H^1(\Omega) \,, \\ (\nabla u, \nabla w) &= -(f, w) && \forall w \in H_0^1(\Omega) \,. \end{aligned}$$

(See Ciarlet (1978), Babuška et al. (1980), and Chen (1997) for appropriate mixed finite element spaces for this problem.)

4 Discontinuous Finite Elements

In Chap. 1, functions used in finite element spaces for the discretization of second-order partial differential equations were continuous across interelement boundaries. In Chap. 2, functions in the finite element spaces were continuous at certain points on interelement boundaries. The functions of this type were also used in Chap. 3 for the vector finite element spaces. In this chapter, we consider the case where the functions in the finite element spaces are totally discontinuous across interelement boundaries, i.e., *discontinuous finite elements*. The *discontinuous Galerkin (DG) finite element method* was originally introduced for a linear *advection (hyperbolic) problem* (Reed-Hill, 1973; LeSaint-Raviart, 1974). This method has established itself as an important alternative for numerically solving advection problems for which the continuous (conforming) finite element method does not work well. An important feature of DG is that it conserves mass locally (cf. (4.3)). This feature has led to increased efforts to use it also for *diffusion problems*, with an ultimate goal of solving advection-diffusion problems (see Chap. 5 for problems of this type). The use of DG with *penalty* (or *stabilization*) for diffusion problems traces back to the 1970's (Douglas-Dupont, 1976; Douglas, 1977), but was then abandoned. The reason is that stability and convergence of DG with penalty depend heavily on the choice of penalty parameters. Recently, as DG has become more popular for advection problems, there have been tremendous efforts to make it work also for the diffusion problems. In this chapter, we introduce the DG method and its various extensions. In Sect. 4.1, we first study DG and its stabilized versions for advection problems. Then, in Sect. 4.2, we show how to extend these methods to diffusion problems. In Sect. 4.3, we discuss the recently developed mixed discontinuous finite element method. Section 4.4 is devoted to theoretical considerations. Finally, bibliographical information is given in Sect. 4.5.

4.1 Advection Problems

We consider the advection problem:

$$\begin{aligned} &\mathbf{b} \cdot \nabla p + Rp = f, && \mathbf{x} \in \Omega \,, \\ &p = g, && \mathbf{x} \in \Gamma_- \,, \end{aligned} \tag{4.1}$$

where the functions $\mathbf{b}$, R, f, and g are given, $\Omega \subset \mathbb{R}^d$ ($d \leq 3$) is a bounded domain with boundary Γ, the *inflow boundary* Γ_- is defined by

$$\Gamma_- = \{\mathbf{x} \in \Gamma : \ (\mathbf{b} \cdot \boldsymbol{\nu})(\mathbf{x}) < 0\} ,$$

and $\boldsymbol{\nu}$ is the outward unit normal to Γ. The advection coefficient $\mathbf{b}$ is assumed to be smooth in $(\mathbf{x}, t)$, and the reaction coefficient R is assumed to be bounded and nonnegative. This problem will be further considered in Sect. 5.2.

4.1.1 DG Methods

For $h > 0$, let K_h be a finite element partition of Ω into elements $\{K\}$; each element $K \in K_h$ has a Lipschitz boundary ∂K (see the definition of a Lipschitz domain in Sect. 1.9.1). Furthermore, K_h is assumed to satisfy the usual minimum angle condition (cf. (1.52)). For the DG method, adjacent elements in K_h are not required to match; a vertex of one element can lie in the interior of the edge or face of another element, for example. Let $\mathcal{E}_h^o$ denote the set of all interior boundaries e in K_h, $\mathcal{E}_h^b$ the set of the boundaries e on Γ, and $\mathcal{E}_h = \mathcal{E}_h^o \cup \mathcal{E}_h^b$. We tacitly assume that $\mathcal{E}_h^o \neq \emptyset$.

Associated with K_h, we define the finite element space

$$V_h = \{v : v \text{ is a bounded function on } \Omega \text{ and } v|_K \in P_r(K), \ K \in K_h\} ,$$

where $P_r(K)$ is the space of polynomials on K of degree at most $r \geq 0$. Note that no continuity across interelement boundaries is required on functions in this space.

To introduce DG, we need some notation. For each $K \in K_h$, we split its boundary ∂K into the inflow and outflow parts by

$$\partial K_- = \{\mathbf{x} \in \partial K : \ (\mathbf{b} \cdot \boldsymbol{\nu})(\mathbf{x}) < 0\} ,$$
$$\partial K_+ = \{\mathbf{x} \in \partial K : \ (\mathbf{b} \cdot \boldsymbol{\nu})(\mathbf{x}) \geq 0\} ,$$

where $\boldsymbol{\nu}$ is the outward unit normal to ∂K. A triangle K with boundary made up of ∂K_- and ∂K_+ is shown is Fig. 4.1. For $e \in \mathcal{E}_h^o$, the left- and right-hand limits on e of a function $v \in V_h$ are defined by

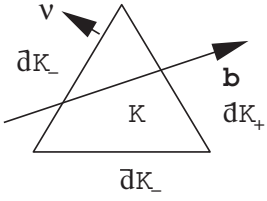

Fig. 4.1. An illustration of ∂K_- and ∂K_+

$$v_-(\mathbf{x}) = \lim_{\epsilon \to 0^-} v\left(\mathbf{x} + \epsilon \mathbf{b}\right), \quad v_+(\mathbf{x}) = \lim_{\epsilon \to 0^+} v\left(\mathbf{x} + \epsilon \mathbf{b}\right) \ ,$$

for $\mathbf{x} \in e$. The jump of v across e is given by

$$[v] = v_+ - v_- \ .$$

For $e \in \mathcal{E}_h^b$, we define (from inside Ω)

$$[v] = v \ .$$

Now, the DG method for (4.1) is defined as follows: For $K \in K_h$, given $p_{h,-}$ on ∂K_-, find $p_h = p_h|_K \in P_r(K)$ such that

$$\begin{aligned} &(\mathbf{b} \cdot \nabla p_h + R p_h, v)_K - \int_{\partial K_-} p_{h,+} v_+ \mathbf{b} \cdot \boldsymbol{\nu} \, d\ell \\ &\quad = (f, v)_K - \int_{\partial K_-} p_{h,-} v_+ \mathbf{b} \cdot \boldsymbol{\nu} \, d\ell \qquad \forall v \in P_r(K) \ , \end{aligned} \tag{4.2}$$

where we recall that

$$(v, w)_K = \int_K vw \, d\mathbf{x}, \quad p_{h,-} = g \text{ on } \Gamma_- \ .$$

Note that (4.2) is the standard finite element method for (4.1) on the element K, with the boundary condition being *weakly* imposed. If $p_{h,-}$ is given on ∂K_-, existence and uniqueness of a solution to (4.2) can be shown as in Chap. 1 (see the remarks following (4.9) later). Equation (4.2) also holds for the continuous problem (4.1) (cf. Sect. 4.4). For a typical triangulation (cf. Fig. 4.2), p_h can be determined first on the triangles K adjacent to Γ_-. Then this process is continued until p_h is found in the whole domain Ω. Thus the computation of (4.2) is local.

If $\mathbf{b}$ is *divergence-free* (or *solenoidal*), i.e., $\nabla \cdot \mathbf{b} = 0$, we use Green's formula (1.19) to see that (cf. Fig. 4.1)

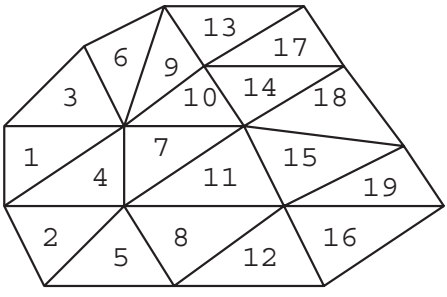

Fig. 4.2. An ordering of computation for DG

$$(\mathbf{b}\cdot\nabla p_h, 1)_K = \int_{\partial K_-} p_{h,+}\mathbf{b}\cdot\boldsymbol{\nu}\; d\ell + \int_{\partial K_+} p_{h,-}\mathbf{b}\cdot\boldsymbol{\nu}\; d\ell\;.$$

We substitute this into (4.2) with $v = 1$ to give

$$(Rp_h, 1)_K + \int_{\partial K_+} p_{h,-}\mathbf{b}\cdot\boldsymbol{\nu}\; d\ell = (f,1)_K - \int_{\partial K_-} p_{h,-}\mathbf{b}\cdot\boldsymbol{\nu}\; d\ell\;, \tag{4.3}$$

which expresses a local conservation property (i.e., the difference between inflow and outflow equals the sum of accumulation of mass).

To express (4.2) in the form used in Chap. 1, we define

$$a_K(v,w) = (\mathbf{b}\cdot\nabla v + Rv, w)_K - \int_{\partial K_-} [v]w_+\mathbf{b}\cdot\boldsymbol{\nu}\; d\ell, \quad K\in K_h\;,$$

and

$$a(v,w) = \sum_{K\in K_h} a_K(v,w)\;.$$

Then (4.2) is expressed as follows: Find $p_h \in V_h$ such that

$$a(p_h, v) = (f,v) \qquad \forall v\in V_h\;, \tag{4.4}$$

where $p_{h,-} = g$ on Γ_-. Before we state stability and convergence results for (4.4), let us consider a couple of examples.

Example 4.1. A one-dimensional example of (4.1) is

$$\begin{aligned} &\frac{dp}{dx} + p = f, \qquad x\in(0,1)\;,\\ &p(0) = g\;. \end{aligned} \tag{4.5}$$

Let $0 = x_0 < x_1 < \ldots < x_M = 1$ be a partition of $(0,1)$ into a set of subintervals $I_i = (x_{i-1}, x_i)$, with length $h_i = x_i - x_{i-1}$, $i = 1,2,\ldots,M$. In this case, (4.2) becomes: For $i = 1,2,\ldots,M$, given $(p_h(x_{i-1}))_-$, find $p_h = p_h|_{I_i} \in P_r(I_i)$ such that

$$\left(\frac{dp_h}{dx} + p_h, v\right)_{I_i} + [p_h(x_{i-1})]\,(v(x_{i-1}))_+ = (f,v)_{I_i} \qquad \forall v\in P_r(I_i)\;,$$

where $(p_h(x_0))_- = g$. In the case $r = 0$, V_h is the space of piecewise constants, and the DG method reduces to: For $i = 1,2,\ldots,M$, find $p_i = (p_h(x_i))_-$ such that

$$\begin{aligned} &\frac{p_i - p_{i-1}}{h_i} + p_i = \frac{1}{h_i}\int_{I_i} f\; dx\;,\\ &p_0 = g\;. \end{aligned} \tag{4.6}$$

Note that (4.6) is nothing but a simple *upwind* finite difference method with an averaged right-hand side.

Example 4.2. Set $R = f = 0$ in the advection problem (4.1). Then (4.1) simplifies to

$$\begin{aligned} \mathbf{b} \cdot \nabla p &= 0, \qquad && \mathbf{x} \in \Omega \,, \\ p &= g, && \mathbf{x} \in \Gamma_- \,. \end{aligned} \tag{4.7}$$

Also, let $r = 0$. Then (4.2) reads: For $K \in K_h$, given $p_{h,-}$ on ∂K_-, find $p_K = p_h|_K$ such that

$$\int_{\partial K_-} p_K \mathbf{b} \cdot \boldsymbol{\nu} \, d\ell = \int_{\partial K_-} p_{h,-} \mathbf{b} \cdot \boldsymbol{\nu} \, d\ell \,;$$

that is,

$$p_K = \frac{\displaystyle\int_{\partial K_-} p_{h,-} \mathbf{b} \cdot \boldsymbol{\nu} \, d\ell}{\displaystyle\int_{\partial K_-} \mathbf{b} \cdot \boldsymbol{\nu} \, d\ell} \,. \tag{4.8}$$

Thus we see that for each $K \in K_h$, the value p_K is determined by a weighted average of the values $p_{h,-}$ on adjoining elements with edges on ∂K_-. As an example, let Ω be a rectangular domain in $\mathbb{R}^2$, K_h consist of rectangles, and $\mathbf{b} > \mathbf{0}$. In this case, for a configuration shown in Fig. 4.3, we see that

$$p_3 = \frac{b_1}{b_1 + b_2} p_1 + \frac{b_2}{b_1 + b_2} p_2 \,,$$

where $p_i = p_h|_{K_i}$, $i = 1, 2, 3$, and $\mathbf{b} = (b_1, b_2)$. Again, in this case, (4.8) corresponds to a usual upwind finite difference method for (4.7).

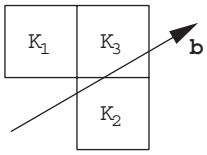

Fig. 4.3. Adjoining rectangles

Let us now state stability and convergence properties of the DG method (4.4). Their proof will be given in Sect. 4.4. We define the norm

$$\|v\|_{\mathbf{b}} = \left(\|R^{1/2} v\|^2_{L^2(\Omega)} + \frac{1}{2} \sum_{K \in K_h} \int_{\partial K_-} [v]^2 |\mathbf{b} \cdot \boldsymbol{\nu}| \, d\ell + \frac{1}{2} \int_{\Gamma_+} v_-^2 \mathbf{b} \cdot \boldsymbol{\nu} \, d\ell \right)^{1/2} .$$

Then, if $\nabla \cdot \mathbf{b} = 0$, it can be shown that

$$a(v, v) = \|v\|^2_{\mathbf{b}} - \frac{1}{2} \int_{\Gamma_-} v_-^2 |\mathbf{b} \cdot \boldsymbol{\nu}| \, d\ell, \qquad v \in V_h \,. \tag{4.9}$$

Using (4.9), existence and uniqueness of a solution to (4.4) can be proven in the usual way. If $R - \nabla \cdot \mathbf{b}/2 \geq 0$ (instead of assuming $\nabla \cdot \mathbf{b} = 0$), the term $\|R^{1/2} v\|_{L^2(\Omega)}$ can be replaced with the quantity $\| (R - \nabla \cdot \mathbf{b}/2)^{1/2} v\|_{L^2(\Omega)}$ in the definition of $\|v\|_{\mathbf{b}}$.

If R is strictly positive with respect to $\mathbf{x} \in \Omega$ (i.e., $R(\mathbf{x}) \geq R_0 > 0$), it can be seen from (4.4) and (4.9) that

$$\|p_h\|_{\mathbf{b}} \leq C \left(\|f\|^2_{L^2(\Omega)} + \int_{\Gamma_-} g^2 |\mathbf{b} \cdot \boldsymbol{\nu}| \, d\ell \right)^{1/2} . \tag{4.10}$$

This is a stability result for (4.4) in terms of data f and g. If the solution p to (4.1) is in $H^{r+1}(K)$ for each $K \in K_h$, an error estimate for (4.4) is given by

$$\begin{aligned} &\|p - p_h\|^2_{L^2(\Omega)} + h \sum_{K \in K_h} \|\mathbf{b} \cdot \nabla (p - p_h)\|^2_{L^2(K)} \\ &\quad \leq C h^{2r+1} \sum_{K \in K_h} \|p\|^2_{H^{r+1}(K)} , \end{aligned} \tag{4.11}$$

for $r \geq 0$. Note that the $L^2(\Omega)$-estimate is half a power of h from being optimal, while the $L^2(\Omega)$-estimate of the derivative in the velocity (or *streamline*) direction is in fact optimal. For general triangulations, this $L^2(\Omega)$-estimate is sharp in the sense that the exponent of h cannot be increased (Johnson, 1994).

We end with a remark that a time-dependent advection problem can be written in the same form as (4.1). To see this, consider the problem

$$c \frac{\partial p}{\partial t} + \mathbf{b} \cdot \nabla p + Rp = f, \quad \mathbf{x} \in \Omega, \ t > 0 ,$$

and set $t = x_0$ and $b_0 = c$. Then we see that

$$\bar{\mathbf{b}} \cdot \nabla_{(t,\mathbf{x})} p + Rp = f ,$$

where $\bar{\mathbf{b}} = (\mathrm{b}_0, \mathbf{b})$ and $\nabla_{(t,\mathbf{x})} = (\frac{\partial}{\partial t}, \nabla_{\mathbf{x}})$. Thus the above development of the DG method for (4.1) applies.

4.1.2 Stabilized DG Methods

We now consider a stabilized DG (SDG) method, which modifies (4.2) as follows: For $K \in K_h$, given $p_{h,-}$ on ∂K_-, find $p_h = p_h|_K \in P_r(K)$ such that

$$\begin{aligned} &(\mathbf{b} \cdot \nabla p_h + R p_h, v + \theta \mathbf{b} \cdot \nabla v)_K - \int_{\partial K_-} p_{h,+} v_+ \mathbf{b} \cdot \boldsymbol{\nu} \, d\ell \\ &\quad = (f, v + \theta \mathbf{b} \cdot \nabla v)_K - \int_{\partial K_-} p_{h,-} v_+ \mathbf{b} \cdot \boldsymbol{\nu} \, d\ell \quad \forall v \in P_r(K) , \end{aligned} \tag{4.12}$$

where θ is a *stabilization parameter*. The difference between (4.2) and (4.12) is that a stabilized term is added in the left- and right-hand sides of (4.12). This stabilized method is also called the *streamline diffusion method* due to the intuition that the added term $\theta\,(\mathbf{b}\cdot\nabla p_h, \mathbf{b}\cdot\nabla v)$ corresponds to the diffusion in the direction of streamlines (or characteristics) (Johnson, 1994). The parameter θ is chosen by the rule: $\theta = \mathcal{O}(h)$, to generate the same convergence rate as for DG. For $r = 0$, DG and SDG are the same.

Now, the bilinear forms $a_K(\cdot,\cdot)$ and $a(\cdot,\cdot)$ are defined by

$$a_K(v,w) = (\mathbf{b}\cdot\nabla v + Rv, w + \theta\mathbf{b}\cdot\nabla w)_K - \int_{\partial K_-} [v] w_+ \mathbf{b}\cdot\boldsymbol{\nu}\; d\ell, \qquad K \in K_h\ ,$$

and

$$a(v,w) = \sum_{K\in K_h} a_K(v,w)\ .$$

Then (4.12) is expressed as follows: Find $p_h \in V_h$ such that

$$a(p_h, v) = \sum_{K\in K_h} (f, v + \theta\mathbf{b}\cdot\nabla v)_K \qquad \forall v \in V_h\ , \tag{4.13}$$

where $p_{h,-} = g$ on Γ_-.

If $1 - \theta R/2 \geq 0$, the norm $\|\cdot\|_{\mathbf{b}}$ is modified to

$$\|v\|_{\mathbf{b}} = \left(\left\| R^{1/2}\,(1-\theta R/2)^{1/2}\, v \right\|^2_{L^2(\Omega)} + \frac{1}{2}\sum_{K\in K_h} \int_{\partial K_-} [v]^2 |\mathbf{b}\cdot\boldsymbol{\nu}|\; d\ell + \frac{1}{2}\sum_{K\in K_h} \|\theta^{1/2}\mathbf{b}\cdot\nabla v\|^2_{L^2(K)} + \frac{1}{2}\int_{\Gamma_+} v_-^2\, \mathbf{b}\cdot\boldsymbol{\nu}\; d\ell \right)^{1/2}\ .$$

Then, if $\mathbf{b}$ satisfies $\nabla\cdot\mathbf{b} = 0$, it can be seen (cf. Sect. 4.4) that

$$a(v,v) \geq \|v\|^2_{\mathbf{b}} - \frac{1}{2}\int_{\Gamma_-} v_-^2\, |\mathbf{b}\cdot\boldsymbol{\nu}|\; d\ell, \qquad v \in V_h\ . \tag{4.14}$$

Moreover, the stability and convergence results (4.10) and (4.11) hold for (4.13) as well.

Example 4.3. We now apply the DG and SDG methods to a one-dimensional advection problem

$$\begin{aligned} &\frac{dp}{dx} = \delta(x - x_c), \qquad x \in (0,1)\ , \\ &p(0) = 1\ , \end{aligned} \tag{4.15}$$

where the location x_c of the Dirac delta function δ is chosen within the interval $(0.3, 0.4)$. The interval $(0,1)$ is divided into ten subintervals of equal

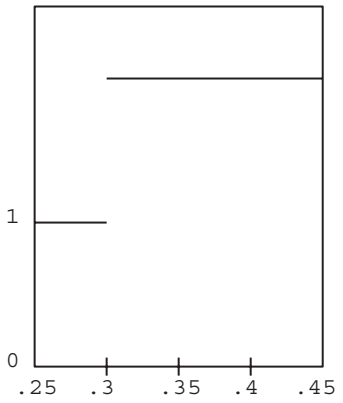

Fig. 4.4. DG and SDG for advection with $r = 0$

length. The approximate solutions by DG and SDG with different degrees r of polynomials are displayed in Figs. 4.4–4.10. From these figures we have the following observations (Hughes et al., 2000):

- **Monotonicity and continuity**. With $r = 0$, DG and SDG are the same, as noted. The constant approximation is the only one that retains monotonicity of the exact solution (cf. Fig. 4.4). For $r > 0$, monotonicity is lost for DG (cf. Figs. 4.5–4.10). SDG improves the approximate solution (cf. Fig. 4.5) and yields monotonicity and continuity in the case $r = 1$ in the limit as $\theta \to \infty$ (cf. Fig. 4.6). For $r > 1$, however, monotonicity is lost for both DG and SDG (cf. Figs. 4.7–4.10). SDG yields continuity in the limit as $\theta \to \infty$, but not monotonicity. In other words, the stabilization cannot ensure an accurate approximation of the solution using higher-order polynomials in the element where the Dirac delta function is located. Thus the method of increasing the order of polynomials is not a good choice in the case of shocks, and the constant approximation most closely reflects the character of the exact solution.
- **Dependence on the location of the Dirac delta function**. The constant approximation is independent of the location x_c of the Dirac delta function within the element (cf. Fig. 4.4). For $r > 0$, however, the approximation for DG and SDG strongly depends on x_c. When the Dirac delta function is close to an upwind node, DG produces more accurate results than SDG (cf. Fig. 4.9). When this function is close to a downwind node, SDG is better than DG. Again, the constant approximation generates the best results.
- **Localization**. From Figs. 4.5–4.10, we see that both DG and SDG localize the non-monotonicity within a single element. This important property is true for advection. For diffusion, however, we will see in the next section that localization is lost, and the error of one element can pollute the solution in other elements (cf. Fig. 4.12). It can be shown that for the

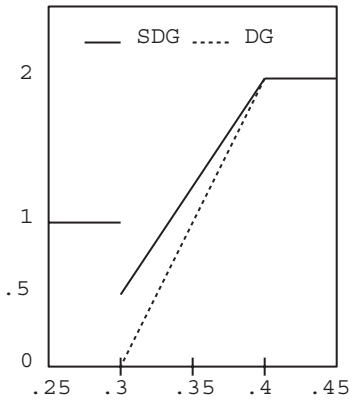

Fig. 4.5. DG and SDG ($\theta = h/2$) for advection with $r = 1$

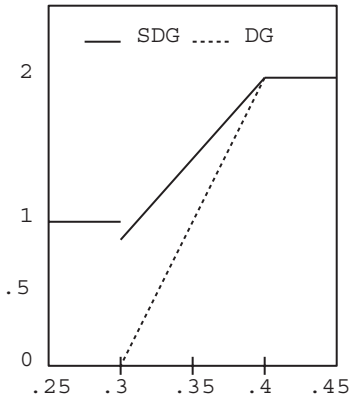

Fig. 4.6. DG and SDG ($\theta = 10h/2$) for advection with $r = 1$

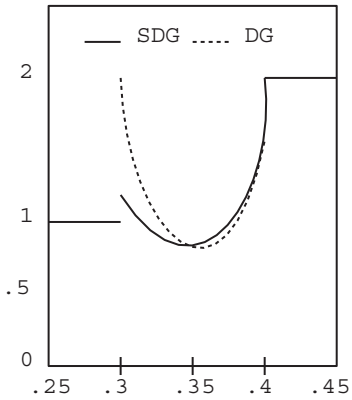

Fig. 4.7. DG and SDG ($\theta = h/2$) for advection with $r = 2$

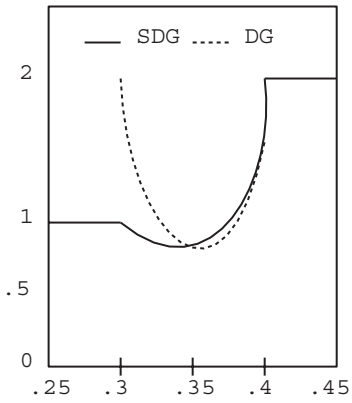

Fig. 4.8. DG and SDG ($\theta = 10h/2$) for advection with $r = 2$

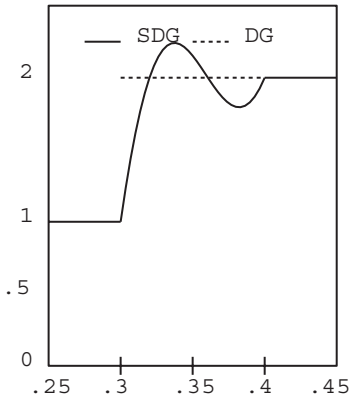

Fig. 4.9. DG and SDG ($\theta = 10h/2,\ x_c = 0.3001$) for advection with $r = 3$

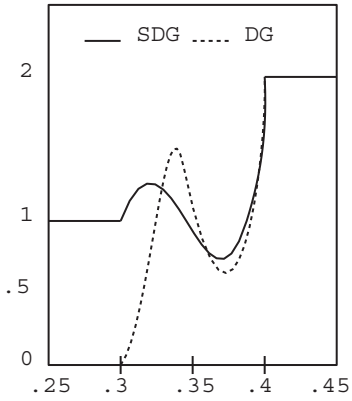

Fig. 4.10. DG and SDG ($\theta = 10h/2,\ x_c = 0.3999$) for advection with $r = 3$

advection problem under consideration, at the downwind node in the element containing the delta function, the solution is exact for all $r \geq 0$. This phenomenon can be seen from Figs. 4.4–4.10, and follows from the local conservation feature of DG.

4.2 Diffusion Problems

In this section, we extend the DG and SDG methods to the diffusion problem

$$\begin{aligned} -\nabla \cdot \big(\mathbf{a}\nabla p\big) + Rp &= f, && \mathbf{x} \in \Omega \ , \\ \mathbf{a}\nabla p \cdot \boldsymbol{\nu} &= g_N, && \mathbf{x} \in \Gamma_N \ , \\ p &= g_D, && \mathbf{x} \in \Gamma_D \ , \end{aligned} \tag{4.16}$$

where Γ_N and Γ_D denote, respectively, the Dirichlet and Neumann parts of the boundary Γ, $\bar{\Gamma}_N \cup \bar{\Gamma}_D = \bar{\Gamma}$, and $\Gamma_N \cap \Gamma_D = \emptyset$. The diffusion tensor $\mathbf{a}$ is assumed to be bounded, symmetric, and uniformly positive-definite in $\mathbf{x} \in \Omega$ (cf. (3.27)), and the reaction coefficient R is assumed to be bounded and nonnegative.

For $h > 0$, let K_h be a finite element partition of Ω into elements $\{K\}$, as in Sect. 4.1.1. For $l \geq 0$, define

$$H^l(K_h) = \big\{ v \in L^2(\Omega) : v|_K \in H^l(K), K \in K_h \big\} \ .$$

The functions in $H^l(K_h)$ are piecewise smooth. With each $e \in \mathcal{E}_h$, we associate a unit normal vector $\boldsymbol{\nu}$. For $e \in \mathcal{E}_h^b$, $\boldsymbol{\nu}$ is just the outward unit normal to Γ. For $e \in \mathcal{E}_h^o$, with $e = \bar{K}_1 \cap \bar{K}_2$, $K_1, K_2 \in K_h$, the direction of $\boldsymbol{\nu}$ is associated with the definition of jumps across e; for $v \in H^l(T_h)$ with $l > 1/2$, if the jump of v across e is defined by

$$[\![v]\!] = (v|_{K_2})|_e - (v|_{K_1})|_e \ , \tag{4.17}$$

then $\boldsymbol{\nu}$ is defined as the unit normal exterior to K_2 (cf. Fig. 4.11). The average of v on e is defined as

$$\{\!\{v\}\!\} = \frac{1}{2}\big((v|_{K_1})|_e + (v|_{K_2})|_e\big) \ . \tag{4.18}$$

As a convention, for $e \in \mathcal{E}_h^b$, the definitions are (from inside Ω)

$$\{\!\{v\}\!\} = v|_e \quad \text{and} \quad [\![v]\!] = \begin{cases} v & \text{if } e \in \mathcal{E}_h^D \ , \\ 0 & \text{if } e \in \mathcal{E}_h^N , \end{cases} \tag{4.19}$$

where $\mathcal{E}_h^D$ and $\mathcal{E}_h^N$ are the sets of edges (respectively, faces) e on Γ_D and Γ_N, respectively.

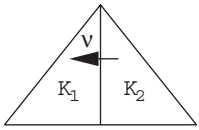

Fig. 4.11. An illustration of $\boldsymbol{\nu}$

Multiplying the first equation of (4.16) by $v \in H^2(K_h)$ and integrating over each element $K \in K_h$, we see that

$$-\left(\nabla \cdot \left(\mathbf{a}\nabla p\right), v\right)_K + (Rp, v)_K = (f, v)_K \, .$$

We apply Green's formula (1.19) to the first term of this equation to have

$$-\left(\mathbf{a}\nabla p \cdot \boldsymbol{\nu}, v\right)_{\partial K} + \left(\mathbf{a}\nabla p, \nabla v\right)_K + (Rp, v)_K = (f, v)_K \, ,$$

so we sum over $K \in K_h$ to obtain

$$-\sum_{K\in K_h} \left(\mathbf{a}\nabla p \cdot \boldsymbol{\nu}, v\right)_{\partial K} + \sum_{K\in K_h} \left[\left(\mathbf{a}\nabla p, \nabla v\right)_K + (Rp, v)_K\right] = \sum_{K\in K_h} (f, v)_K \, . \tag{4.20}$$

Note that the boundary integrals in (4.20) can be split as follows:

$$\begin{aligned}\sum_{K\in K_h} \left(\mathbf{a}\nabla p \cdot \boldsymbol{\nu}, v\right)_{\partial K} = & \sum_{e\in \mathcal{E}_h^D} \left(\mathbf{a}\nabla p \cdot \boldsymbol{\nu}, v\right)_e + \sum_{e\in \mathcal{E}_h^N} \left(\mathbf{a}\nabla p \cdot \boldsymbol{\nu}, v\right)_e \\ & + \sum_{e\in \mathcal{E}_h^o} \left[\left(\mathbf{a}\nabla p \cdot \boldsymbol{\nu}, v\right)_{\partial K_1 \cap e} + \left(\mathbf{a}\nabla p \cdot \boldsymbol{\nu}, v\right)_{\partial K_2 \cap e}\right] \, ,\end{aligned}$$

where $e = \partial K_1 \cap \partial K_2$, $K_1, K_2 \in K_h$. For simplicity of notation, we write

$$\begin{aligned}\sum_{e\in \mathcal{E}_h^D} \left(\mathbf{a}\nabla p \cdot \boldsymbol{\nu}, v\right)_e &= \left(\mathbf{a}\nabla p \cdot \boldsymbol{\nu}, v\right)_{\Gamma_D} \, , \\ \sum_{e\in \mathcal{E}_h^N} \left(\mathbf{a}\nabla p \cdot \boldsymbol{\nu}, v\right)_e &= \left(\mathbf{a}\nabla p \cdot \boldsymbol{\nu}, v\right)_{\Gamma_N} \, .\end{aligned} \tag{4.21}$$

For $e = \partial K_1 \cap \partial K_2$, with the jump definition in (4.17) and a corresponding unit normal $\boldsymbol{\nu}$ on e (exterior to K_2), we see that

$$\begin{aligned}&\left(\mathbf{a}\nabla p \cdot \boldsymbol{\nu} v\right)_{\partial K_1 \cap e} + \left(\mathbf{a}\nabla p \cdot \boldsymbol{\nu} v\right)_{\partial K_2 \cap e} \\ &\quad = \left(\mathbf{a}\nabla p \; v\right)_{\partial K_2 \cap e} \cdot \boldsymbol{\nu} - \left(\mathbf{a}\nabla p \; v\right)_{\partial K_1 \cap e} \cdot \boldsymbol{\nu} \, .\end{aligned}$$

Using the algebraic identity

$$\eta\xi - \zeta\sigma = \frac{1}{2}(\eta + \xi)(\zeta - \sigma) + \frac{1}{2}(\eta - \xi)(\zeta + \sigma) \, ,$$

we have, on $e = \partial K_1 \cap \partial K_2$,

$$\begin{aligned}(\mathbf{a}\nabla p\ v)_{\partial K_2 \cap e} \cdot \boldsymbol{\nu} - (\mathbf{a}\nabla p\ v)_{\partial K_1 \cap e} \cdot \boldsymbol{\nu} \\ = \{\!\{\mathbf{a}\nabla p \cdot \boldsymbol{\nu}\}\!\}[\![v]\!] + [\![\mathbf{a}\nabla p \cdot \boldsymbol{\nu}]\!]\{\!\{v\}\!\}\ .\end{aligned}$$

Applying these results, the integrals on interior boundaries become

$$\begin{aligned}\sum_{e \in \mathcal{E}_h^o} \left[(\mathbf{a}\nabla p \cdot \boldsymbol{\nu}, v)_{\partial K_1 \cap e} + (\mathbf{a}\nabla p \cdot \boldsymbol{\nu}, v)_{\partial K_2 \cap e}\right] \\ = \sum_{e \in \mathcal{E}_h^o} \left[(\{\!\{\mathbf{a}\nabla p \cdot \boldsymbol{\nu}\}\!\}, [\![v]\!])_e + ([\![\mathbf{a}\nabla p \cdot \boldsymbol{\nu}]\!], \{\!\{v\}\!\})_e\right]\ .\end{aligned} \tag{4.22}$$

Consequently, we apply (4.21), (4.22), and the Neumann boundary condition to (4.20) to see that

$$\begin{aligned}\sum_{K \in K_h} \left[(\mathbf{a}\nabla p, \nabla v)_K + (Rp, v)_K\right] - \sum_{e \in \mathcal{E}_h^o} (\{\!\{\mathbf{a}\nabla p \cdot \boldsymbol{\nu}\}\!\}, [\![v]\!])_e \\ - \sum_{e \in \mathcal{E}_h^o} ([\![\mathbf{a}\nabla p \cdot \boldsymbol{\nu}]\!], \{\!\{v\}\!\})_e - (\mathbf{a}\nabla p \cdot \boldsymbol{\nu}, v)_{\Gamma_D} \\ = \sum_{K \in K_h} (f, v)_K + (g_N, v)_{\Gamma_N}\ .\end{aligned} \tag{4.23}$$

Note that if the fluxes $\mathbf{a}\nabla p \cdot \boldsymbol{\nu}$ are continuous almost everywhere in Ω (e.g., when $p \in H^2(\Omega)$), we have

$$\sum_{e \in \mathcal{E}_h^o} ([\![\mathbf{a}\nabla p \cdot \boldsymbol{\nu}]\!], \{\!\{v\}\!\})_e = 0 \qquad \forall v \in H^2(K_h)\ .$$

Then (4.23) reduces to

$$\begin{aligned}\sum_{K \in K_h} \left[(\mathbf{a}\nabla p, \nabla v)_K + (Rp, v)_K\right] - \sum_{e \in \mathcal{E}_h^o} (\{\!\{\mathbf{a}\nabla p \cdot \boldsymbol{\nu}\}\!\}, [\![v]\!])_e \\ - (\mathbf{a}\nabla p \cdot \boldsymbol{\nu}, v)_{\Gamma_D} = \sum_{K \in K_h} (f, v)_K + (g_N, v)_{\Gamma_N}\ .\end{aligned} \tag{4.24}$$

We introduce the bilinear forms $b(\cdot,\cdot) : H^2(K_h) \times H^2(K_h) \to \mathbb{R}$ and $J(\cdot,\cdot) : H^2(K_h) \times H^2(K_h) \to \mathbb{R}$ by

$$\begin{aligned}b(p, v) &= \sum_{K \in K_h} \left[(\mathbf{a}\nabla p, \nabla v)_K + (Rp, v)_K\right]\ , \\ J(p, v) &= \sum_{e \in \mathcal{E}_h^o} (\{\!\{\mathbf{a}\nabla p \cdot \boldsymbol{\nu}\}\!\}, [\![v]\!])_e + (\mathbf{a}\nabla p \cdot \boldsymbol{\nu}, v)_{\Gamma_D}\ .\end{aligned} \tag{4.25}$$

Also, we define the linear form $L : H^2(K_h) \to \mathbb{R}$ by

$$L(v) = \sum_{K \in K_h} (f, v)_K + (g_N, v)_{\Gamma_N} . \tag{4.26}$$

Then a discontinuous weak formulation of (4.16) is

$$b(p, v) - J(p, v) = L(v) \qquad \forall v \in H^2(K_h) . \tag{4.27}$$

The definition of subsequent DG methods is based on (4.27).

4.2.1 Symmetric DG Method

If p is continuous in Ω, the jump $[p]$ vanishes on each $e \in \mathcal{E}_h^o$, so

$$\sum_{e \in \mathcal{E}_h^o} (\{\!\{\mathbf{a}\nabla v \cdot \boldsymbol{\nu}\}\!\}, [p])_e = 0 \qquad \forall v \in H^2(K_h) . \tag{4.28}$$

Note that when $p \in H^1(\Omega) \cap H^2(K_h)$, (4.28) remains true. Also, the Dirichlet boundary condition can be imposed weakly:

$$(\mathbf{a}\nabla v \cdot \boldsymbol{\nu}, p)_{\Gamma_D} = (\mathbf{a}\nabla v \cdot \boldsymbol{\nu}, g_D)_{\Gamma_D} \qquad \forall v \in H^2(K_h) . \tag{4.29}$$

Then, for $p \in H^1(\Omega) \cap H^2(K_h)$ and $p = g_D$ on Γ_D, it follows from (4.28) and (4.29) that

$$J(v, p) = (\mathbf{a}\nabla v \cdot \boldsymbol{\nu}, g_D)_{\Gamma_D} \qquad \forall v \in H^2(K_h) . \tag{4.30}$$

We now define the bilinear form $a_-(\cdot, \cdot) : H^2(K_h) \times H^2(K_h) \to \mathbb{R}$ by

$$a_-(p, v) = b(p, v) - J(p, v) - J(v, p) ,$$

and the linear form $L_- : H^2(K_h) \to \mathbb{R}$ by

$$L_-(v) = L(v) - (\mathbf{a}\nabla v \cdot \boldsymbol{\nu}, g_D)_{\Gamma_D} \qquad \forall v \in H^2(K_h) . \tag{4.31}$$

The weak formulation, based on these two forms, for (4.16) is

$$a_-(p, v) = L_-(v) \qquad \forall v \in H^2(K_h) . \tag{4.32}$$

Therefore, we see that if $p \in H^2(\Omega)$ is the solution of (4.16), it satisfies (4.32). The converse is also true: If $p \in H^1(\Omega) \cap H^2(K_h)$ is a solution to (4.32), it also satisfies (4.16) (cf. Exercise 4.2).

Let V_h be a (discontinuous) finite element space associated with K_h, as in Sect. 4.1.1. The discrete analogue of (4.32) consists of finding $p_h \in V_h$ such that

$$a_-(p_h, v) = L_-(v) \qquad \forall v \in V_h . \tag{4.33}$$

This method was developed by Delves-Hall (1979) with the objective of accelerating convergence of iterative algorithms. It was called the *global element method*. Note that while $J(\cdot, \cdot)$ is non-symmetric, $a_-(\cdot, \cdot)$ is symmetric. One advantage of this method is that the linear system of algebraic equations arising from (4.33) is symmetric. On the other hand, the stiffness matrix of this system is not guaranteed to be positive semi-definite. For a time-dependent problem, this drawback may imply that some eigenvalues have negative real parts, which causes the method to be unconditionally unstable.

4.2.2 Symmetric Interior Penalty DG Method

To overcome the disadvantage of the symmetric DG method, *penalty* (*stabilization*) terms were added (Douglas-Dupont, 1976; Wheeler, 1978; Arnold, 1982). We introduce the penalty bilinear form

$$J^\theta(p,v) = \sum_{e\in\mathcal{E}_h^o} \theta_e \left([\![p]\!], [\![v]\!]\right)_e + \sum_{e\in\mathcal{E}_h^D} \theta_e \left(p,v\right)_e \ ,$$

where θ_e denotes a penalty parameter, which depends on e and the polynomial degree used in V_h; i.e., $\theta_e = \theta(h_e, r)$. Now, the bilinear form $a_-(\cdot,\cdot)$ is augmented by

$$a_-^\theta(p,v) = a_-(p,v) + J^\theta(p,v) \ .$$

The symmetric interior penalty formulation is defined by finding $p \in H^2(K_h)$ such that

$$a_-^\theta(p,v) = L_-^\theta(v) \qquad \forall v \in H^2(K_h) \ , \tag{4.34}$$

where

$$L_-^\theta(v) = L_-(v) + \sum_{e\in\mathcal{E}_h^D} \theta_e \left(g_D, v\right)_e \ .$$

The corresponding DG method is to find $p_h \in V_h$ such that

$$a_-^\theta(p_h,v) = L_-^\theta(v) \qquad \forall v \in V_h \ . \tag{4.35}$$

A similar penalty method was used by Baker (1977) for treating fourth order equations, where the interior penalty was utilized to impose (weakly) the continuity of partial derivatives across interelement boundaries on continuous finite elements. The penalty idea was introduced by Nitsche (1971) to stablize the finite element method and to improve error estimates.

We introduce the norm

$$\begin{aligned} \|v\|_h^2 &= b(v,v) + \sum_{e\in\mathcal{E}_h^o} \frac{1}{\theta_e} \left(\{\!\{\mathbf{a}\nabla v\cdot\boldsymbol{\nu}\}\!\}, \{\!\{\mathbf{a}\nabla v\cdot\boldsymbol{\nu}\}\!\}\right)_e \\ &+ J^\theta(v,v) + \sum_{e\in\mathcal{E}_h^D} \frac{1}{\theta_e} \left(\mathbf{a}\nabla v\cdot\boldsymbol{\nu}, \mathbf{a}\nabla v\cdot\boldsymbol{\nu}\right)_e , \quad v \in H^2(K_h) \ . \end{aligned} \tag{4.36}$$

With this norm, the next theorem holds (Arnold, 1982).

Theorem 4.1. *The bilinear form $a_-^\theta(\cdot,\cdot)$ is continuous in the norm $\|\cdot\|_h$:*

$$\left|a_-^\theta(v,w)\right| \le C\|v\|_h\|w\|_h \qquad \forall v,w \in H^2(K_h) \ , \tag{4.37}$$

where $0 < C \le 2$ is a constant. With the choice of the penalty parameters $\theta_e = \mathcal{C}\max\{1, r^2\}/h$, $e \in \mathcal{E}_h$, where $\mathcal{C}$ is a constant and r is the polynomial degree used in V_h, there exists a positive constant $\mathcal{C}_0$ such that for $\mathcal{C} > \mathcal{C}_0 > 0$,

$$a_-^\theta(v,v) \geq a_* \|v\|_h^2 \qquad \forall v \in V_h \ , \tag{4.38}$$

where the constant $a_* > 0$ is independent of h and r. Furthermore, with these chosen penalty parameters, if $p \in H^{r+1}(K)$, $K \in K_h$, then the following optimal error estimate holds:

$$\begin{aligned} &\|p - p_h\|_{L^2(\Omega)}^2 + h^2 \sum_{K \in K_h} \|\nabla(p - p_h)\|_{\mathbf{L}^2(K)}^2 \\ &\quad \leq C h^{2(r+1)} \sum_{K \in K_h} \|p\|_{H^{r+1}(K)}^2 \ . \end{aligned} \tag{4.39}$$

It follows from this theorem that $a_-^\theta(\cdot,\cdot)$ is coercive with respect to the norm $\|\cdot\|_h$ in the finite element space V_h. For $\mathcal{C} > \mathcal{C}_0 > 0$, using (4.38), we thus see that (4.35) has a unique solution. Moreover, the matrix corresponding to the left-hand side of (4.35) is symmetric and positive definite. Also, a stability result for p_h in terms of the data f, g_N, and g_D can be proven using (4.38) (cf. Exercise 4.6).

4.2.3 Non-Symmetric DG Method

A DG method different from (4.33) was introduced by Oden et al. (1998). We define the bilinear form $a_+(\cdot,\cdot) : H^2(K_h) \times H^2(K_h) \to \mathbb{R}$ by

$$a_+(p,v) = b(p,v) - J(p,v) + J(v,p) \ ,$$

and the linear form $L_+ : H^2(K_h) \to \mathbb{R}$ by

$$L_+(v) = L(v) + (\mathbf{a}\nabla v \cdot \boldsymbol{\nu}, g_D)_{\Gamma_D} \qquad \forall v \in H^2(K_h) \ . \tag{4.40}$$

The weak formulation for (4.16) is defined by

$$a_+(p,v) = L_+(v) \qquad \forall v \in H^2(K_h) \ . \tag{4.41}$$

The difference between (4.32) and (4.41) is just by a sign. Note that $a_+(\cdot,\cdot)$ is non-symmetric. The corresponding DG method is to find $p_h \in V_h$ such that

$$a_+(p_h,v) = L_+(v) \qquad \forall v \in V_h \ . \tag{4.42}$$

We observe that

$$a_+(v,v) = b(v,v) \geq 0 \qquad \forall v \in H^2(K_h) \ . \tag{4.43}$$

Hence we see that this bilinear form is positive semi-definite. Equation (4.43) also implies that $a_+(\cdot,\cdot)$ is coercive with respect to the norm induced by $b(\cdot,\cdot)$ (the energy seminorm). If R is strictly positive, then this energy seminorm is a norm, and existence and uniqueness of a solution to (4.42) is guaranteed.

When R equals zero, existence and uniqueness is guaranteed only when $r \geq 2$ (Rivière et al., 1999).

The next theorem can be found in Rivière et al. (1999).

Theorem 4.2. *Let the norm $\|\cdot\|_h$ be defined by* (4.36). *Then*

$$|a_+(v,w)| \leq \|v\|_h \|w\|_h \qquad \forall v, w \in H^2(K_h) \,. \tag{4.44}$$

Moreover, if $p \in H^{r+1}(K)$, $K \in K_h$, then

$$\begin{aligned} \|p - p_h\|^2_{L^2(\Omega)} + h \sum_{K \in K_h} \|\nabla(p - p_h)\|^2_{L^2(K)} \\ \leq C h^{2r+1} \sum_{K \in K_h} \|p\|^2_{H^{r+1}(K)} \,. \end{aligned} \tag{4.45}$$

Inequality (4.44) says that $a_+(\cdot,\cdot)$ is continuous with respect to the norm $\|\cdot\|_h$. This property holds for the bilinear form $a_-(\cdot,\cdot)$, too. Estimate (4.45) yields an optimal error estimate for the derivative of the solution and a "virtually" optimal estimate for the solution itself in the L^2-norm. Since $a_+(\cdot,\cdot)$ is nonsymmetric, as noted, the matrix corresponding to the left-hand side of (4.42) is also nonsymmetric.

4.2.4 Non-Symmetric Interior Penalty DG Method

As in the symmetric case, a penalty term can be also added to the bilinear form $a_+(\cdot,\cdot)$ (Rivière et al., 1999). The new bilinear and linear forms are

$$\begin{aligned} a^\theta_+(p,v) &= a_+(p,v) + J^\theta(p,v) \,, \\ L^\theta_+(v) &= L_+(v) + \sum_{e \in \mathcal{E}^D_h} \theta_e \left(g_D, v\right)_e \,. \end{aligned}$$

As a result, the non-symmetric penalty formulation consists of finding $p \in H^2(K_h)$ such that

$$a^\theta_+(p,v) = L^\theta_+(v) \qquad \forall v \in H^2(K_h) \,. \tag{4.46}$$

The discrete analogue of (4.46) is to determine $p_h \in V_h$ such that

$$a^\theta_+(p_h,v) = L^\theta_+(v) \qquad \forall v \in V_h \,. \tag{4.47}$$

Theorem 4.3. *With the definition of the norm $\|\cdot\|_h$ in* (4.36), *there is a constant $0 < C \leq 2$, independent of h and r, such that*

$$\left|a^\theta_+(v,w)\right| \leq C \|v\|_h \|w\|_h \qquad \forall v, w \in H^2(K_h) \,. \tag{4.48}$$

Furthermore, with the penalty parameters $\theta_e = \mathcal{C} \max\{1, r^2\}/h$, $e \in \mathcal{E}_h$, for any $\mathcal{C} > 0$ there is a constant $a_* > 0$, independent of h and r, such that

$$a_+^\theta(v,v) \geq a_* \|v\|_h^2 \qquad \forall v \in V_h \ . \tag{4.49}$$

Finally, the optimal error estimate in (4.39) *remains true for* (4.47).

For the non-symmetric penalty DG method, it is straightforward to see that

$$a_+^\theta(v,v) = b(v,v) + J^\theta(v,v) \qquad \forall v \in H^2(K_h) \ , \tag{4.50}$$

so (4.49) follows. The derivation of estimate (4.39) for any $\mathcal{C} > 0$ can be found in Rivière et al. (1999).

We now present numerical experiments using the DG methods for a diffusion problem. These experiments follow Hughes et al. (2000).

Example 4.4. We consider the one-dimensional diffusion problem

$$\begin{aligned} &\frac{d^2p}{dx^2} = \delta(x - x_c), \qquad x \in (0,1) \ , \\ &p(0) = p(1) = 0 \ , \end{aligned} \tag{4.51}$$

where the location x_c of the Dirac delta function δ varies within the interval $(0.6, 0.7)$. Again, as in Example 4.3, the interval $(0,1)$ is divided into ten subintervals of equal length. The approximate solutions by the nonsymmetric DG method and its penalty version with $r = 2$ are shown in Figs. 4.12–4.14. From the three figures we make the following observations:

- The nonsymmetric DG method for a pure diffusion problem is stable for $r = 2$. However, as seen in Fig. 4.12, the approximation can be quite inaccurate in some cases. Increasing the value of the penalty parameter θ leads to better approximations (cf. Figs 4.13 and 4.14).

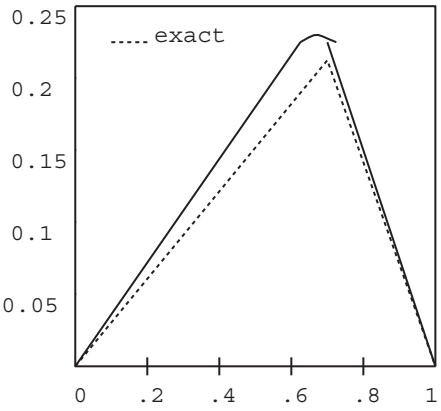

Fig. 4.12. $r = 2$, $\theta = 0$, and $x_c = 0.69$

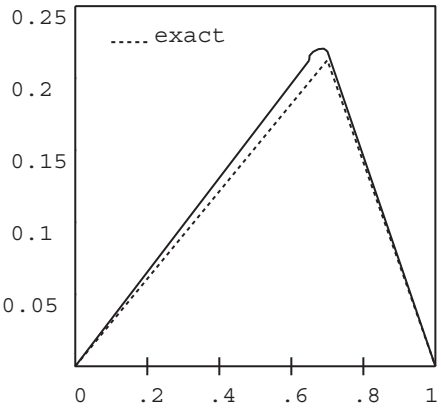

Fig. 4.13. $r = 2$, $\theta = 1/h$, and $x_c = 0.69$

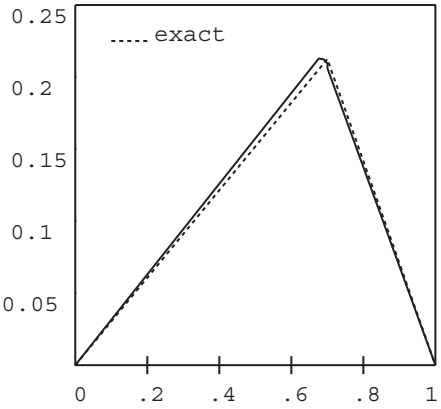

Fig. 4.14. $r = 2$, $\theta = 10/h$, and $x_c = 0.69$

- For the diffusion problem, the advantageous localization property of the DG method encountered in the advection case is lost, and an off-centered Dirac delta function in a single element causes a deteriorated approximation globally. This problem can be fixed by increasing θ.
- As discussed earlier, the symmetric penalty DG method is stable only for a sufficiently large $\mathcal{C}$ in the definition of θ. In contrast, the non-symmetric method is stable for all $\mathcal{C} > 0$. This is a desirable property, since selecting a suitable $\mathcal{C}$ may be hard without knowledge of the smallest eigenvalue of discrete problems. Note that the condition number of the stiffness matrix increases as $\mathcal{C}$ increases. Thus it is important not to choose the penalty parameter too large. We plot the smallest and largest eigenvalues of the discrete problem as a function of θ for the symmetric and non-symmetric penalty DG methods in Figs. 4.15 and 4.16, where the cases $r = 1, 2, 3$ are

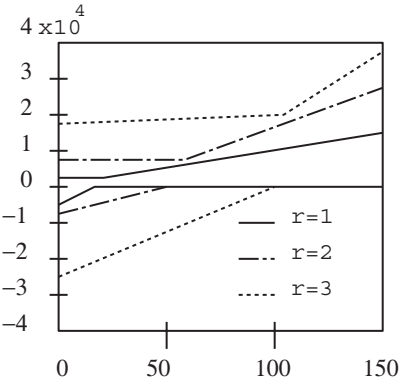

Fig. 4.15. The symmetric method

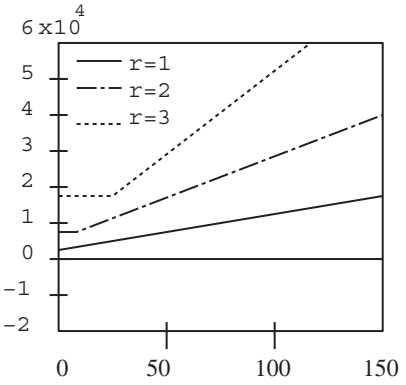

Fig. 4.16. The non-symmetric method

displayed. Note that the symmetric method is indefinite for θ too small and the critical value of θ increases with the polynomial degree. (On the scale of the graphs, the smallest eigenvalue often plots as zero, even though it is positive.)

4.2.5 Remarks

The four DG methods presented so far for the diffusion problem (4.16) are quite similar, except for a plus or minus sign in front of the term $J(v,p)$ and the addition of a penalty term $J^\theta(p,v)$ or not, but they have very different stability and convergence properties. Little can be obtained for the symmetric DG because its bilinear form is not guaranteed to be positive semi-definite. The symmetric interior penalty DG is an augmentation of the symmetric DG with a penalty term. Continuity of the bilinear form in this penalty method and coercivity in the finite element space are shown. Moreover, an

error estimate optimal with respect to h is proven. A major drawback of this method is that its stability and convergence depend on the choice of the penalty parameter (for a sufficiently large value of this parameter). The non-symmetric DG differs from the symmetric DG by a change of sign. With a strictly positive reaction coefficient, a stability result is obtained. For a pure diffusion problem, existence and uniqueness of a solution can be obtained only under the assumption that the polynomial degree used in the discrete space is greater than one. An optimal convergence rate is shown for the solution flux, while the rate deteriorates for the solution itself. The limitation of the symmetric penalty DG for a sufficiently large value of the penalty parameter is remedied by the non-symmetric penalty DG. The non-symmetric penalty formulation results in a *robust* (e.g., in terms of stability and the choice of penalty parameters) method that seems to produce the best discontinuous approximation to diffusion problems from the numerical experiments in the previous subsection. One disadvantage of this method is that its bilinear form is non-symmetric.

The number of unknowns of a discrete problem is a good indicator for the efficiency of a numerical method. The DG method can exploit a finite element space of piecewise constants, which is impossible for the continuous finite element method developed in Chap. 1. For the degrees of polynomials commonly utilized in the finite element space of the continuous method, DG seems inefficient. Following Hughes et al. (2000), Table 4.1 gives an overview of the ratio of the number of unknowns in the DG to the number of unknowns in the continuous method for different polynomial degrees r and commonly employed two- and three-dimensional geometric elements. In the case of triangles and tetrahedra, the ratio is based on regular grids obtained from subdivisions of quadrilateral and hexahedral grids, respectively. Note that, in the limit as $r \to \infty$, this ratio approaches one. That is, the DG method of very high order has a number of unknowns analogous to the corresponding continuous method.

Table 4.1. The ratio of numbers of unknowns

r	Quadrilateral	Triangle	Hexahedron	Tetrahedron
1	4	6	8	20
2	2.25	3	3.38	7.14
3	1.78	2.22	2.37	4.35
∞	1	1	1	1

4.3 Mixed Discontinuous Finite Elements

In this section, we introduce a numerical method that combines the ideas of the mixed finite element method in Chap. 3 and of the DG method in the previous sections of this chapter. As an introduction, we begin with a one-dimensional diffusion problem.

4.3.1 A One-Dimensional Problem

The one-dimensional reduction of (4.16) takes the form

$$\begin{aligned} -\frac{d}{dx}\left(a\frac{dp}{dx}\right) + Rp &= f \qquad \text{in } \Omega\,, \\ a\frac{dp}{dx}\nu &= g_N \qquad \text{on } \Gamma_N\,, \\ p &= g_D \qquad \text{on } \Gamma_D\,, \end{aligned} \tag{4.52}$$

where Ω now is a bounded interval. As in Chap. 3, to define a mixed weak formulation, we introduce the auxiliary variable

$$u = a\frac{dp}{dx} \qquad \text{in } \Omega\,, \tag{4.53}$$

so that the first equation of (4.52) becomes

$$-\frac{du}{dx} + Rp = f \qquad \text{in } \Omega\,. \tag{4.54}$$

In the next three subsections, we assume that a is strictly positive; the case where a is only nonnegative will be discussed in Sect. 4.3.1.4.

For $h > 0$, let K_h be a partition of Ω into subintervals with a maximum mesh size h. With each $e \in \mathcal{E}_h$, we associate a unit normal vector ν_e as in Sect. 4.2. For $e \in \mathcal{E}_h^b$, ν_e is just the outer unit normal to Γ; i.e., for the left end, $\nu = -1$ and for the right end, $\nu = 1$. For $e \in \mathcal{E}_h^o$, it is chosen pointing to the element with lower index (cf. Fig. 4.17); i.e., $\nu = -1$ at all interior points. This is just for notational convenience; other choices are possible. Also, for $v \in H^l(K_h)$ with $l > 1/2$, we define its average and jump at $e \in \mathcal{E}_h^o$ as follows:

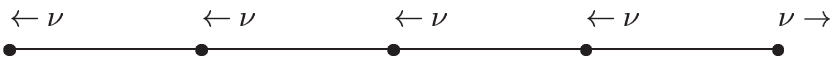

Fig. 4.17. An illustration of the unit normal vector ν

$$\{\!\{v\}\!\} = \frac{1}{2}\big((v|_{K_1})(e) + (v|_{K_2})(e)\big), \quad [\![v]\!] = (v|_{K_2})(e) - (v|_{K_1})(e)\ ,$$

where $e = \bar{K}_1 \cap \bar{K}_2$ (ν at e is exterior to K_2). For $e \in \mathcal{E}_h^b$, the convention (4.19) is used. The use of ν is for the convenience of extending the one-dimensional case to multiple dimensions.

Multiplying (4.54) by $v \in H^1(K_h)$, integrating the resulting equation on each $K \in K_h$, and using integration by parts, we see that

$$\left(u, \frac{dv}{dx}\right)_K + (Rp, v)_K - uv\big|_{\partial K} = (f, v)_K. \tag{4.55}$$

Assume that u is continuous in Ω. Then, sum (4.55) over all $K \in K_h$ and use the Neumann boundary condition in (4.52) to give

$$\begin{aligned} \sum_{K\in K_h} \left(u, \frac{dv}{dx}\right)_K + (Rp, v) - \sum_{e\in\mathcal{E}_h} u(e)\nu_e[\![v]\!](e) \\ = \sum_{e\in\Gamma_N} g_N(e)v(e) + (f, v), \qquad v \in H^1(K_h)\ . \end{aligned} \tag{4.56}$$

Similarly, invert a in (4.53), multiply by $\tau \in H^1(K_h)$, integrate on $K \in K_h$, and sum the resulting equation over all $K \in K_h$ to see that

$$\sum_{K\in K_h} \left(a^{-1}u - \frac{dp}{dx}, \tau\right)_K = 0\ . \tag{4.57}$$

With the assumption that p is continuous in Ω (so $[\![p]\!](e) = 0$, $e \in \mathcal{E}_h^o$) and the Dirichlet boundary condition in (4.52), (4.57) becomes

$$\begin{aligned} \sum_{K\in K_h} \left(a^{-1}u - \frac{dp}{dx}, \tau\right)_K + \sum_{e\in\mathcal{E}_h} [\![p]\!](e)\{\!\{\tau\nu\}\!\}(e) \\ = \sum_{e\in\Gamma_D} g_D(e)\tau(e)\nu_e, \qquad \tau \in H^1(K_h)\ . \end{aligned} \tag{4.58}$$

Equations (4.56) and (4.58) form the weak formulation on which the subsequent discontinuous methods are based. We see that if u and p is a solution of (4.53) and (4.54) with the boundary condition in (4.52), then it satisfies (4.56) and (4.58); the converse also holds if p is sufficiently smooth (e.g., $p \in H^2(\Omega)$; cf. Exercise 4.10). The term over $\mathcal{E}_h$ in the left-hand side of (4.56) is called the *consistent* term (it comes from integration by parts), while the corresponding term in (4.58) is termed the *symmetric* term (which is added).

4.3.1.1 The First Mixed Discontinuous Method

Let $V_h \times W_h$ be a pair of finite element spaces for the approximation of u and p, respectively. They are finite dimensional and defined locally on each element

$K \in K_h$, so let $V_h(K) = V_h|_K$ and $W_h(K) = W_h|_K$. Neither continuity constraint nor boundary data are imposed on $V_h \times W_h$.

The first mixed discontinuous method for (2.1) is: Find $u_h \in V_h$ and $p_h \in W_h$ such that

$$\begin{aligned}
\sum_{K\in K_h} \left(u_h, \frac{dv}{dx}\right)_K + (Rp_h, v) - \sum_{e\in\mathcal{E}_h} \{\!\{u_h\nu\}\!\}(e)[\![v]\!](e) \\
= \sum_{e\in\Gamma_N} g_N(e)v(e) + (f, v), \quad v \in W_h \ , \\
\sum_{K\in K_h} \left(a^{-1}u_h - \frac{dp_h}{dx}, \tau\right)_K + \sum_{e\in\mathcal{E}_h} [\![p_h]\!](e)\{\!\{\tau\nu\}\!\}(e) \\
= \sum_{e\in\Gamma_D} g_D(e)\tau(e)\nu_e, \quad \tau \in V_h \ .
\end{aligned} \tag{4.59}$$

With the definition of the bilinear forms

$$\begin{aligned}
A(u_h, \tau) &= \sum_{K\in K_h} (a^{-1}u_h, \tau)_K \ , \\
B(\tau, v) &= \sum_{K\in K_h} \left(\tau, \frac{dv}{dx}\right)_T - \sum_{e\in\mathcal{E}_h} \{\!\{\tau\nu\}\!\}(e)[\![v]\!](e) \ ,
\end{aligned}$$

system (4.59) is of the form

$$\begin{aligned}
B(u_h, v) + (Rp_h, v) &= \sum_{e\in\Gamma_N} g_N(e)v(e) + (f, v), \quad v \in W_h \ , \\
A(u_h, \tau) - B(\tau, p_h) &= \sum_{e\in\Gamma_D} g_D(e)\tau(e)\nu_e, \qquad\qquad \tau \in V_h \ .
\end{aligned} \tag{4.60}$$

If we take $v = p_h$ and $\tau = u_h$ in (4.60) and add the two equations, the left-hand side of the resulting sum is

$$A(u_h, u_h) + (Rp_h, p_h) = \|a^{-1/2}u_h\|^2_{L^2(\Omega)} + \|R^{1/2}p_h\|^2_{L^2(\Omega)}. \tag{4.61}$$

Thus uniqueness of u_h follows. If R is strictly positive, uniqueness of p_h also follows from (4.61). Therefore, existence and uniqueness of a solution to (4.59) is shown. Note that the system corresponding to the left-hand side of (4.59) is symmetric after changing a sign in either of the two equations. But this would alter the property (4.61). As seen in the subsequent analysis, if (4.59) is written in nonmixed form (or the standard Galerkin version), positive definiteness and symmetry can be preserved simultaneously. Thus, in terms of implementation, it is desirable to write (4.59) in nonmixed form. However, we emphasize that the mixed formulation naturally stabilizes the discontinuous finite element method; see the discussion at the end of this

subsection. The case $R \equiv 0$ is more complicated; existence and uniqueness of a solution depends on the type of boundary conditions used in (4.52) (Chen-Chen, 2003).

The proof of Theorems 4.4–4.6 can be found in Chen et al. (2003A).

Theorem 4.4. *With the choice*

$$V_h(K) = W_h(K) = P_r(K), \qquad K \in K_h, \quad r \geq 0 \,, \tag{4.62}$$

it holds that

$$\begin{aligned} &\|p - p_h\|^2_{L^2(\Omega)} + \|u - u_h\|^2_{L^2(\Omega)} \\ &\quad \leq Ch^{2r} \left(\sum_{K \in K_h} \|p\|^2_{H^{r+1}(K)} + \sum_{K \in K_h} \|u\|^2_{H^{r+1}(K)} \right) . \end{aligned} \tag{4.63}$$

For r even, an optimal order in h of convergence occurs:

$$\begin{aligned} &\|p - p_h\|^2_{L^2(\Omega)} + \|u - u_h\|^2_{L^2(\Omega)} \\ &\quad \leq Ch^{2(r+1)} \left(\sum_{K \in K_h} \|p\|^2_{H^{r+1}(K)} + \sum_{K \in K_h} \|u\|^2_{H^{r+1}(K)} \right) , \end{aligned} \tag{4.64}$$

provided K_h is a uniform partition.

Note that estimate (4.63) gives a suboptimal order in h of convergence. For an odd r, it is sharp for (4.59).

While method (4.59) is in mixed form, it can be implemented (if desired) in nonmixed form. We introduce the coefficient-dependent $L^2(\Omega)$-projection $P_h : L^2(\Omega) \to V_h$ by

$$\left(a^{-1}(w - P_h w), \tau\right) = 0 \,, \qquad \forall \tau \in V_h \,, \tag{4.65}$$

for $w \in L^2(\Omega)$, and the operator $R_h : H^1(K_h) \to V_h$ by

$$\begin{aligned} \sum_{K \in K_h} (a^{-1} R_h(v), \tau)_K = &- \sum_{e \in \mathcal{E}_h} [v](e) \{\tau \nu\}(e) \\ &+ \sum_{e \in \Gamma_D} g_D(e) \tau(e) \nu_e, \qquad \tau \in V_h \,, \end{aligned} \tag{4.66}$$

for $v \in H^1(K_h)$. Note that R_h depends on g_D; for notational convenience, we omit this dependence. Using (4.65) and (4.66), (4.59) can be rewritten as follows: Find $p_h \in W_h$ such that

$$\begin{aligned}
&\sum_{K\in K_h}\left(P_h\left(a\frac{dp_h}{dx}\right),\frac{dv}{dx}\right)_K+(Rp_h,v)\\
&\quad-\sum_{e\in\mathcal{E}_h}[p_h](e)\left\{\!\!\left\{P_h\left(a\frac{dv}{dx}\right)\nu\right\}\!\!\right\}(e)\\
&\quad-\sum_{e\in\mathcal{E}_h}\left\{\!\!\left\{\left(P_h\left(a\frac{dp_h}{dx}\right)+R_h(p_h)\right)\nu\right\}\!\!\right\}(e)[v](e)\\
&\qquad=-\sum_{e\in\Gamma_D}g_D(e)P_h\left(a\frac{dv}{dx}\right)(e)\nu_e+\sum_{e\in\Gamma_N}g_N(e)v(e)\\
&\qquad\quad+(f,v),\qquad v\in W_h\ ,
\end{aligned}\tag{4.67}$$

with u_h given by

$$u_h=P_h\left(a\frac{dp_h}{dx}\right)+R_h(p_h).\tag{4.68}$$

To see the relationship between (4.67) and the DG methods in the previous section, we consider the case where a is piecewise constant. In this case, (4.67) becomes: Find $p_h\in W_h$ satisfying

$$\begin{aligned}
&\sum_{K\in K_h}\left(a\frac{dp_h}{dx},\frac{dv}{dx}\right)_K+(Ru_h,v)-\sum_{e\in\mathcal{E}_h}[p_h](e)\left\{\!\!\left\{a\frac{dv}{dx}\nu\right\}\!\!\right\}(e)\\
&\quad-\sum_{e\in\mathcal{E}_h}\left\{\!\!\left\{a\frac{dp_h}{dx}\nu\right\}\!\!\right\}(e)[v](e)+\sum_{K\in K_h}\left(a^{-1}R_h(p_h),R_h(v)\right)_K\\
&\quad=-\sum_{e\in\Gamma_D}g_D(e)\left(\left(a\frac{dv}{dx}\right)(e)-R_h(p_h)(e)\right)\nu_e\\
&\qquad+\sum_{e\in\Gamma_N}g_N(e)v(e)+(f,v),\quad v\in W_h\ .
\end{aligned}\tag{4.69}$$

Observe that without the term involving R_h, (4.69) is just the symmetric DG method introduced in Sect. 4.2.1. With a positive sign in front of the fourth term in the left-hand side of (4.69) (and without the R_h term), it is the non-symmetric method in Sect. 4.2.3. The R_h term naturally comes from the mixed formulation, and stabilizes the DG methods in the previous section. Although R_h appears, equation (4.69) can be evaluated virtually in almost the same amount of work as in the evaluation of the DG methods in the previous section. This is due to the definition of R_h in (4.66), where the matrix associated with the left-hand side can be diagonal if the basis functions of V_h are appropriately chosen. In addition, (4.66) is defined on each element and is thus totally local. The equations of this type can be implemented in a parallel fashion.

4.3.1.2 The Second Mixed Discontinuous Method

Let $\Omega = (l_1, l_2)$, and for $v \in H^1(K_h)$, we define the one-sided limits at nodes $e \in \mathcal{E}_h^o$:

$$v(e^+) = \lim_{x \to e^+} v(x), \quad v(e^-) = \lim_{x \to e^-} v(x) \ .$$

At the endpoints l_1 and l_2, the limits are defined from inside Ω. We now define the second mixed discontinuous method for (4.52): Find $u_h \in V_h$ and $p_h \in W_h$ such that

$$\begin{aligned} \sum_{K \in K_h} \left(u_h, \frac{dv}{dx} \right)_K + (Rp_h, v) - \sum_{e \in \mathcal{E}_h} u_h(e^+) \nu_e [\![v]\!](e) \\ = \sum_{e \in \Gamma_N} g_N(e) v(e) + (f, v), \quad v \in W_h \ , \\ \sum_{K \in K_h} \left(a^{-1} u_h - \frac{dp_h}{dx}, \tau \right)_K + \sum_{e \in \mathcal{E}_h} [\![p_h]\!](e) \tau(e^+) \nu_e \\ = \sum_{e \in \Gamma_D} g_D(e) \tau(e) \nu_e, \quad \tau \in V_h \ . \end{aligned} \tag{4.70}$$

Note that the second method differs from the first one in that the averaged quantities in (4.59) are replaced by the right-hand sided limits. For (4.70), existence and uniqueness of a solution and convergence can be shown in a similar way as for (4.59). In particular, the convergence result (4.63) holds for (4.70) (Chen et al., 2003A). For the present method, if the Dirichlet and Neumann boundary conditions occur at $x = l_1$ and $x = l_2$, respectively, we are able to obtain the optimal convergence rate (4.64), no matter whether r is even or odd. To improve the convergence rate for other boundary conditions, we can adopt the penalty idea in the previous section. With this idea, (4.70) is modified as follows: Find $u_h \in V_h$ and $p_h \in W_h$ such that

$$\begin{aligned} \sum_{K \in K_h} \left(u_h, \frac{dv}{dx} \right)_K + (Rp_h, v) - \sum_{e \in \mathcal{E}_h} u_h(e^+) \nu_e [\![v]\!](e) \\ + \theta_{l_2} p_h(l_2) v(l_2) = \theta_{l_2} g_D(l_2) v(l_2) + \sum_{e \in \Gamma_N} g_N(e) v(e) + (f, v), \quad v \in W_h \ , \\ \sum_{K \in K_h} \left(a^{-1} u_h - \frac{dp_h}{dx}, \tau \right)_K + \sum_{e \in \mathcal{E}_h} [\![p_h]\!](e) \tau(e^+) \nu_e \\ + \theta_{l_1} u_h(l_1) \tau(l_1) = \theta_{l_1} g_N(l_1) \tau(l_1) + \sum_{e \in \Gamma_D} g_D(e) \tau(e) \nu_e \ , \quad \tau \in V_h \ , \end{aligned} \tag{4.71}$$

where $\theta_{l_1} \geq 0$ and $\theta_{l_2} \geq 0$ are penalty parameters. Note that we only penalize at the endpoints.

Theorem 4.5. *If the following choices are made for the penalty parameters in* (4.71):

- $\theta_{l_1} = \theta_{l_2} = 0$ *if the Dirichlet and Neumann boundary conditions are imposed at* $x = l_1$ *and* $x = l_2$*, respectively,*
- $\theta_{l_1} = \mathcal{C}\max(1, r)/h$ *and* $\theta_{l_2} = 0$ *if the Neumann boundary conditions are imposed at both* $x = l_1$ *and* $x = l_2$*,*
- $\theta_{l_1} = 0$ *and* $\theta_{l_2} = \mathcal{C}\max(1, r)/h$ *if the Dirichlet boundary conditions are imposed at both* $x = l_1$ *and* $x = l_2$*,*
- *or* $\theta_{l_1} = \theta_{l_2} = \mathcal{C}\max(1, r)/h$ *if the Neumann and Dirichlet boundary conditions are imposed at* $x = l_1$ *and* $x = l_2$*, respectively,*

then the optimal error estimate (4.64) *with any* $r \geq 0$ *holds for all* $\mathcal{C} > 0$.

We mention that (4.70) and (4.71) can be also written in nonmixed form as in (4.67) (Chen et al., 2003A).

4.3.1.3 The Third Mixed Discontinuous Method

The third method is analogous to the second one. We recall that the averaged quantities in (4.59) were replaced by the right-hand sided limits in (4.71). In the third method, they are replaced by the left-hand sided limits. That is, the third mixed discontinuous method for (4.52) is: Find $u_h \in V_h$ and $p_h \in W_h$ such that

$$\begin{aligned}
&\sum_{K\in K_h}\left(u_h, \frac{dv}{dx}\right)_K + (Rp_h, v) - \sum_{e\in\mathcal{E}_h} u_h(e^-)\nu_e[v](e) \\
&\quad + \theta_{l_1} p_h(l_1)v(l_1) = \theta_{l_1} g_D(l_1)v(l_1) + \sum_{e\in\Gamma_N} g_N(e)v(e) + (f, v), \qquad v \in W_h\ , \\
&\sum_{K\in K_h}\left(a^{-1}u_h - \frac{dp_h}{dx}, \tau\right)_K + \sum_{e\in\mathcal{E}_h} [p_h](e)\tau(e^-)\nu_e \\
&\quad + \theta_{l_2} u_h(l_2)\tau(l_2) = \theta_{l_2} g_N(l_2)\tau(l_2) + \sum_{e\in\Gamma_D} g_D(e)\tau(e)\nu_e\ , \quad \tau \in V_h\ .
\end{aligned} \tag{4.72}$$

Again, we only penalize at the endpoints. Existence and uniqueness of a solution to (4.72) can be proven as for (4.59). As for convergence, the next theorem holds.

Theorem 4.6. *If the following choices for the penalty parameters in* (4.72) *are made:*

- $\theta_{l_1} = \theta_{l_2} = 0$ *if the Neumann and Dirichlet boundary conditions are imposed at* $x = l_1$ *and* $x = l_2$*, respectively,*
- $\theta_{l_1} = 0$ *and* $\theta_{l_2} = \mathcal{C}\max(1, r)/h$ *if the Neumann boundary conditions are imposed at both* $x = l_1$ *and* $x = l_2$*,*
- $\theta_{l_1} = \mathcal{C}\max(1, r)/h$ *and* $\theta_{l_2} = 0$ *if the Dirichlet boundary conditions are imposed at both* $x = l_1$ *and* $x = l_2$*,*

- *or* $\theta_{l_1} = \theta_{l_2} = \mathcal{C}\max(1, r)/h$ *if the Dirichlet and Neumann boundary conditions are imposed at* $x = l_1$ *and* $x = l_2$*, respectively,*

then the error bound (4.64) *with any* $r \geq 0$ *holds for* (4.72) *for all* $\mathcal{C} > 0$.

System (4.72) can be implemented in nonmixed form as for the first and second methods (Chen et al., 2003A).

Example 4.5. We present a numerical example for the mixed discontinuous methods discussed in this section. We solve (4.52) with $a = R = 1$ on the unit interval $(0, 1)$, with Dirichlet conditions at both endpoints. The first method (4.59) is used on uniform grids, and the estimates $p - p_h$ and $u - u_h$ in the L^2-norm for polynomials of degree one to six for successively refined grids are shown in Tables 4.2 and 4.3. From these two tables we see that the convergence is of order h^r and h^{r+1} for odd r and even r, respectively. This agrees with the theoretical results in (4.63) and (4.64).

We now solve the same problem using the second mixed discontinuous method (4.70). The case where the Dirichlet and Neumann boundary conditions occur at $x = 0$ and $x = 1$, respectively, is first experimented. The estimates $p - p_h$ and $u - u_h$ in the L^2-norm are given in Tables 4.4 and 4.5. Note that error estimates of optimal order h^{r+1} in h are shown for both. This agrees with the discussion in Sect. 4.3.1.2.

Table 4.2. The estimates in h for p_h for the first method

$1/h$	16		32		64	
r	Error	Order	Error	Order	Error	Order
1	3.0095e-02	1.05	1.4959e-02	1.00	7.4807e-03	1.00
2	2.6498e-04	3.09	3.2652e-05	3.02	4.0695e-06	3.00
3	2.8625e-05	3.03	3.5602e-06	3.00	4.4458e-07	3.00
4	1.2424e-07	5.09	3.8248e-09	5.02	1.1909e-10	5.01
5	9.4656e-09	5.03	2.9432e-10	5.01	9.2076e-12	5.00
6	2.8156e-11	7.10	2.1064e-13	7.06	1.6334e-15	7.01

Table 4.3. The estimates in h for u_h for the first method

$1/h$	16		32		64	
r	Error	Order	Error	Order	Error	Order
1	1.7101e-01	1.01	8.5382e-02	1.00	4.2677e-02	1.00
2	8.6296e-04	3.08	1.0645e-04	3.02	1.3262e-05	3.00
3	1.2269e-04	3.02	1.5290e-05	3.00	1.9098e-06	3.00
4	4.2128e-07	5.08	1.2989e-08	5.02	4.0452e-10	5.00
5	3.8028e-08	5.02	1.1840e-09	5.01	3.7233e-11	5.00
6	9.7283e-11	7.08	7.5150e-13	7.02	5.8745e-15	7.00

Table 4.4. The estimates in h for p_h for the second method

$1/h$	16		32		64	
r	Error	Order	Error	Order	Error	Order
1	6.6299e-03	2.00	1.6587e-03	2.00	4.1474e-04	2.00
2	2.0847e-04	2.99	2.6104e-05	3.00	3.2644e-06	3.00
3	5.0346e-06	3.99	3.1512e-07	4.00	1.9702e-08	4.00
4	9.7993e-08	4.99	3.0660e-09	5.00	9.6280e-10	5.00
5	1.5947e-09	5.99	2.5175e-11	5.99	3.9321e-13	6.00
6	2.2331e-11	6.99	1.7444e-13	7.00	1.3660e-15	7.00

Table 4.5. The estimates in h for u_h for the second method

$1/h$	16		32		64	
r	Error	Order	Error	Order	Error	Order
1	4.1575e-02	1.99	1.0417e-02	2.00	2.6056e-03	2.00
2	1.3100e-03	2.99	1.6402e-04	3.00	2.0511e-05	3.00
3	3.1639e-05	3.99	1.9800e-06	4.00	1.2379e-07	4.00
4	6.1579e-07	4.99	1.9265e-08	5.00	6.0241e-10	5.00
5	1.0021e-08	5.99	1.5682e-10	6.00	2.4491e-12	6.00
6	1.4002e-10	6.99	1.0962e-12	7.00	8.5470e-15	7.00

Table 4.6. The estimates in h for p_h

$1/h$	16		32		64	
r	Error	Order	Error	Order	Error	Order
1	1.0931e-00	1.40	3.9329e-01	1.47	1.3965e-01	1.49
2	1.1037e-02	3.46	9.8249e-04	3.49	8.7021e-05	3.50
3	1.8424e-03	3.41	1.6551e-04	3.48	1.4688e-05	3.49
4	9.0700e-06	5.46	2.0167e-07	5.50	4.4637e-09	5.50
5	9.0002e-07	5.41	2.0201e-08	5.47	4.4533e-10	5.50
6	2.9402e-09	7.47	1.6639e-11	7.47	9.2451e-14	7.50

The case where the Dirichlet boundary conditions occur at both $x = 0$ and $x = 1$ is next experimented. The estimates $p - p_h$ in the L^2-norm for different values of r are displayed in Table 4.6. These estimates are asymptotically of order $h^{r+1/2}$ for odd r and $h^{r+1+1/2}$ for even r. They are better than those obtained in Sect. 4.3.1.2. We emphasize that the convergence rates are not optimal in h for odd r. Similar numerical results are observed for the variable u_h.

4.3.1.4 A Generalization

We have assumed so far that a is strictly positive. In this subsection, we consider the case where a is only nonnegative. Let $\kappa \geq 0$ satisfy

$$a = \kappa^2. \tag{4.73}$$

Then (4.52) is expressed in the form

$$\begin{aligned} &-\frac{d(\kappa u)}{dx} + Rp = f, \quad u = \kappa \frac{dp}{dx} && \text{in } \Omega\ , \\ &p = g_D && \text{on } \Gamma_D\ , \\ &\kappa u \nu = g_N && \text{on } \Gamma_N\ . \end{aligned} \tag{4.74}$$

The corresponding weak formulation is defined as follows:

$$\begin{aligned} &\sum_{K \in K_h} \left(\kappa u, \frac{dv}{dx} \right)_K + (Rp, v) - \sum_{e \in \mathcal{E}_h} \{\!\{\kappa u \nu\}\!\}(e)[\![v]\!](e) \\ &\qquad = \sum_{e \in \Gamma_N} g_N(e) v(e) + (f, v), \qquad v \in H^1(K_h)\ , \\ &\sum_{K \in K_h} \left(u - \kappa \frac{dp}{dx}, \tau \right)_K + \sum_{e \in \mathcal{E}_h} [\![p]\!](e) \{\!\{\kappa \tau \nu\}\!\}(e) \\ &\qquad = \sum_{e \in \Gamma_D} g_D(e) \kappa \tau(e) \nu_e, \qquad \tau \in H^1(K_h)\ . \end{aligned} \tag{4.75}$$

Based on (4.75), the three mixed discontinuous methods in the previous subsections can be defined accordingly. Moreover, similar stability and convergence results hold (see the next subsection).

4.3.2 Multi-Dimensional Problems

We now extend the mixed discontinuous methods in the previous subsection to multi-dimensional problems. As an example, we develop the first method for the problem

$$\begin{aligned} &-\nabla \cdot \big(\mathbf{a} \big(\nabla p + \mathbf{b} p \big) \big) + Rp = f, && \mathbf{x} \in \Omega\ , \\ &\mathbf{a} \left(\nabla p + \mathbf{b} p \right) \cdot \boldsymbol{\nu} = g_N, && \mathbf{x} \in \Gamma_N\ , \\ &p = g_D, && \mathbf{x} \in \Gamma_D\ , \end{aligned} \tag{4.76}$$

where $\Omega \subset \mathbb{R}^d$ $(d \leq 3)$ is a bounded domain, $\mathbf{a}$ is a bounded, symmetric, and positive *semi-definite* tensor, $\mathbf{b}$ and $R \geq 0$ are bounded functions, $f \in L^2(\Omega)$,

$g_D \in H^{1/2}(\Gamma_D)$, and $g_N \in H^{-1/2}(\Gamma_N)$. To be general, note that an advection term is included in (4.76). We assume that (4.76) has a unique solution.

Under the present assumption on $\mathbf{a}$, formulation (4.75) is utilized. Since $\mathbf{a}$ is a symmetric, positive semi-definite tensor, there is a tensor $\boldsymbol{\kappa}$, which has the same property as $\mathbf{a}$, such that

$$\mathbf{a} = \boldsymbol{\kappa}\boldsymbol{\kappa} \, . \tag{4.77}$$

With this splitting, the first equation of (4.76) can be written as

$$\begin{aligned} -\nabla \cdot (\boldsymbol{\kappa}\mathbf{u}) + Rp &= f, \qquad \mathbf{x} \in \Omega \, , \\ \mathbf{u} &= \boldsymbol{\kappa} \left(\nabla p + \mathbf{b} p \right) , \qquad \mathbf{x} \in \Omega \, . \end{aligned} \tag{4.78}$$

Although the variable $\mathbf{u}$ may not have a physical meaning in applications, the variable $\boldsymbol{\kappa}\mathbf{u}$ does have a physical meaning, such as a fluid velocity. Once $\mathbf{u}$ is computed, the latter variable can be obtained in a free manner.

The following weak formulation for (4.78), with the boundary conditions in (4.76), can be derived as in (4.56) and (4.58):

$$\begin{aligned} \sum_{K \in K_h} (\boldsymbol{\kappa}\mathbf{u}, \nabla v)_K + (Rp, v) - \sum_{e \in \mathcal{E}_h} \left(\{\!\{ \boldsymbol{\kappa}\mathbf{u} \cdot \boldsymbol{\nu} \}\!\}, [v] \right)_e \\ = \sum_{e \in \mathcal{E}_h^N} (g_N, v)_e + (f, v), \qquad v \in H^1(K_h) \, , \\ \sum_{K \in K_h} \left(\mathbf{u} - (\boldsymbol{\kappa} \nabla p + \bar{\mathbf{b}} p), \boldsymbol{\tau} \right)_K + \sum_{e \in \mathcal{E}_h} \left([p], \{\!\{ \boldsymbol{\kappa}\boldsymbol{\tau} \cdot \boldsymbol{\nu} \}\!\} \right)_e \\ = \sum_{e \in \mathcal{E}_h^D} (g_D, \boldsymbol{\kappa}\boldsymbol{\tau} \cdot \boldsymbol{\nu})_e, \qquad \boldsymbol{\tau} \in \mathbf{H}^1(K_h) \, , \end{aligned} \tag{4.79}$$

where $\bar{\mathbf{b}} = \boldsymbol{\kappa}\mathbf{b}$ and we assume that $\boldsymbol{\kappa}\mathbf{u} \cdot \boldsymbol{\nu}$ and p are continuous across interelement boundaries.

For $h > 0$, let K_h be a finite element partition of Ω into elements $\{K\}$, as in Sect. 4.1.1, and let $\mathbf{V}_h \times W_h$ be a pair of finite element spaces for approximating $\mathbf{u}$ and p, respectively, associated with K_h. With the same notation as in the previous section, the first mixed discontinuous method for (4.76) is: Find $\mathbf{u}_h \in \mathbf{V}_h$ and $p_h \in W_h$ such that

$$\begin{aligned} \sum_{K \in K_h} (\boldsymbol{\kappa}\mathbf{u}_h, \nabla v)_K + (Rp_h, v) - \sum_{e \in \mathcal{E}_h} \left(\{\!\{ \boldsymbol{\kappa}\mathbf{u}_h \cdot \boldsymbol{\nu} \}\!\}, [v] \right)_e \\ = \sum_{e \in \mathcal{E}_h^N} (g_N, v)_e + (f, v), \qquad v \in W_h \, , \\ \sum_{K \in K_h} \left(\mathbf{u}_h - (\boldsymbol{\kappa} \nabla p_h + \bar{\mathbf{b}} p_h), \boldsymbol{\tau} \right)_K + \sum_{e \in \mathcal{E}_h} \left([p_h], \{\!\{ \boldsymbol{\kappa}\boldsymbol{\tau} \cdot \boldsymbol{\nu} \}\!\} \right)_e \\ = \sum_{e \in \mathcal{E}_h^D} (g_D, \boldsymbol{\kappa}\boldsymbol{\tau} \cdot \boldsymbol{\nu})_e, \qquad \boldsymbol{\tau} \in \mathbf{V}_h \, . \end{aligned} \tag{4.80}$$

We make the assumption

$$(\boldsymbol{\tau},\boldsymbol{\tau}) - (\bar{\mathbf{b}}v,\boldsymbol{\tau}) + (Rv,v) \geq C_1 \left(\|\boldsymbol{\tau}\|^2_{\mathbf{L}^2(\Omega)} + (Rv,v) \right) \quad \forall \boldsymbol{\tau} \in \mathbf{L}^2(\Omega),\ v \in L^2(\Omega)\ , \tag{4.81}$$

where C_1 is a positive constant. Inequality (4.81) immediately implies that if $R = 0$ a.e. on Ω, then $\bar{\mathbf{b}} = \mathbf{0}$ a.e. on Ω. With this assumption, we can check existence and uniqueness of a solution to (4.80). Setting $f = g_N = g_D = 0$ in (4.80), it follows from the choices of $v = p_h$ and $\boldsymbol{\tau} = \mathbf{u}_h$ and the addition of the two equations in (4.80) that

$$(\mathbf{u}_h,\mathbf{u}_h) - (\bar{\mathbf{b}}p_h,\mathbf{u}_h) + (Rp_h,p_h) = 0\ ,$$

which, by (4.81), immediately implies uniqueness of $\mathbf{u}_h$. Also, by (4.81), if R is strictly positive, uniqueness of p_h follows. Therefore, existence and uniqueness of a solution to (4.80) is shown. If $R \equiv 0$, the reader should refer to Chen-Chen (2003) on existence and uniqueness of a solution.

Theorem 4.7. *With the choice of the finite element spaces V_h and W_h*

$$\mathbf{V}_h(K) = (P_r(K))^d\ , \quad W_h(K) = P_r(K), \quad r \geq 0\ ,$$

the following error estimate holds for (4.80):

$$\begin{aligned} &\|p - p_h\|^2_{L^2(\Omega)} + \|\mathbf{u} - \mathbf{u}_h\|^2_{\mathbf{L}^2(\Omega)} \\ &\quad \leq Ch^{2r} \left(\sum_{K \in K_h} \|p\|^2_{H^{r+1}(K)} + \sum_{K \in K_h} \|\mathbf{u}\|^2_{\mathbf{H}^{r+1}(K)} \right)\ . \end{aligned} \tag{4.82}$$

Observe that estimate (4.82) gives a suboptimal order of convergence in h. The next theorem gives an optimal convergence result.

Theorem 4.8. *Assume that Ω is a rectangular domain, K_h is a Cartesian product of uniform grids in each of the coordinate directions, and*

$$\mathbf{V}_h(K) = (Q_r(K))^d\ , \quad W_h(K) = Q_r(K), \quad r \geq 0\ ,$$

where $Q_r(K)$ is the space of tensor products of one-dimensional polynomials of degree r on K. Then, if r is even,

$$\begin{aligned} &\|p - p_h\|^2_{L^2(\Omega)} + \|\mathbf{u} - \mathbf{u}_h\|^2_{\mathbf{L}^2(\Omega)} \\ &\quad \leq Ch^{2(r+1)} \left(\sum_{K \in K_h} \|p\|^2_{H^{r+1}(K)} + \sum_{K \in K_h} \|\mathbf{u}\|^2_{\mathbf{H}^{r+1}(K)} \right)\ . \end{aligned} \tag{4.83}$$

The proof of Theorems 4.7 and 4.8 can be found in Chen (2001A). As in Sect. 4.3.1.1, system (4.80) can be also implemented in nonmixed form. Let $\mathbf{P}_h : \mathbf{L}^2(\Omega) \to \mathbf{V}_h$ denote the L^2-projection:

$$(\mathbf{w} - \mathbf{P}_h\mathbf{w}, \boldsymbol{\tau}) = 0, \qquad \forall \boldsymbol{\tau} \in \mathbf{V}_h \,, \tag{4.84}$$

for $\mathbf{w} \in \mathbf{L}^2(\Omega)$. Also, we define the operator $\mathbf{R}_h : H^1(K_h) \to \mathbf{V}_h$ by

$$\begin{aligned} \sum_{K \in K_h} (\mathbf{R}_h(v), \boldsymbol{\tau})_K = & - \sum_{e \in \mathcal{E}_h} \left([\![v]\!], \{\!\{\boldsymbol{\kappa}\boldsymbol{\tau} \cdot \boldsymbol{\nu}\}\!\}\right)_e \\ & + \sum_{e \in \mathcal{E}_h^D} \left(g_D, \boldsymbol{\kappa}\boldsymbol{\tau} \cdot \boldsymbol{\nu}\right)_e , \qquad \boldsymbol{\tau} \in \mathbf{V}_h \,, \end{aligned} \tag{4.85}$$

for $v \in H^1(K_h)$. With these two operators, system (4.80) is rewritten: Find $p_h \in W_h$ such that

$$\begin{aligned} & \sum_{K \in K_h} \big(\mathbf{P}_h\big(\boldsymbol{\kappa}(\nabla p_h + \mathbf{b}p_h)\big), \boldsymbol{\kappa}\nabla v\big)_K + (Rp_h, v) \\ & - \sum_{e \in \mathcal{E}_h} \left([\![p_h]\!], \{\!\{\boldsymbol{\kappa}\mathbf{P}_h(\boldsymbol{\kappa}\nabla v) \cdot \boldsymbol{\nu}\}\!\}\right)_e \\ & - \sum_{e \in \mathcal{E}_h} \left(\{\!\{\boldsymbol{\kappa}\left(\mathbf{P}_h\big(\boldsymbol{\kappa}(\nabla p_h + \mathbf{b}p_h)\big) + \mathbf{R}_h(p_h)\right) \cdot \boldsymbol{\nu}\}\!\}, [\![v]\!]\right)_e \\ & = - \sum_{e \in \mathcal{E}_h^D} \left(g_D, \boldsymbol{\kappa}\mathbf{P}_h(\boldsymbol{\kappa}\nabla v) \cdot \boldsymbol{\nu}\right)_e + \sum_{e \in \mathcal{E}_h^N} (g_N, v)_e + (f, v), \quad v \in W_h \,, \end{aligned}$$

with $\mathbf{u}_h$ given by

$$\mathbf{u}_h = \mathbf{P}_h\big(\boldsymbol{\kappa}(\nabla p_h + \mathbf{b}p_h)\big) + \mathbf{R}_h(p_h) \,.$$

We remark that if $\mathbf{a}$ is positive definite, the first equation of (4.76) can be written as

$$\begin{aligned} -\nabla \cdot \mathbf{u} + Rp &= f, \qquad \mathbf{x} \in \Omega \,, \\ \mathbf{a}^{-1}\mathbf{u} &= \nabla p + \mathbf{b}p, \qquad \mathbf{x} \in \Omega \,. \end{aligned} \tag{4.86}$$

Then a mixed discontinuous method similar to (4.59) can be considered, and analogous results as those for (4.80) hold (Chen, 2001A).

4.3.3 Nonlinear Problems

We consider the mixed discontinuous methods for a nonlinear problem:

$$\begin{aligned} & \frac{\partial p}{\partial t} - \nabla \cdot \big(\mathbf{a}(p)\big(\nabla p + \mathbf{b}(p)\big)\big) = f(p), && \mathbf{x} \in \Omega \,, \\ & \mathbf{a}(p)\left(\nabla p + \mathbf{b}(p)\right) \cdot \boldsymbol{\nu} = g_N, && \mathbf{x} \in \Gamma_N \,, \\ & p = g_D, && \mathbf{x} \in \Gamma_D \,, \end{aligned} \tag{4.87}$$

where the coefficients $\mathbf{a}$, $\mathbf{b}$, and f depend on the solution itself. These coefficients are assumed to be Lipschitz continuous in p.

As in (4.77), if the tensor $\mathbf{a}$ is assumed to be symmetric, positive semi-definite, there exists a symmetric tensor $\boldsymbol{\kappa}$ such that

$$\mathbf{a}(p) = \boldsymbol{\kappa}(p)\boldsymbol{\kappa}(p) \; . \tag{4.88}$$

Using (4.88), the first equation of (4.87) can be written as

$$\begin{aligned} &\frac{\partial p}{\partial t} - \nabla \cdot (\boldsymbol{\kappa}(p)\mathbf{u}) = f(p), \qquad \mathbf{x} \in \Omega \; , \\ &\mathbf{u} = \boldsymbol{\kappa}(p) \left(\nabla p + \mathbf{b}(p)\right), \qquad \mathbf{x} \in \Omega \; . \end{aligned} \tag{4.89}$$

A mixed weak formulation for (4.89), with the boundary conditions in (4.87), is given by

$$\begin{aligned} &\left(\frac{\partial p}{\partial t}, v\right) + \sum_{K \in K_h} (\boldsymbol{\kappa}(p)\mathbf{u}, \nabla v)_K - \sum_{e \in \mathcal{E}_h} (\{\!\{\boldsymbol{\kappa}(p)\mathbf{u} \cdot \boldsymbol{\nu}\}\!\}, [\![v]\!])_e \\ &\qquad = \sum_{e \in \mathcal{E}_h^N} (g_N, v)_e + (f(p), v), \qquad v \in H^1(K_h) \; , \\ &\sum_{K \in K_h} \left(\mathbf{u} - (\boldsymbol{\kappa}(p)\nabla p + \bar{\mathbf{b}}(p)), \boldsymbol{\tau}\right)_K + \sum_{e \in \mathcal{E}_h} \left([\![p]\!], \{\!\{\boldsymbol{\kappa}(p)\boldsymbol{\tau} \cdot \boldsymbol{\nu}\}\!\}\right)_e \\ &\qquad = \sum_{e \in \mathcal{E}_h^D} (g_D, \boldsymbol{\kappa}(p)\boldsymbol{\tau} \cdot \boldsymbol{\nu})_e, \qquad \boldsymbol{\tau} \in \mathbf{H}^1(K_h) \; , \end{aligned} \tag{4.90}$$

where $\bar{\mathbf{b}}(p) = \boldsymbol{\kappa}(p)\mathbf{b}(p)$. Accordingly, the first mixed discontinuous method for (4.87) is: Find $\mathbf{u}_h \in \mathbf{V}_h$ and $p_h \in W_h$ such that

$$\begin{aligned} &\left(\frac{\partial p_h}{\partial t}, v\right) + \sum_{K \in K_h} (\boldsymbol{\kappa}(p_h)\mathbf{u}_h, \nabla v)_K \\ &\qquad - \sum_{e \in \mathcal{E}_h} \left(\{\!\{\boldsymbol{\kappa}(p_h)\mathbf{u} \cdot \boldsymbol{\nu}\}\!\}, [\![v]\!]\right)_e \\ &\qquad\quad = \sum_{e \in \mathcal{E}_h^N} (g_N, v)_e + (f(p_h), v), \qquad v \in W_h \; , \\ &\sum_{K \in K_h} \left(\mathbf{u}_h - (\boldsymbol{\kappa}(p_h)\nabla p_h + \bar{\mathbf{b}}(p_h)), \boldsymbol{\tau}\right)_K \\ &\qquad + \sum_{e \in \mathcal{E}_h} \left([\![p_h]\!], \{\!\{\boldsymbol{\kappa}(p_h)\boldsymbol{\tau} \cdot \boldsymbol{\nu}\}\!\}\right)_e \\ &\qquad\quad = \sum_{e \in \mathcal{E}_h^D} (g_D, \boldsymbol{\kappa}(p_h)\boldsymbol{\tau} \cdot \boldsymbol{\nu})_e, \qquad \boldsymbol{\tau} \in \mathbf{V}_h \; . \end{aligned} \tag{4.91}$$

The second and third mixed discontinuous methods can be defined in a similar way. A time discretization in (4.91) can be carried out by an Euler approach

as in Sect. 1.7 or by an Eulerian-Lagrangian approach (cf. Chap. 5). In the latter approach, the time differentiation and advection terms are combined through a characteristic tracking scheme. The various solution techniques (e.g., linearization, implicit time approximation, and explicit time approximation) developed in Sect. 1.8 for the standard finite element method can be applied to (4.91). Finally, with appropriate assumptions on the coefficients and solution of (4.87), stability and convergence results similar to those in the linear case can be shown (Cockburn-Shu, 1998).

4.4 Theoretical Considerations

In this section, as an example, we present a theoretical analysis for the DG method in Sect. 4.1.1 and its stabilized version in Sect. 4.1.2.

4.4.1 DG Methods

We first analyze the DG method in Sect. 4.1.1: Find $p_h \in V_h$ such that

$$a(p_h, v) = (f, v) \qquad \forall v \in V_h \ , \tag{4.92}$$

where $p_{h,-} = g$ on Γ_-,

$$a_K(v, w) = (\mathbf{b} \cdot \nabla v + Rv, w)_K - \int_{\partial K_-} [v] w_+ \mathbf{b} \cdot \boldsymbol{\nu} \, d\ell, \quad K \in K_h \ ,$$

and

$$a(v, w) = \sum_{K \in K_h} a_K(v, w) \ .$$

If the exact solution p of (4.1) is continuous in Ω, $[p]\mathbf{b} \cdot \boldsymbol{\nu}|_e = 0$, $e \in \mathcal{E}_h^o$, so p satisfies the equation

$$a(p, v) = (f, v) \qquad \forall v \in V_h \ , \tag{4.93}$$

where $p_- = g$ on Γ_-. Then we subtract (4.92) from (4.93) to see that

$$a(p - p_h, v) = 0 \qquad \forall v \in V_h \ . \tag{4.94}$$

We recall the norm $\| \cdot \|_{\mathbf{b}}$ defined in Sect. 4.1.1:

$$\|v\|_{\mathbf{b}} = \left(\|R^{1/2} v\|_{L^2(\Omega)}^2 + \frac{1}{2} \sum_{K \in K_h} \int_{\partial K_-} [v]^2 |\mathbf{b} \cdot \boldsymbol{\nu}| \, d\ell + \frac{1}{2} \int_{\Gamma_+} v_-^2 \mathbf{b} \cdot \boldsymbol{\nu} \, d\ell \right)^{1/2} .$$

Lemma 4.9. *For any $v \in H^1(K_h)$, if $\nabla \cdot \mathbf{b} = 0$, we have*

$$a(v, v) = \|v\|_{\mathbf{b}}^2 - \frac{1}{2} \int_{\Gamma_-} v_-^2 |\mathbf{b} \cdot \boldsymbol{\nu}| \, d\ell \ .$$

Proof. It follows from Green's formula (1.19) and $\nabla \cdot \mathbf{b} = 0$ that

$$(\mathbf{b} \cdot \nabla v, v)_K = \frac{1}{2} \int_{\partial K_+} v_-^2 \mathbf{b} \cdot \boldsymbol{\nu} \, d\ell - \frac{1}{2} \int_{\partial K_-} v_+^2 |\mathbf{b} \cdot \boldsymbol{\nu}| \, d\ell \,,$$

which, together with the definition of $a(\cdot, \cdot)$, implies

$$\begin{aligned} a(v,v) = \sum_{K \in K_h} \Big\{ & \frac{1}{2} \int_{\partial K_+} v_-^2 \mathbf{b} \cdot \boldsymbol{\nu} \, d\ell - \frac{1}{2} \int_{\partial K_-} v_+^2 |\mathbf{b} \cdot \boldsymbol{\nu}| \, d\ell \\ & + \int_{\partial K_-} (v_+ - v_-) v_+ |\mathbf{b} \cdot \boldsymbol{\nu}| \, d\ell \Big\} + \|R^{1/2} v\|_{L^2(\Omega)}^2 \,. \end{aligned} \tag{4.95}$$

Note that each side of ∂K_+ coincides with a side of $\partial K'_-$ for an adjacent element K', except for $\partial K_+ \subset \Gamma_+$, and similarly with $+$ and $-$ reversed. Thus we see that

$$\begin{aligned} \sum_{K \in K_h} \int_{\partial K_+} v_-^2 \mathbf{b} \cdot \boldsymbol{\nu} \, d\ell = & \sum_{K \in K_h} \int_{\partial K_-} v_-^2 |\mathbf{b} \cdot \boldsymbol{\nu}| \, d\ell \\ & + \int_{\Gamma_+} v_-^2 \mathbf{b} \cdot \boldsymbol{\nu} \, d\ell - \int_{\Gamma_-} v_-^2 |\mathbf{b} \cdot \boldsymbol{\nu}| \, d\ell \,. \end{aligned} \tag{4.96}$$

We combine (4.95) and (4.96) to obtain

$$\begin{aligned} a(v,v) = & \frac{1}{2} \sum_{K \in K_h} \int_{\partial K_-} [v]^2 |\mathbf{b} \cdot \boldsymbol{\nu}| \, d\ell \\ & + \frac{1}{2} \int_{\Gamma_+} v_-^2 \mathbf{b} \cdot \boldsymbol{\nu} \, d\ell - \frac{1}{2} \int_{\Gamma_-} v_-^2 |\mathbf{b} \cdot \boldsymbol{\nu}| \, d\ell + \|R^{1/2} v\|_{L^2(\Omega)}^2 \,, \end{aligned}$$

which yields the desired result. □

This lemma implies equation (4.9). Also, existence and uniqueness of a solution to (4.92) follows from this lemma, and if R is strictly positive, inequality (4.10) can be shown. A careful check of the above proof shows that if $R - \nabla \cdot \mathbf{b}/2 \geq 0$ (instead of the assumption $\nabla \cdot \mathbf{b} = 0$), the term $\|R^{1/2} v\|_{L^2(\Omega)}$ can be replaced with $\| \left(R - \nabla \cdot \mathbf{b}/2\right)^{1/2} v \|_{L^2(\Omega)}$.

We now prove the error estimate (4.11). For simplicity, we focus on the case $r = 0$. For a general r, the reader may refer to Johnson and Pitkäranta (1986).

Theorem 4.10. *Let p and p_h be the respective solutions of* (4.93) *and* (4.92) *with $r = 0$. Then there exists a positive constant C, independent of h, such that*

$$\|p - p_h\|_{\mathbf{b}}^2 \leq Ch \sum_{K \in K_h} \|p\|_{H^1(K)}^2 \,.$$

Proof. Define $\bar{p}_h \in V_h$ by

$$\bar{p}_h|_K = \frac{1}{|K|}(p,1)_K, \qquad K \in K_h \ ;$$

i.e., $\bar{p}_h|_K$ is the mean value of p over each $K \in K_h$. Note that $(p-p_h)_- = 0$ on Γ_-. Then we use Lemma 4.9 with $v = p - p_h$ and (4.94) to see that

$$\begin{aligned}
\|p-p_h\|_{\mathbf{b}}^2 &= a(p-p_h, p-p_h) \\
&= a(p-p_h, p-\bar{p}_h) + a(p-p_h, \bar{p}_h - p_h) \\
&= a(p-p_h, p-\bar{p}_h) \\
&= \sum_{K\in K_h} (\mathbf{b}\cdot\nabla(p-p_h) + R(p-p_h), (p-\bar{p}_h))_K \\
&\quad - \int_{\partial K_-} [p-p_h](p-\bar{p}_h)_+ \mathbf{b}\cdot\boldsymbol{\nu}\, d\ell \ .
\end{aligned}$$

Since p_h is piecewise constant, it follows from Cauchy's inequality (1.10) that

$$\begin{aligned}
\|p-p_h\|_{\mathbf{b}}^2 &\le \left(\|\mathbf{b}\cdot\nabla p\|_{L^2(\Omega)} + \|p-p_h\|_{L^2(\Omega)}\right) \|p-\bar{p}_h\|_{L^2(\Omega)} \\
&\quad + \left(\sum_{K\in K_h} \int_{\partial K_-} [p-p_h]^2 |\mathbf{b}\cdot\boldsymbol{\nu}|\, d\ell\right)^{1/2} \\
&\quad \times \left(\sum_{K\in K_h} \int_{\partial K_-} |p-\bar{p}_h|^2 |\mathbf{b}\cdot\boldsymbol{\nu}|\, d\ell\right)^{1/2} .
\end{aligned} \tag{4.97}$$

Applying the approximation properties of $\bar{p}_h$, we have

$$\begin{aligned}
&\|p-\bar{p}_h\|_{L^2(\Omega)}^2 \le Ch^2 \sum_{K\in K_h} \|p\|_{H^1(K)}^2 \ , \\
&\sum_{K\in K_h} \int_{\partial K_-} |p-\bar{p}_h|^2 |\mathbf{b}\cdot\boldsymbol{\nu}|\, d\ell \le Ch \sum_{K\in K_h} \|p\|_{H^1(K)}^2 \ .
\end{aligned} \tag{4.98}$$

Finally, we combine (4.97), (4.98), and the definition of the norm $\|\cdot\|_{\mathbf{b}}$ to obtain the desired result. □

Note that Theorem 4.10 implies (4.11) with $r = 0$.

4.4.2 Stabilized DG Methods

We now study the stabilized DG method in Sect. 4.1.2: Find $p_h \in V_h$ such that

$$a(p_h, v) = \sum_{K\in K_h} (f, v + \theta\mathbf{b}\cdot\nabla v)_K \qquad \forall v \in V_h \ , \tag{4.99}$$

where $p_{h,-} = g$ on Γ_- and

$$a_K(v,w) = (\mathbf{b}\cdot\nabla v + Rv, w + \theta\mathbf{b}\cdot\nabla w)_K - \int_{\partial K_-} [v] w_+ \mathbf{b}\cdot\boldsymbol{\nu}\, d\ell, \qquad K \in K_h\ .$$

Now, the norm $\|\cdot\|_{\mathbf{b}}$ is defined by

$$\|v\|_{\mathbf{b}} = \left(\left\| R^{1/2}\,(1-\theta R/2)^{1/2}\, v \right\|^2_{L^2(\Omega)} + \frac{1}{2}\sum_{K\in K_h} \int_{\partial K_-} [v]^2 |\mathbf{b}\cdot\boldsymbol{\nu}|\, d\ell + \frac{1}{2}\sum_{K\in K_h} \|\theta^{1/2}\mathbf{b}\cdot\nabla v\|^2_{L^2(K)} + \frac{1}{2}\int_{\Gamma_+} v_-^2\, \mathbf{b}\cdot\boldsymbol{\nu}\, d\ell \right)^{1/2}.$$

As in the previous subsection, the exact solution p of (4.1) satisfies

$$a(p,v) = \sum_{K\in K_h} (f, v + \theta\mathbf{b}\cdot\nabla v)_K \qquad \forall v \in V_h\ , \tag{4.100}$$

where $p_- = g$ on Γ_-. Subtracting (4.99) from (4.100) yields

$$a(p - p_h, v) = 0 \qquad \forall v \in V_h\ . \tag{4.101}$$

We only prove (4.14); estimate (4.11) for method (4.99) can be shown in the same way as in Theorem 4.10.

Lemma 4.11. *For any $v \in H^1(K_h)$, if $\nabla\cdot\mathbf{b} = 0$, we have*

$$a(v,v) \geq \|v\|^2_{\mathbf{b}} - \frac{1}{2}\int_{\Gamma_-} v_-^2\, |\mathbf{b}\cdot\boldsymbol{\nu}|\, d\ell\ .$$

Proof. It suffices to treat the term

$$(\mathbf{b}\cdot\nabla v + Rv, \theta\mathbf{b}\cdot\nabla v)_K\ ,$$

since all other terms appeared in the bilinear form $a(\cdot,\cdot)$ in the DG method (4.92). Note that

$$(\mathbf{b}\cdot\nabla v + Rv, \theta\mathbf{b}\cdot\nabla v)_K = \left\|\theta^{1/2}\mathbf{b}\cdot\nabla v\right\|^2_{L^2(K)} + \left(Rv, \theta\mathbf{b}\cdot\nabla v\right)_K\ ,$$

so, using Cauchy's inequality (1.10), we see that

$$(\mathbf{b}\cdot\nabla v + Rv, \theta\mathbf{b}\cdot\nabla v)_K \geq \frac{1}{2}\left(\left\|\theta^{1/2}\mathbf{b}\cdot\nabla v\right\|^2_{L^2(K)} - \left\|\theta^{1/2} Rv\right\|^2_{L^2(K)} \right)\ .$$

This inequality, together with the proof of Lemma 4.9, implies the desired result. □

4.5 Bibliographical Remarks

The definition of the DG and SDG methods given in Sect. 4.1 and their analysis presented in Sect. 4.4 follow those given by Johnson (1994). The computational results discussed in Sects. 4.1 and 4.2 were obtained by Hughes et al. (2000). For the original definition of the non-symmetric DG method and its penalty version for diffusion problems in Sect. 4.2, the reader should refer to Oden et al. (1998) and Rivière et al. (1999), respectively. The content of Sect. 4.3 is essentially based on Chen et al. (2003A) and Chen (2001A). The first and second mixed discontinuous methods described in Sects. 4.3.1.1 and 4.3.1.2 were originally developed by Cockburn-Shu (1998) and Brooks-Hughes (1982), with slightly different forms, where these methods were termed the *local discontinuous Galerkin methods*. Finally, the book edited by Cockburn et al. (2000) contains some of the recent developments on the discontinuous finite element method.

4.6 Exercises

4.1. Write a code to solve problem (4.1) approximately using the discontinuous finite element method developed in Sect. 4.1.1, with $r = 0$ and 1. Use $\mathbf{b} = (1,1)$, $R = 0$, $f(x_1, x_2) = 2\pi^2(\sin(2\pi x_1)\cos(2\pi x_2) + \cos(2\pi x_1)\sin(2\pi x_2))$, $g = 0$, and a uniform partition of $\Omega = (0,1) \times (0,1)$ as given in Fig. 1.7. Also, compute the errors

$$\|p - p_h\| = \left(\int_\Omega (p - p_h)^2 \, d\mathbf{x}\right)^{1/2},$$

$$\left(\sum_{K \in K_h} \|\mathbf{b} \cdot \nabla(p - p_h)\|^2_{L^2(K)}\right)^{1/2}$$

$$= \left(\sum_{K \in K_h} \int_K |\mathbf{b} \cdot \nabla(p - p_h)|^2 \, d\mathbf{x}\right)^{1/2},$$

with $h = 0.1$, 0.01, and 0.001, and compare them. Here p and p_h are the exact and approximate solutions, respectively, and h is the mesh size in the x_1- and x_2-directions.

4.2. Show that if $p \in H^1(\Omega) \cap H^2(K_h)$ is a solution to (4.32), then it satisfies (4.16).

4.3. Prove the boundedness of $a^\theta_-(\cdot,\cdot)$ in (4.37).

4.4. Define $\theta_e = \mathcal{C} \max\{1, r^2\}/h$, $e \in \mathcal{E}_h$. Prove that there is a positive constant $\mathcal{C}_0$ such that (4.38) holds for $\mathcal{C} > \mathcal{C}_0$.

4.5. Define $\theta_e = \mathcal{C} \max\{1, r^2\}/h$, $e \in \mathcal{E}_h$. Show that there is a positive constant $\mathcal{C}_0$ such that the matrix corresponding to the left-hand side of (4.35) is symmetric and positive definite for $\mathcal{C} > \mathcal{C}_0$.

4.6. Let the approximate solution $p_h \in V_h$ satisfy (4.35). Bound p_h in the norm $\|\cdot\|_h$ given in (4.36) in terms of the data f, g_N, and g_D in (4.16).
4.7. Prove the boundedness of $a_+(\cdot,\cdot)$ in (4.44).
4.8. Prove the boundedness of $a_+^\theta(\cdot,\cdot)$ in (4.48).
4.9. Define $\theta_e = \mathcal{C}\max\{1, r^2\}/h$, $e \in \mathcal{E}_h$. Show that (4.49) holds for any $\mathcal{C} > 0$.
4.10. Show that if u and p is a solution of (4.56) and (4.58) and if $p \in H^2(\Omega)$, then p satisfies (4.52).
4.11. Let V_h, W_h, P_h, and R_h be defined as in (4.62), (4.65), and (4.66). Show that (4.59) can be expressed by (4.67), with u_h given by (4.68).
4.12. Introduce appropriate projection operators as in (4.65) and (4.66) to prove that (4.70) can be written in nonmixed form in terms of p_h.
4.13. Derive (4.75) from (4.74).
4.14. Based on (4.75), define the first mixed discontinuous method and then write it in nonmixed form in terms of p_h by introducing appropriate projection operators (cf. Sect. 4.3.1.1).
4.15. Derive (4.79) from (4.78) and the boundary conditions in problem (4.76).
4.16. Let $\mathbf{P}_h$ and $\mathbf{R}_h$ be defined as in (4.84) and (4.85), respectively. Write (4.80) in nonmixed form in terms of p_h.
4.17. Show that if $R - \nabla\cdot\mathbf{b}/2 \geq 0$ (instead of the assumption $\nabla\cdot\mathbf{b} = 0$ in Lemma 4.9), Lemma 4.9 remains valid provided that the term $\|R^{1/2}v\|_{L^2(\Omega)}$ is replaced by $\|\left(R - \nabla\cdot\mathbf{b}/2\right)^{1/2} v\|_{L^2(\Omega)}$ in the definition of the norm $\|\cdot\|_{\mathbf{b}}$.
4.18. Prove the error estimate (4.39) for the symmetric interior penalty DG method (4.35). (If necessary, consult Arnold (1982).)
4.19. Prove the error estimate (4.39) for the nonsymmetric interior penalty DG method (4.47). (If necessary, consult Rivière et al. (1999).)

5 Characteristic Finite Elements

In this chapter, we consider an application of the finite element method to the *reaction-diffusion-advection problem*:

$$\frac{\partial(cp)}{\partial t} + \nabla \cdot \big(\mathbf{b}p - \mathbf{a}\nabla p\big) + Rp = f \ , \tag{5.1}$$

for the unknown solution p, where c, $\mathbf{b}$ (vector), $\mathbf{a}$ (tensor), R, and f are given functions. Note that (5.1) involves advection ($\mathbf{b}$), diffusion ($\mathbf{a}$), and reaction (R). Many problems arise in this form, e.g., saturation problems for multiphase flow in porous media (cf. Chap. 9), transport problems for contaminants in groundwater, and density problems for semiconductor modeling (cf. Chap. 10), to name a few. Problems of this type were considered in the preceding chapter.

When diffusion dominates advection, the finite element method developed in Chap. 1 performs well for (5.1). When advection dominates diffusion, however, it does not work well. In particular, it exhibits excessive nonphysical oscillations when the solution to (5.1) is not smooth. Standard upstream weighting approaches have been applied to the finite element method with the purpose of eliminating the nonphysical oscillations, but these approaches smear sharp fronts in the solution and suffer from grid-orientation difficulties. Although extremely fine mesh refinement is possible to overcome some of these difficulties, it is not feasible due to the excessive computational effort involved.

Many numerical methods have been developed for solving (5.1) where advection dominates, such as the *optimal spatial method*. This method employs an *Eulerian approach* that is based on the minimization of the error in the approximation of spatial derivatives and the use of optimal test functions satisfying a local adjoint problem (Brooks-Hughes, 1982; Barrett-Morton, 1984). It yields an upstream bias in the resulting approximation and has the features: (i) time truncation errors dominate the solution; (ii) the solution has significant numerical diffusion and phase errors; (iii) the *Courant number* is generally restricted to be less than one (see (5.43) for the definition of this number).

Other Eulerian methods such as the *Petrov-Galerkin finite element method* have been developed to use nonzero spatial truncation errors to cancel temporal errors and thereby reduce the overall truncation errors (Christie et al.,

1976; Westerink-Shea, 1989). While these methods improve accuracy in the approximation of the solution, they still suffer from a strict Courant number limitation.

Another class of numerical methods for the solution of (5.1) are the *Eulerian-Lagrangian methods*. Because of the Lagrangian nature of advection, these methods treat the advection by a characteristic tracking approach. They have shown great potential. This class is rich and bears a variety of names, the *method of characteristics* (Garder et al., 1984), the *modified method of characteristics* (Douglas-Russell, 1982), the *transport diffusion method* (Pironneau, 1982), the *Eulerian-Lagrangian method* (Neuman, 1981), the *operator splitting method* (Espedal-Ewing, 1987), the *Eulerian-Lagrangian localized adjoint method* (Celia et al., 1990; Russell, 1990), the *characteristic mixed finite element method* (Yang, 1992; Arbogast-Wheeler, 1995), and the *Eulerian-Lagrangian mixed discontinuous method* (Chen, 2002B). The common features of this class are: (i) the Courant number restriction of the purely Eulerian methods is alleviated because of the Lagrangian nature of the advection step; (ii) since the spatial and temporal dimensions are coupled through the characteristic tracking, the effect of time truncation errors present in the optimal spatial method is greatly reduced; (iii) they produce non-oscillatory solutions without numerical diffusion, using reasonably large time steps on grids no finer than necessary to resolve the solution on the moving fronts. In this chapter, we describe the Eulerian-Lagrangian methods. Especially, we discuss the modified method of characteristics (cf. Sect. 5.2), the Eulerian-Lagrangian method (cf. Sect. 5.3), the characteristic mixed method (cf. Sect. 5.4), and the Eulerian-Lagrangian mixed discontinuous method (cf. Sect. 5.5). Other characteristic methods either are similar to these methods or can be deduced from them (Chen, 2002C). In Sect. 5.6, nonlinear problems are considered. In Sect. 5.7, we further comment on the characteristic finite element method. Section 5.8 is devoted to theoretical considerations. Finally, bibliographical information is given in Sect. 1.9.

5.1 An Example

We consider an example of (5.1):

$$c\frac{\partial p}{\partial t} + \mathbf{b}\cdot\nabla p - \nabla\cdot\big(\mathbf{a}\nabla p\big) + Rp = f, \quad \mathbf{x}\in\Omega,\ t>0\ ,$$

where $\Omega\subset\mathbb{R}^d$ $(d\le 3)$ is a bounded domain with boundary Γ. In this section, we briefly study its hyperbolic part

$$c\frac{\partial p}{\partial t} + \mathbf{b}\cdot\nabla p + Rp = f, \quad \mathbf{x}\in\Omega,\ t>0\ , \tag{5.2}$$

and first consider its steady-state version

$$\mathbf{b} \cdot \nabla p + Rp = f, \quad \mathbf{x} \in \Omega \,. \tag{5.3}$$

That is, in (5.3) all functions are assumed to be independent of time t. Problem (5.3) was also considered in Sect. 4.1. The *characteristic curves* (or *characteristics*) corresponding to the given velocity field $\mathbf{b} = (b_1, b_2, \ldots, b_d)$ are the curves $\mathbf{x}(s)$ defined by

$$\frac{dx_i}{ds} = b_i(\mathbf{x}), \qquad i = 1, 2, \ldots, d \,,$$

where $\mathbf{x}(s) = (x_1(s), x_2(s), \ldots, x_d(s))$ and these characteristics are parametrized by the parameter s. In vector form, we have

$$\frac{d\mathbf{x}}{ds} = \mathbf{b}(\mathbf{x}) \,.$$

In the context of fluid dynamics, these curves are called the *streamlines* associated with $\mathbf{b}$. If $\mathbf{b}$ is Lipschitz-continuous (i.e., $\|\mathbf{b}(\mathbf{x}) - \mathbf{b}(\mathbf{y})\| \le C\|\mathbf{x} - \mathbf{y}\|$, for all $\mathbf{x}$, $\mathbf{y} \in \Omega$ and for some constant C, where $\|\cdot\|$ is the Euclidean norm), for a given point $\mathbf{x}_- \in \Omega$ there exists a unique characteristic $\mathbf{x}(s)$ passing through $\mathbf{x}_-$. For such a characteristic $\mathbf{x}(s)$, it follows from the chain rule that

$$\frac{dp(\mathbf{x})}{ds} = \nabla p \cdot \frac{d\mathbf{x}}{ds} = \mathbf{b} \cdot \nabla p \,;$$

consequently, (5.3) reduces to

$$\frac{dp(\mathbf{x})}{ds} + Rp(\mathbf{x}) = f \,. \tag{5.4}$$

Hence, along each characteristic $\mathbf{x}(s)$, (5.3) becomes an ordinary differential equation. If p is known at a point on $\mathbf{x}(s)$, then p can be determined at other points on $\mathbf{x}(s)$ by integrating (5.4). In general, p is prescribed on the *inflow boundary* Γ_-

$$\Gamma_- = \{\mathbf{x} \in \Gamma : \ (\mathbf{b} \cdot \boldsymbol{\nu})(\mathbf{x}) < 0\} \,,$$

where $\boldsymbol{\nu}$ is the outward unit normal to the boundary Γ. The solution p at any point $\mathbf{x} \in \Omega$ can be found by integrating along the characteristic through $\mathbf{x}$ starting on Γ_- (cf. Fig. 5.1). In particular, for (5.3) this implies that effects are propagated along characteristics.

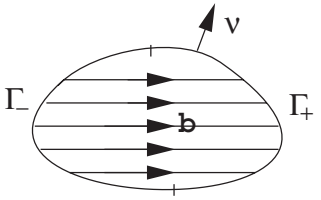

Fig. 5.1. Characteristics

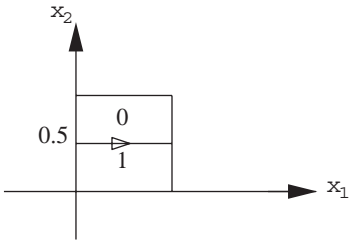

Fig. 5.2. A discontinuous solution

We observe that a solution of (5.3) may be discontinuous across a characteristic. For example, if the boundary datum is discontinuous at some point $\mathbf{x}_- \in \Gamma_-$, then p is discontinuous across the whole characteristic passing through $\mathbf{x}_-$. As an example, consider the problem on the unit square

$$\begin{aligned} &\frac{\partial p}{\partial x_1} = 0, && 0 < x_i < 1,\ i = 1, 2\ , \\ &p(0, x_2) = 1, && 0 < x_2 < \frac{1}{2}\ , \\ &p(0, x_2) = 0, && \frac{1}{2} < x_2 < 1\ . \end{aligned}$$

This problem is a special case of (5.3) where $\mathbf{b} = (1, 0)$ and $R = 0$. It is obvious that the solution to this problem is (cf. Fig. 5.2)

$$\begin{aligned} &p(x_1, x_2) = 1, && 0 < x_1 < 1,\ 0 < x_2 < \frac{1}{2}\ , \\ &p(x_1, x_2) = 0, && 0 < x_1 < 1,\ \frac{1}{2} < x_2 < 1\ . \end{aligned}$$

For the time-dependent problem (5.2), set $t = x_0$ and $b_0 = c$ so that this problem is rewritten as

$$\bar{\mathbf{b}} \cdot \nabla_{(t,\mathbf{x})} p + Rp = f\ , \tag{5.5}$$

where $\bar{\mathbf{b}} = (\mathrm{b}_0, \mathbf{b})$ and $\nabla_{(t,\mathbf{x})} = (\frac{\partial}{\partial t}, \nabla)$. Equation (5.5) has the same form as (5.3), and thus the discussion on (5.3) applies to (5.5).

5.2 The Modified Method of Characteristics

5.2.1 A One-Dimensional Model Problem

The modified method of characteristics (MMOC) was independently developed by Douglas-Russell (1982) and Pironneau (1982) and is based on a non-divergence form of (5.1). It was called the *transport-diffusion method*

by Pironneau. In the engineering literature the name *Eulerian-Lagrangian method* is often used (Neuman, 1981).

For the purpose of introduction, we consider a one-dimensional model problem on the whole real line:

$$\begin{aligned} c(x)\frac{\partial p}{\partial t} + b(x)\frac{\partial p}{\partial x} - \frac{\partial}{\partial x}\left(a(x,t)\frac{\partial p}{\partial x}\right) + R(x,t)p = f(x,t)\ , \\ x \in \mathbb{R},\ t > 0\ , \\ p(x,0) = p_0(x), \qquad x \in \mathbb{R}\ . \end{aligned} \tag{5.6}$$

Set

$$\psi(x) = \left(c^2(x) + b^2(x)\right)^{1/2}\ .$$

Assume that

$$c(x) > 0, \qquad x \in \mathbb{R}\ ,$$

so $\psi(x) > 0$, $x \in \mathbb{R}$. Let the characteristic direction associated with the hyperbolic part of (5.6), $c\partial p/\partial t + b\partial p/\partial x$, be denoted by $\tau(x)$, so

$$\frac{\partial}{\partial \tau(x)} = \frac{c(x)}{\psi(x)}\frac{\partial}{\partial t} + \frac{b(x)}{\psi(x)}\frac{\partial}{\partial x}\ .$$

Then (5.6) can be rewritten as

$$\begin{aligned} \psi(x)\frac{\partial p}{\partial \tau} - \frac{\partial}{\partial x}\left(a(x,t)\frac{\partial p}{\partial x}\right) + R(x,t)p = f(x,t)\ , \\ x \in \mathbb{R},\ t > 0\ , \\ p(x,0) = p_0(x), \qquad x \in \mathbb{R}\ . \end{aligned} \tag{5.7}$$

We assume that the coefficients a, b, c, and R are bounded and satisfy

$$\left|\frac{b(x)}{c(x)}\right| + \left|\frac{d}{dx}\left(\frac{b(x)}{c(x)}\right)\right| \le C, \qquad x \in \mathbb{R}\ ,$$

where C is a positive constant. We introduce the linear space (cf. Sect. 1.2)

$$V = W^{1,2}(\mathbb{R})\ .$$

We also recall the scalar product in $L^2(\mathbb{R})$

$$(v, w) = \int_{\mathbb{R}} v(x)w(x)\ dx\ .$$

Now, multiplying the first equation of (5.7) by any $v \in V$ and applying integration by parts in space, problem (5.7) can be written in the equivalent variational form

$$\begin{aligned} &\left(\psi\frac{\partial p}{\partial\tau}, v\right) + \left(a\frac{\partial p}{\partial x}, \frac{dv}{dx}\right) + (Rp, v) = (f, v), \quad v \in V,\ t > 0\ , \\ &p(x, 0) = p_0(x), \qquad\qquad x \in \mathbb{R}\ . \end{aligned} \tag{5.8}$$

Let $0 = t^0 < t^1 < \ldots < t^n < \ldots$ be a partition in time, with $\Delta t^n = t^n - t^{n-1}$. For a generic function v of time, set $v^n = v(t^n)$. The characteristic derivative is approximated in the following way: Let

$$\check{x}_n = x - \frac{\Delta t^n}{c(x)} b(x)\ , \tag{5.9}$$

and note that, at $t = t^n$,

$$\begin{aligned} \psi\frac{\partial p}{\partial\tau} &\approx \psi(x)\frac{p(x, t^n) - p(\check{x}_n, t^{n-1})}{\left((x - \check{x}_n)^2 + (\Delta t^n)^2\right)^{1/2}} \\ &= c(x)\frac{p(x, t^n) - p(\check{x}_n, t^{n-1})}{\Delta t^n}\ . \end{aligned} \tag{5.10}$$

Namely, a backtracking algorithm is used to approximate the characteristic derivative; $\check{x}_n$ is the foot (at level t^{n-1}) of the characteristic corresponding to x at the head (at level t^n) (cf. Fig. 5.3).

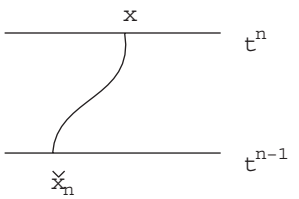

Fig. 5.3. An illustration of the definition $\check{x}_n$

Let V_h be a finite element subspace of $V \cap W^{1,\infty}(\mathbb{R})$ (cf. Chap. 1). Because we are considering the whole line, V_h is necessarily infinite-dimensional. In practice, we can assume that the support of p_0 is compact, the portion of the line on which we need to know p is bounded, and p is very small outside that set. Then V_h can be taken to be finite-dimensional.

The MMOC for (5.6) is defined: For $n = 1, 2, \ldots$, find $p_h^n \in V_h$ such that

$$\begin{aligned} &\left(c\frac{p_h^n - \check{p}_h^{n-1}}{\Delta t^n}, v\right) + \left(a^n\frac{dp_h^n}{dx}, \frac{dv}{dx}\right) \\ &\qquad + (R^n p_h^n, v) = (f^n, v) \qquad \forall v \in V_h\ , \end{aligned} \tag{5.11}$$

where

$$\check{p}_h^{n-1} = p_h\left(\check{x}_n, t^{n-1}\right) = p_h\left(x - \frac{\Delta t^n}{c(x)} b(x), t^{n-1}\right). \tag{5.12}$$

The initial approximation p_h^0 can be defined as any reasonable approximation of p_0 in V_h such as the interpolant of p_0 in V_h.

Note that (5.11) determines $\{p_h^n\}$ uniquely in terms of the data p_0 and f (at least, for reasonable a and R such that a is uniformly positive with respect to x and t and R is nonnegative). This can be seen as follows: Since (5.11) is a finite-dimensional system, it suffices to show uniqueness of the solution. Let $f = p_0 = 0$, and take $v = p_h^n$ in (5.11) to see that

$$\left(c\frac{p_h^n - \check{p}_h^{n-1}}{\Delta t^n}, p_h^n\right) + \left(a^n \frac{dp_h^n}{dx}, \frac{dp_h^n}{dx}\right) + (R^n p_h^n, p_h^n) = 0\ ;$$

with an induction assumption that $p_h^{n-1} = 0$, this equation implies $p_h^n = 0$.

It is obvious that the linear system arising from (5.11) is symmetric positive definite (cf. Sect. 1.1.1), even in the presence of the advection term. This system has an improved condition number of order (cf. Sect. 1.10 and Exercise 5.1)

$$\mathcal{O}\big(1 + \max_{x\in\mathbb{R},\ t\geq 0} |a(x,t)|h^{-2}\Delta t\big), \quad \Delta t = \max_{n=1,2,\ldots} \Delta t^n\ .$$

Thus the system arising from (5.11) is well suited for the iterative linear solution algorithms discussed in Sect. 1.10.

We end with a remark on a convergence result for (5.11). Let $V_h \subset V$ be a finite element space (cf. Chap. 1) with the following approximation property:

$$\inf_{v_h\in V_h} \left(\|v - v_h\|_{L^2(\mathbb{R})} + h\|v - v_h\|_{W^{1,2}(\mathbb{R})}\right) \leq Ch^{r+1}|v|_{W^{r+1,2}(\mathbb{R})}\ , \tag{5.13}$$

where the constant $C > 0$ is independent of h and $r > 0$ is an integer; refer to Sect. 1.2 for the definition of spaces and their norms. Then, under appropriate assumptions on the smoothness of the solution p and a suitable choice of p_h^0 it can be shown (Douglas-Russell, 1982) that

$$\begin{aligned}\max_{1\leq n\leq N} \left(\|p^n - p_h^n\|_{L^2(\mathbb{R})} + h\|p^n - p_h^n\|_{W^{1,2}(\mathbb{R})}\right) \\ \leq C(p)\left(h^{r+1} + \Delta t\right)\ ,\end{aligned} \tag{5.14}$$

where N is an integer such that $t^N = T < \infty$ and $J = (0, T]$ is the time interval of interest; see Sect. 5.8 for more information.

This result, by itself, is not different from what we have obtained with the standard finite element method in Chap. 1. However, the constant C is greatly improved when the MMOC is applied to (5.6). In time, C depends on a norm of $\frac{\partial^2 p}{\partial t^2}$ with the standard method, but on a norm of $\frac{\partial^2 p}{\partial \tau^2}$ with the

MMOC. The latter norm is much smaller, and thus long time steps with large Courant numbers (see their definition in the next section) are possible. The reader may refer to Sect. 5.8 for more details.

Some matters are raised by (5.11) and its analogues for more complicated differential problems considered later. The first concern is the backtracking scheme that determines $\check{x}_n$ and a numerical quadrature rule that computes the associated integral. For the problem considered in this subsection, this matter can be resolved; the required computations can be performed exactly. For more complicated problems, there are some discussions by Russell-Trujillo (1990). The second matter is the treatment of boundary conditions. In this section, we work on the whole line or on periodic boundary conditions (see the next subsection). For a bounded domain, if a backtracked characteristic crosses a boundary of the domain, it is not obvious what is the meaning of $\check{x}_n$ or of $p_h(\check{x}_n)$. The last matter, and perhaps the greatest drawback of the MMOC, is its failure to conserve mass. This issue will be discussed in detail in Sect. 5.2.4.

5.2.2 Periodic Boundary Conditions

In the previous subsection, (5.6) was considered on the whole line. For a bounded interval, say, $(0,1)$, the MMOC has a difficulty handling general boundary conditions. In this case, it is normally developed for *periodic boundary conditions* (cf. Exercise 5.2):

$$p(0,t) = p(1,t), \quad \frac{\partial p}{\partial x}(0,t) = \frac{\partial p}{\partial x}(1,t) \; . \tag{5.15}$$

These conditions are also called *cyclic boundary conditions*. In this case, we assume that all functions in (5.6) are spatially $(0,1)$-periodic. Accordingly, the linear space V is modified by

$$V = \{v \in H^1(I) : v \text{ is } I\text{-periodic}\}, \quad I = (0,1) \; .$$

With this modification, the developments in (5.8) and (5.11) remain unchanged.

5.2.3 Extension to Multi-Dimensional Problems

We now extend the MMOC to (5.1) defined on a multi-dimensional domain. Let $\Omega \subset \mathbb{R}^d$ $(d \le 3)$ be a rectangle (respectively, a rectangular parallelepiped), and assume that (5.1) is Ω-periodic; i.e., all functions in (5.1) are spatially Ω-periodic. We write (5.1) in nondivergence form:

$$\begin{aligned} c(\mathbf{x})\frac{\partial p}{\partial t} + \mathbf{b}(\mathbf{x},t)\cdot\nabla p - \nabla\cdot\big(\mathbf{a}(\mathbf{x},t)\nabla p\big) \\ + R(\mathbf{x},t)p = f(\mathbf{x},t), \quad \mathbf{x}\in\Omega, \; t>0 \; , \\ p(\mathbf{x},0) = p_0(\mathbf{x}), \quad \mathbf{x}\in\Omega \; . \end{aligned} \tag{5.16}$$

Let

$$\psi(\mathbf{x},t) = \left(c^2(\mathbf{x}) + \|\mathbf{b}(\mathbf{x},t)\|^2\right)^{1/2} ,$$

where $\|\mathbf{b}\|^2 = b_1^2 + b_2^2 + \cdots + b_d^2$, with $\mathbf{b} = (b_1, b_2, \ldots, b_d)$. Assume that

$$c(\mathbf{x}) > 0, \qquad \mathbf{x} \in \Omega .$$

Now, the characteristic direction corresponding to the hyperbolic part of (5.16), $c\partial p/\partial t + \mathbf{b} \cdot \nabla p$, is $\boldsymbol{\tau}$, so

$$\frac{\partial}{\partial \boldsymbol{\tau}} = \frac{c(\mathbf{x})}{\psi(\mathbf{x},t)} \frac{\partial}{\partial t} + \frac{1}{\psi(\mathbf{x},t)} \mathbf{b}(\mathbf{x},t) \cdot \nabla .$$

With this definition, (5.16) becomes

$$\begin{aligned} \psi(\mathbf{x},t)\frac{\partial p}{\partial \boldsymbol{\tau}} - \nabla \cdot \left(\mathbf{a}(\mathbf{x},t)\nabla p\right) + R(\mathbf{x},t)p = f(\mathbf{x},t) , \\ \mathbf{x} \in \Omega, \ t > 0 , \\ p(\mathbf{x},0) = p_0(\mathbf{x}), \qquad \mathbf{x} \in \Omega . \end{aligned} \tag{5.17}$$

We define the linear space

$$V = \{v \in H^1(\Omega) : v \text{ is } \Omega\text{-periodic}\} .$$

Recall the notation

$$(v,w)_S = \int_S v(\mathbf{x}) w(\mathbf{x}) \, d\mathbf{x} .$$

If $S = \Omega$, we omit it in this notation. Now, applying Green's formula (1.19) in space and the periodic boundary conditions, (5.17) is written in the equivalent variational form

$$\begin{aligned} \left(\psi \frac{\partial p}{\partial \boldsymbol{\tau}}, v\right) + (\mathbf{a}\nabla p, \nabla v) + (Rp, v) = (f, v), \qquad v \in V, \ t > 0 , \\ p(\mathbf{x},0) = p_0(\mathbf{x}), \qquad \mathbf{x} \in \Omega . \end{aligned} \tag{5.18}$$

The characteristic is approximated by

$$\check{\mathbf{x}}_n = \mathbf{x} - \frac{\Delta t^n}{c(\mathbf{x})} \mathbf{b}(\mathbf{x}, t^n) . \tag{5.19}$$

Furthermore, we see that, at $t = t^n$,

$$\begin{aligned} \psi \frac{\partial p}{\partial \boldsymbol{\tau}} &\approx \psi(\mathbf{x}, t^n) \frac{p(\mathbf{x}, t^n) - p(\check{\mathbf{x}}_n, t^{n-1})}{\left(\|\mathbf{x} - \check{\mathbf{x}}_n\|^2 + (\Delta t^n)^2\right)^{1/2}} \\ &= c(\mathbf{x}) \frac{p(\mathbf{x}, t^n) - p(\check{\mathbf{x}}_n, t^{n-1})}{\Delta t^n} . \end{aligned} \tag{5.20}$$

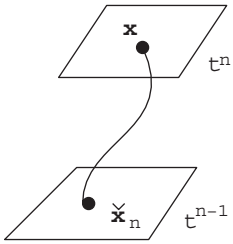

Fig. 5.4. An illustration of the definition $\check{\mathbf{x}}_n$

A backtracking algorithm similar to that employed in one dimension is used to approximate the characteristic derivative (cf. Fig. 5.4).

Let $V_h \subset V$ be a finite element space associated with a regular partition K_h of Ω (cf. Chap. 1). The MMOC for (5.16) is given: For $n = 1, 2, \ldots$, find $p_h^n \in V_h$ such that

$$\left(c\frac{p_h^n - \check{p}_h^{n-1}}{\Delta t^n}, v\right) + (\mathbf{a}^n \nabla p_h^n, \nabla v) + (R^n p_h^n, v) = (f^n, v) \quad \forall v \in V_h \,, \tag{5.21}$$

where

$$\check{p}_h^{n-1} = p_h\left(\check{\mathbf{x}}_n, t^{n-1}\right) = p_h\left(\mathbf{x} - \frac{\Delta t^n}{c(\mathbf{x})}\mathbf{b}(\mathbf{x}, t^n), t^{n-1}\right). \tag{5.22}$$

The remarks made in Sect. 5.2.1 for (5.11) also apply to (5.21). In particular, existence and uniqueness of a solution for reasonable choices of $\mathbf{a}$ and R can be shown in the same way (cf. Exercise 5.3), and the error estimate (5.14) under appropriate assumptions on p holds for (5.21) (cf. Sect. 5.8)

$$\max_{1 \le n \le N} \left(\|p^n - p_h^n\|_{L^2(\Omega)} + h\|p^n - p_h^n\|_{H^1(\Omega)}\right) \le C(p)\left(h^{r+1} + \Delta t\right) \,,$$

provided an approximation property similar to (5.13) holds for V_h in the multiple dimensions.

5.2.4 Discussion of a Conservation Relation

We discuss the MMOC in the simple case where

$$R = f = 0, \quad \nabla \cdot \mathbf{b} = 0 \qquad \text{in } \Omega \,. \tag{5.23}$$

That is, $\mathbf{b}$ is *divergence-free* (or *solenoidal*). Application of condition (5.23), the periodicity assumption, and the divergence theorem (1.17) to (5.16) yields the *conservation relation*

$$\int_\Omega c(\mathbf{x}) p(\mathbf{x}, t)\, d\mathbf{x} = \int_\Omega c(\mathbf{x}) p_0(\mathbf{x})\, d\mathbf{x}, \qquad t > 0 \,. \tag{5.24}$$

In applications, it is desirable to conserve at least a discrete form of this relation in any numerical approximation of (5.16). However, in general, the MMOC does not conserve it. To see this, we take $v = 1$ in (5.21) and apply (5.23) to give

$$\begin{aligned} \int_\Omega c(\mathbf{x}) p(\mathbf{x}, t^n)\, d\mathbf{x} &= \int_\Omega c(\mathbf{x}) p(\check{\mathbf{x}}_n, t^{n-1})\, d\mathbf{x} \\ &\neq \int_\Omega c(\mathbf{x}) p(\mathbf{x}, t^{n-1})\, d\mathbf{x} \,. \end{aligned} \tag{5.25}$$

For each n, define the transformation

$$\mathbf{G}(\mathbf{x}) \equiv \mathbf{G}(\mathbf{x}, t^n) = \mathbf{x} - \frac{\Delta t^n}{c(\mathbf{x})} \mathbf{b}(\mathbf{x}, t^n) \,. \tag{5.26}$$

We assume that $\mathbf{b}/c$ has bounded first partial derivatives in space. Then, for $d = 3$, the *Jacobian of this transformation*, $\mathbf{J}(\mathbf{G})$, is

$$\begin{pmatrix} 1 - \dfrac{\partial}{\partial x_1}\left(\dfrac{b_1^n}{c}\right)\Delta t^n & -\dfrac{\partial}{\partial x_2}\left(\dfrac{b_1^n}{c}\right)\Delta t^n & -\dfrac{\partial}{\partial x_3}\left(\dfrac{b_1^n}{c}\right)\Delta t^n \\ -\dfrac{\partial}{\partial x_1}\left(\dfrac{b_2^n}{c}\right)\Delta t^n & 1 - \dfrac{\partial}{\partial x_2}\left(\dfrac{b_2^n}{c}\right)\Delta t^n & -\dfrac{\partial}{\partial x_3}\left(\dfrac{b_2^n}{c}\right)\Delta t^n \\ -\dfrac{\partial}{\partial x_1}\left(\dfrac{b_3^n}{c}\right)\Delta t^n & -\dfrac{\partial}{\partial x_2}\left(\dfrac{b_3^n}{c}\right)\Delta t^n & 1 - \dfrac{\partial}{\partial x_1}\left(\dfrac{b_3^n}{c}\right)\Delta t^n \end{pmatrix},$$

and its determinant equals (cf. Exercise 5.4)

$$|\mathbf{J}(\mathbf{G})| = 1 - \nabla \cdot \left(\frac{\mathbf{b}^n}{c}\right)\Delta t^n + \mathcal{O}((\Delta t^n)^2). \tag{5.27}$$

Thus, even in the case where c is constant, for the second equality of (5.25) to hold requires that the Jacobian of the transformation (5.26) be identically one. While this is true for constant c and $\mathbf{b}$, it cannot be expected to be true for variable coefficients. In the case where c is constant and $\nabla \cdot \mathbf{b} = 0$, it follows from (5.27) that the determinant of this transformation is $1 + \mathcal{O}((\Delta t^n)^2)$, so a systematic error of size $\mathcal{O}\left((\Delta t^n)^2\right)$ should be expected. On the other hand, if $\nabla \cdot (\mathbf{b}/c) \neq 0$, the determinant is $1 + \mathcal{O}(\Delta t^n)$ and a systematic error of size $\mathcal{O}(\Delta t^n)$ can occur. In particular, in using the MMOC in the solution of a two-phase immiscible flow problem (cf. Chap. 9), Douglas et al. (1997) found that conservation of mass failed by as much as 10% in simulations with stochastic rock properties and about half that much with uniform rock properties. Errors of this magnitude obscure the relevance of numerical approximations to an unacceptable level and motivate the search for a modification of the MMOC that both conserves (5.24) and is at most very little more expensive computationally than the MMOC. A new method, the *modified method of characteristics with adjusted advection*, was defined

by Douglas et al. (1997) to satisfy these criteria. This method is defined from the MMOC by perturbing the foot of characteristics in an *ad hoc* fashion. We do not introduce this method in this chapter. Instead, we describe the *Eulerian-Lagrangian localized adjoint method* (ELLAM) in the next section, since the idea of the ELLAM will be used in the definition of other two methods studied.

5.3 The Eulerian-Lagrangian Localized Adjoint Method

5.3.1 A One-Dimensional Model Problem

To sketch the idea of the ELLAM (Celia et al., 1990; Russell, 1990), we consider a one-dimensional reaction-diffusion-advection problem:

$$\frac{\partial(cp)}{\partial t} + \frac{\partial}{\partial x}\left(bp - a\frac{\partial p}{\partial x}\right) + Rp = f, \quad x \in I,\ t > 0\ , \tag{5.28}$$

where $I = (0,1)$ is the space interval. Furthermore, let c and b be constant. An extension to a general case will be considered in the next subsection. We consider the boundary conditions

$$\begin{aligned} &p(0,t) = g_0(t) \text{ or } \left(bp - a\frac{\partial p}{\partial x}\right)(0,t) = g_0(t), \qquad t > 0\ , \\ &p(1,t) = g_1(t) \text{ or } \left(bp - a\frac{\partial p}{\partial x}\right)(1,t) = g_1(t), \qquad t > 0\ , \end{aligned} \tag{5.29}$$

where g_0 and g_1 are given. The initial condition is the same as in (5.6):

$$p(x,0) = p_0(x), \qquad x \in I\ .$$

Below let $\Gamma = \{0,1\}$, i.e., the boundary of I.

The origin of the ELLAM can be seen by considering (5.28) in a space-time framework and in divergence form. For any $x \in I$ and two times $0 \le t^{n-1} < t^n$, the hyperbolic part of problem (5.28), as in the previous section, $c\partial p/\partial t + b\partial p/\partial x$, defines the characteristics $\check{x}_n(x,t)$ along the interstitial velocity $\varphi = b/c$:

$$\check{x}_n(x,t) = x - \varphi(t^n - t), \qquad t \in [\check{t}(x), t^n]\ , \tag{5.30}$$

where $\check{t}(x) = t^{n-1}$ if $\check{x}_n(x,t)$ does not backtrack to the boundary Γ for $t \in [t^{n-1}, t^n]$; $\check{t}(x) \in (t^{n-1}, t^n]$ is the time instant when $\check{x}_n(x,t)$ intersects Γ, i.e., $\check{x}_n(x, \check{t}(x)) \in \Gamma$, otherwise. Note that this characteristic emanates backward from x at t^n; see Fig. 5.5. If $b > 0$, the characteristics at the right boundary ($x = 1, t \in J^n = (t^{n-1}, t^n]$) are defined by

$$\check{x}_n(1,\theta) = 1 - \varphi(t - \theta), \qquad \theta \in [t^{n-1}, t]\ . \tag{5.31}$$

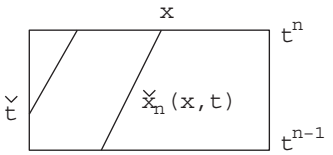

Fig. 5.5. An illustration of characteristics for constant c and b

Similarly, we can define the characteristics at the left boundary ($x = 0, t \in J^n$) if $b < 0$. For simplicity of exposition, we focus on the case where $b > 0$.

For a positive integer M, let $0 = x_0 < x_1 < \ldots < x_M = 1$ be a partition K_h of I into subintervals $I_i = (x_{i-1}, x_i)$, with length $h_i = x_i - x_{i-1}$, $i = 1, 2, \ldots, M$. Set $h = \max\{h_i : i = 1, 2, \ldots, M\}$.

For each subinterval $I_i \in K_h$, let $\check{I}_i(t)$ indicate the trace-back of I_i to time t, $t \in J^n$:

$$\check{I}_i(t) = \{x \in I : x = \check{x}_n(y,t) \text{ for some } y \in I_i\} .$$

Also, let $\mathcal{I}_i^n$ be the space-time region that follows the characteristics (cf. Fig. 5.6):

$$\mathcal{I}_i^n = \{(x,t) \in I \times J : t \in J^n \text{ and } x \in \check{I}_i(t)\} .$$

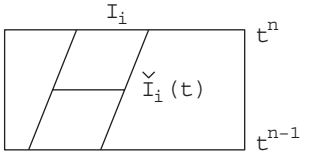

Fig. 5.6. The definition of $\mathcal{I}_i^n$

5.3.1.1 Interior ELLAM Formulation

Let $\mathcal{I}_i^n \cap (\Gamma \times J^n) = \emptyset$. We multiply (5.28) by a smooth test function $v(x,t)$ and integrate over $\mathcal{I}_i^n$. With $\boldsymbol{\tau} = (b, c)$ indicating the characteristic direction and application of Green's formula (1.19) in space and time, the hyperbolic part of (5.28) gives

$$\begin{aligned}
\int_{\mathcal{I}_i^n} \left(c\frac{\partial p}{\partial t} + b\frac{\partial p}{\partial x} \right) v \; dx \; dt &= \int_{\mathcal{I}_i^n} \left(\frac{\partial p}{\partial x}, \frac{\partial p}{\partial t} \right) \cdot \boldsymbol{\tau} v \; dx \; dt \\
&= \int_{\partial \mathcal{I}_i^n} p\boldsymbol{\tau} \cdot \boldsymbol{\nu}_{\mathcal{I}_i^n} v \; d\ell - \int_{\mathcal{I}_i^n} p \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial t} \right) \cdot \boldsymbol{\tau} v \; dx \; dt \\
&\quad - \int_{\mathcal{I}_i^n} p\boldsymbol{\tau} \cdot \left(\frac{\partial v}{\partial x}, \frac{\partial v}{\partial t} \right) \; dx \; dt ,
\end{aligned}$$

where $\boldsymbol{\nu}_{\mathcal{I}_i^n}$ denotes the unit normal to $\mathcal{I}_i^n$. Using the facts that $\boldsymbol{\tau} \cdot \boldsymbol{\nu}_{\mathcal{I}_i^n} = 0$ on the space-time edges $\partial\mathcal{I}_i^n \cap (\check{I}_i \times J^n)$ and b and c are constants, we see that

$$\begin{aligned}
&\int_{\mathcal{I}_i^n} \left(c\frac{\partial p}{\partial t} + b\frac{\partial p}{\partial x} \right) v \; dx \; dt \\
&\quad = \int_{I_i} cp^n v^n \; dx - \int_{\check{I}_i(t^{n-1})} cp^{n-1} v^{n-1,+} \; dx \\
&\qquad - \int_{\mathcal{I}_i^n} p\boldsymbol{\tau} \cdot \left(\frac{\partial v}{\partial x}, \frac{\partial v}{\partial t} \right) \; dx \; dt \; ,
\end{aligned} \tag{5.32}$$

where $v^{n-1,+} = v(x, t^{n-1,+}) = \lim_{\epsilon \to 0^+} v(x, t^{n-1} + \epsilon)$ to take account of the fact that $v(x,t)$ can be discontinuous at time levels. Analogously, the diffusion part of (5.28) yields, by Green's formula (1.19) in space,

$$\begin{aligned}
\int_{\mathcal{I}_i^n} \frac{\partial}{\partial x} \left(a\frac{\partial p}{\partial x} \right) v \; dx \; dt &= \int_{J^n} \int_{\check{I}_i(t)} \frac{\partial}{\partial x} \left(a\frac{\partial p}{\partial x} \right) v \; dx \; dt \\
&= \int_{J^n} \left(\int_{\partial \check{I}_i(t)} a\frac{\partial p}{\partial x} \nu_{\check{I}_i(t)} v \; d\ell - \int_{\check{I}_i(t)} a\frac{\partial p}{\partial x}\frac{\partial v}{\partial x} \; dx \right) dt \; .
\end{aligned} \tag{5.33}$$

The test function v is chosen by the following rule:

$$\boldsymbol{\tau} \cdot \left(\frac{\partial v}{\partial x}, \frac{\partial v}{\partial t} \right) = c\frac{\partial v}{\partial t} + b\frac{\partial v}{\partial x} = 0 \quad \text{ on } \mathcal{I}_i^n \; ; \tag{5.34}$$

that is, it is constant along characteristics. Using (5.28) and (5.34) and adding (5.32) and (5.33) yield

$$\begin{aligned}
&\int_{I_i} cp^n v^n \; dx + \int_{\mathcal{I}_i^n} a\frac{\partial p}{\partial x}\frac{\partial v}{\partial x} \; dx \; dt + \int_{\mathcal{I}_i^n} Rpv \; dx \; dt \\
&\quad - \int_{J^n} \int_{\partial \check{I}_i(t)} a\frac{\partial p}{\partial x} \nu_{\check{I}_i(t)} v \; d\ell \; dt \\
&\qquad = \int_{\check{I}_i(t^{n-1})} cp^{n-1} v^{n-1,+} \; dx + \int_{\mathcal{I}_i^n} fv \; dx \; dt \; .
\end{aligned} \tag{5.35}$$

This is an interior ELLAM formulation.

5.3.1.2 Left Boundary

There are four different types of elements at the left boundary, as illustrated in Fig. 5.7. We study the first type in detail; the second has the form of the first with $\hat{x}_0 = x_i$, where $\hat{x}_0$ is the head (at level t^n) of the characteristic corresponding to x_0 at the foot (at level t^{n-1}), as shown in Fig. 5.7, and

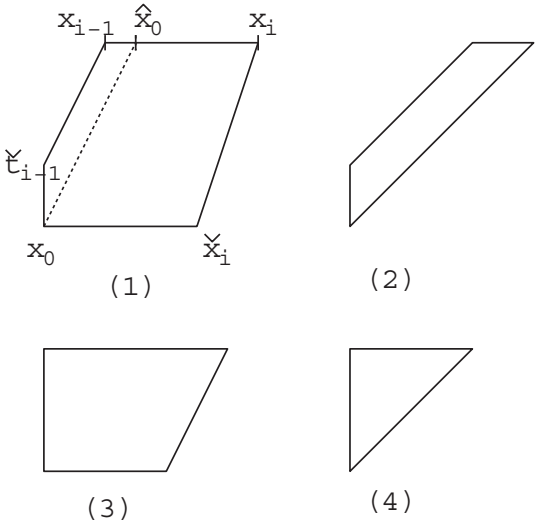

Fig. 5.7. Four types of elements meeting the left boundary

the third and fourth have the form of the first and second, respectively, with $i = 1$.

Using (5.34), as for the development of (5.32), we see that (cf. Fig. 5.7 and Exercise 5.5)

$$\begin{aligned} &\int_{\mathcal{I}_i^n} \left(c\frac{\partial p}{\partial t} + b\frac{\partial p}{\partial x} \right) v \; dx \; dt \\ &= \int_{I_i} cp^n v^n \; dx - \int_{x_0}^{\check{x}_i} cp^{n-1} v^{n-1,+} \; dx - \int_{t^{n-1}}^{\check{t}_{i-1}} (bpv)|_{x=0} \; dt \; . \end{aligned} \tag{5.36}$$

Similarly, we see that

$$\begin{aligned} &\int_{\mathcal{I}_i^n} \frac{\partial}{\partial x} \left(a\frac{\partial p}{\partial x} \right) v \; dx \; dt \\ &= \int_{J^n \times \partial \check{I}_i(t) \setminus (t^{n-1}, \check{t}_{i-1})} a\frac{\partial p}{\partial x} \nu_{\check{I}_i(t)} v \; d\ell \; dt \\ &\quad - \int_{t^{n-1}}^{\check{t}_{i-1}} \left(a\frac{\partial p}{\partial x} v \right) \Big|_{x=0} dt - \int_{\mathcal{I}_i^n} a\frac{\partial p}{\partial x}\frac{\partial v}{\partial x} \; dx \; dt \; . \end{aligned} \tag{5.37}$$

If a flux boundary condition is used at the left-hand end, we combine (5.28), (5.29), (5.36), and (5.37) to get

$$\begin{aligned}
&\int_{I_i} cp^n v^n \; dx + \int_{\mathcal{I}_i^n} a\frac{\partial p}{\partial x}\frac{\partial v}{\partial x} \; dx \; dt + \int_{\mathcal{I}_i^n} Rpv \; dx \; dt \\
&\quad - \int_{J^n \times \partial \check{I}_i(t) \setminus (t^{n-1}, \check{t}_{i-1})} a\frac{\partial p}{\partial x}\nu_{\check{I}_i(t)} v \; d\ell \; dt \\
&= \int_{x_0}^{\check{x}_i} cp^{n-1} v^{n-1,+} \; dx + \int_{\mathcal{I}_i^n} fv \; dx \; dt \\
&\quad + \int_{t^{n-1}}^{\check{t}_{i-1}} g_0(t) v(0,t) \; dt \; .
\end{aligned} \tag{5.38}$$

If a Dirichlet boundary condition occurs at $x = 0$, a different treatment from (5.37) is employed. We use backward Euler time integration along characteristics to see that (cf. (5.30) and Fig. 5.7)

$$\begin{aligned}
\int_{\mathcal{I}_i^n} \frac{\partial}{\partial x}\left(a\frac{\partial p}{\partial x}\right) v \; dx \; dt &= \int_{J^n} \int_{\check{I}_i(t)} \frac{\partial}{\partial x}\left(a\frac{\partial p}{\partial x}\right) v \; dx \; dt \\
&= \int_{I_i} \frac{\partial}{\partial x}\left(a^n \frac{dp^n}{dx}\right) v^n \Delta t^n(x) \; dx \; ,
\end{aligned} \tag{5.39}$$

where $\Delta t^n(x) = t^n - \check{t}(x)$ if $x < \hat{x}_0$, taking account of the reduced elapsed time along a characteristic that meets the boundary. The x-dependent Δt^n seems quite appropriate, since the diffusion at each point is weighted by the length of time over which it acts. Applying integration by parts to the last term of (5.39), we see that

$$\begin{aligned}
\int_{I_i} \frac{\partial}{\partial x}\left(a^n \frac{dp^n}{dx}\right) v^n \Delta t^n(x) \; dx &= \left(a^n \frac{dp^n}{dx} v^n \Delta t^n(x)\right)\Big|_{x_{i-1}}^{x_i} \\
&\quad - \int_{I_i} a^n \frac{dp^n}{dx}\left(\frac{dv^n}{dx}\Delta t^n(x) + v^n \frac{d\Delta t^n}{dx}\right) dx \; .
\end{aligned} \tag{5.40}$$

Note that $\Delta t^n(x_0) = 0$. Also, it follows from (5.30) that

$$\frac{d\Delta t^n}{dx} = \frac{1}{\varphi} \text{ if } x < \hat{x}_0 \text{ and } \frac{d\Delta t^n}{dx} = 0 \text{ if } x \geq \hat{x}_0 \; . \tag{5.41}$$

Now, we combine (5.28), (5.29), (5.36), and (5.39)–(5.41) to have

$$\begin{aligned}
&\int_{I_i} cp^n v^n \; dx + \int_{I_i} a^n \frac{dp^n}{dx}\frac{dv^n}{dx}\Delta t^n(x) \; dx + \int_{\mathcal{I}_i^n} Rpv \; dx \; dt \\
&\quad + \int_{x_{i-1}}^{\check{x}_0} a^n \frac{dp^n}{dx}\frac{v^n}{\varphi} \; dx - \left(a^n \frac{dp^n}{dx} v^n \Delta t^n(x)\right)\Big|_{x_{i-1}}^{x_i} \\
&= \int_{x_0}^{\check{x}_i} cp^{n-1} v^{n-1,+} \; dx + \int_{\mathcal{I}_i^n} fv \; dx \; dt \\
&\quad + \int_{t^{n-1}}^{\check{t}_{i-1}} b(t) g_0(t) v(0,t) \; dt \; .
\end{aligned} \tag{5.42}$$

Note that the factor of $1/\varphi$ appears in (5.42). This does not cause trouble, because the integration is over an interval of length at most $\varphi\Delta t^n$. As in Chap. 1, the Dirichlet boundary condition is essential so that $p^n(0)$ is not solved for and is assigned from a boundary datum. The flux boundary condition is a natural condition, so $p^n(0)$ needs to be obtained as an unknown. We have not discussed a Neumann boundary condition. In practice, it is unlikely that this condition would be physically appropriate for problem (5.28).

5.3.1.3 Right Boundary and Hyperbolic Case

The treatment of the right boundary is somewhat more involved. We define the *Courant number*

$$Ku = \frac{\varphi\Delta t^n}{h} , \tag{5.43}$$

and let $[Ku]$ be the integer part of this number. For $j = M, M+1, \ldots, M+[Ku]-1$, set

$$t_j = t^n - \frac{(j-M)h}{\varphi} ,$$

and $t_{M+[Ku]} = t^{n-1}$. Thus $[t^{n-1}, t^n]$ is divided into $[Ku]$ subintervals, backward in time, with the first $[Ku]-1$ of length $\Delta t^n/Ku$ and the last of length $\big(Ku-[Ku]+1\big)\Delta t^n/Ku$ (up to twice the size of the others). Alternatively, we may set $t_{M+[Ku]} = t^n - [Ku]h/\varphi$ and if $Ku - [Ku] > 0$, we define $t_{M+[Ku]+1} = t^{n-1}$. The treatment of these two cases is similar, and we consider the former case.

There are two types of elements at the right-hand end; see Fig. 5.8. Because the second has the form of the first with $t_j = t^{n-1}$, we study the first only. As for (5.36), we have (cf. Exercise 5.6)

$$\begin{aligned} &\int_{\mathcal{I}_j^n} \left(c\frac{\partial p}{\partial t} + b\frac{\partial p}{\partial x} \right) v \; dx \; dt \\ &\quad = -\int_{\check{x}_{j-1}}^{\check{x}_j} cp^{n-1}v^{n-1,+} \; dx + \int_{t_j}^{t_{j-1}} (bpv)|_{x=1} \; dt , \end{aligned} \tag{5.44}$$

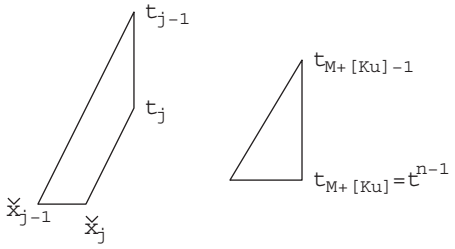

Fig. 5.8. Two types of elements meeting the right boundary

and as for (5.37),

$$\begin{aligned}
&\int_{\mathcal{I}_i^n} \frac{\partial}{\partial x}\left(a\frac{\partial p}{\partial x}\right) v \; dx \; dt \\
&\quad = \int_{J^n \times \partial \check{I}_j(t) \setminus (t_j, t_{j-1})} a\frac{\partial p}{\partial x} \nu_{\check{I}_j(t)} v \; d\ell \; dt \\
&\qquad + \int_{t_j}^{t_{j-1}} \left(a\frac{\partial p}{\partial x} v\right)\Big|_{x=1} dt - \int_{\mathcal{I}_i^n} a\frac{\partial p}{\partial x}\frac{\partial v}{\partial x} \; dx \; dt \; .
\end{aligned} \tag{5.45}$$

We combine (5.28), (5.44), and (5.45) to obtain

$$\begin{aligned}
&\int_{t_j}^{t_{j-1}} \left(bpv - a\frac{\partial p}{\partial x} v\right)\Big|_{x=1} dt + \int_{\mathcal{I}_i^n} a\frac{\partial p}{\partial x}\frac{\partial v}{\partial x} \; dx \; dt \\
&\quad + \int_{\mathcal{I}_j^n} Rpv \; dx \; dt - \int_{J^n \times \partial \check{I}_j(t) \setminus (t_j, t_{j-1})} a\frac{\partial p}{\partial x} \nu_{\check{I}_i(t)} v \; d\ell \; dt \\
&= \int_{\check{x}_{j-1}}^{\check{x}_j} cp^{n-1} v^{n-1,+} \; dx + \int_{\mathcal{I}_j^n} fv \; dx \; dt \; .
\end{aligned} \tag{5.46}$$

If a flux boundary condition is used at the right-hand end, we combine (5.29) and (5.46) to see that

$$\begin{aligned}
&\int_{\mathcal{I}_i^n} a\frac{\partial p}{\partial x}\frac{\partial v}{\partial x} \; dx \; dt + \int_{\mathcal{I}_j^n} Rpv \; dx \; dt \\
&\qquad - \int_{J^n \times \partial \check{I}_j(t) \setminus (t_j, t_{j-1})} a\frac{\partial p}{\partial x} \nu_{\check{I}_i(t)} v \; d\ell \; dt \\
&\quad = \int_{\check{x}_{j-1}}^{\check{x}_j} cp^{n-1} v^{n-1,+} \; dx + \int_{\mathcal{I}_j^n} fv \; dx \; dt \\
&\qquad - \int_{t_j}^{t_{j-1}} g_1(t) v(1,t) \; dt \; .
\end{aligned} \tag{5.47}$$

For a Dirichlet boundary, we use (5.29) and (5.46) to get

$$\begin{aligned}
&\int_{\mathcal{I}_i^n} a\frac{\partial p}{\partial x}\frac{\partial v}{\partial x} \; dx \; dt - \int_{t_j}^{t_{j-1}} \left(a\frac{\partial p}{\partial x} v\right)\Big|_{x=1} dt \\
&\quad + \int_{\mathcal{I}_j^n} Rpv \; dx \; dt - \int_{J^n \times \partial \check{I}_j(t) \setminus (t_j, t_{j-1})} a\frac{\partial p}{\partial x} \nu_{\check{I}_i(t)} v \; d\ell \; dt \\
&= \int_{\check{x}_{j-1}}^{\check{x}_j} cp^{n-1} v^{n-1,+} \; dx + \int_{\mathcal{I}_j^n} fv \; dx \; dt \\
&\qquad - \int_{t_j}^{t_{j-1}} b(t) g_1(t) v(1,t) \; dt \; .
\end{aligned} \tag{5.48}$$

In this case we see that $\partial p/\partial x$ is also an unknown at the right-hand boundary.

In the *purely hyperbolic case* where $a = 0$, equation (5.46) becomes

$$\begin{aligned}\int_{t_j}^{t_{j-1}} (bpv)\,\big|_{x=1}\; dt + \int_{\mathcal{I}_j^n} Rpv\; dx\; dt \\ = \int_{\check{x}_{j-1}}^{\check{x}_j} cp^{n-1}v^{n-1,+}\; dx + \int_{\mathcal{I}_j^n} fv\; dx\; dt\;.\end{aligned} \tag{5.49}$$

Thus the ELLAM naturally handles the hyperbolic case without an artificial boundary condition.

5.3.1.4 Conservation of Mass

In addition to the requirement (5.34) for the test functions v, we also assume that their sum over $I \times J^n$ is identically one. Then the addition of (5.35)–(5.37) and (5.46) implies

$$\begin{aligned}\int_I cp^n\; dx - \int_I cp^{n-1}\; dx + \int_{I\times J^n} Rp\; dx\; dt = \int_{I\times J^n} f\; dx\; dt \\ - \int_{J^n} \left(bp - a\frac{\partial p}{\partial x}\right)\Big|_{x=1} dt + \int_{J^n} \left(bp - a\frac{\partial p}{\partial x}\right)\Big|_{x=0} dt\;,\end{aligned} \tag{5.50}$$

where we assumed the continuity of $a\partial p/\partial x$ on $I \times J^n$. Relation (5.50) is precisely the statement of global mass conservation. To obtain (5.50) exactly in an implementation, some care needs to be taken in the consistent evaluation of integrals; the integrals at level t^{n-1} in the ELLAM formulation must be evaluated so that they sum to the integral at this time level in (5.50).

5.3.1.5 Test Functions

So far we have not specified the space-time test functions in the ELLAM formulation. The most natural choice is continuous piecewise-linear elements, which we considered in Chap. 1. Of course, nothing prevents us using test functions of higher order (cf. Exercise 5.7).

From the analysis in Sects. 5.3.1.1–5.3.1.4, the only requirements on test functions are that they satisfy (5.34) and that their sum over $I \times J^n$ is identically one. We assume that $\mathcal{I}_i^n \cup \mathcal{I}_{i+1}^n$ does not meet the boundary of I. Then a linear test function v satisfying such assumptions on $\mathcal{I}_i^n \cup \mathcal{I}_{i+1}^n$ is

$$v(x,t) = \begin{cases} \dfrac{x - x_{i-1}}{h_i} + \varphi\dfrac{t^n - t}{h_i}, & (x,t) \in \mathcal{I}_i^n\;, \\ \dfrac{x_{i+1} - x}{h_{i+1}} - \varphi\dfrac{t^n - t}{h_{i+1}}, & (x,t) \in \mathcal{I}_{i+1}^n\;, \\ 0, & \text{all other } (x,t)\;. \end{cases} \tag{5.51}$$

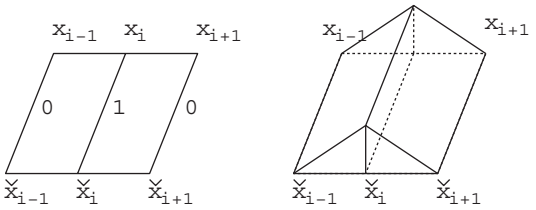

Fig. 5.9. Interior test functions

This definition is shown in Fig. 5.9.

The definition of a test function near the boundary of I is somewhat more involved. For a flux condition at the left end, the test functions, through v_2, are illustrated in Fig. 5.10. If this boundary condition is Dirichlet, the test functions are displayed in Fig. 5.11, since there is no degree of freedom at x_0 (cf. (5.42)). The only test function on $\mathcal{I}_1^n$ is $v_1 \equiv 1$. In these two figures, $1 < Ku < 2$ is considered for illustration.

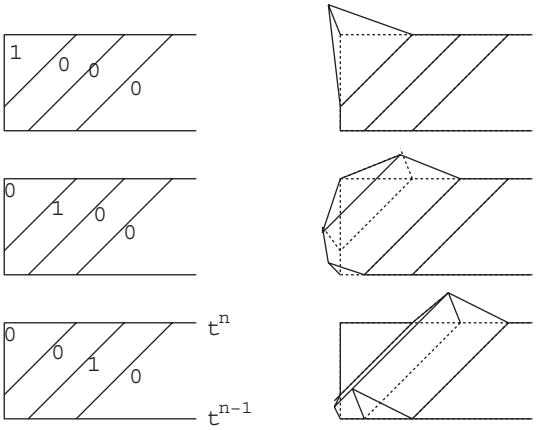

Fig. 5.10. Test functions for the left-hand flux boundary

At the right end, because the solution at point $(x_M, t_{M+[Ku]}) = (x_M, t^{n-1})$ is known from the previous time level, we do not solve for an unknown associated with $t_{M+[Ku]}$, so the element $\mathcal{I}^n_{M+[Ku]}$ has the single test function $v_{M+[Ku]-1} \equiv 1$, instead of two, as shown in Fig. 5.12, where the case $2 < Ku < 3$ is illustrated for the last two test functions v_M and v_{M+1}.

5.3.1.6 An ELLAM Procedure

We consider the case where the left and right boundaries are of the flux type in detail. We add (5.35), (5.38), and (5.47) and use the continuity of $a\partial p/\partial x$

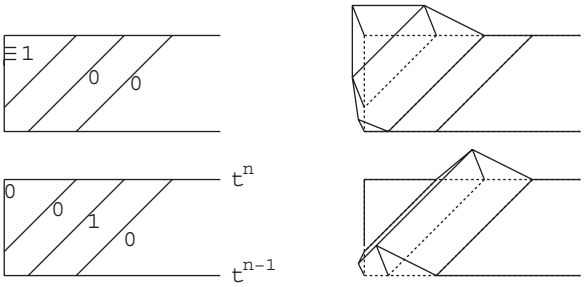

Fig. 5.11. Test functions for the left-hand Dirichlet boundary

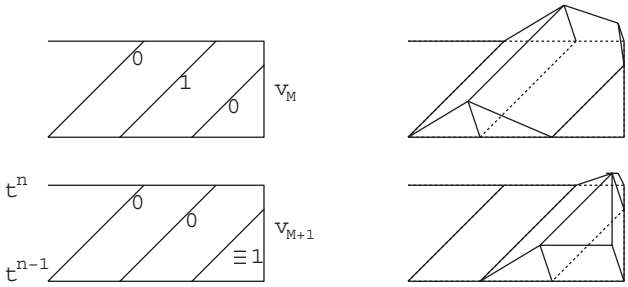

Fig. 5.12. Test functions for the right-hand boundary

on $I \times J^n$ to see that

$$\begin{aligned}
&\int_I cp^n v^n \, dx + \int_{J^n}\int_I a\frac{\partial p}{\partial x}\frac{\partial v}{\partial x} \, dx \, dt + \int_{J^n}\int_I Rpv \, dx \, dt \\
&\quad = \int_I cp^{n-1}v^{n-1,+} \, dx + \int_{J^n}\int_I fv \, dx \, dt \\
&\qquad + \int_{J^n} g_0(t)v(0,t) \, dt - \int_{J^n} g_1(t)v(1,t) \, .
\end{aligned} \tag{5.52}$$

If we apply backward Euler time integration along characteristics to the diffusion, reaction, and source term in (5.52), we obtain

$$\begin{aligned}
&\int_I cp^n v^n \, dx + \int_I \Delta t^n(x) a^n \frac{dp^n}{dx}\frac{dv^n}{dx} \, dx \\
&\quad + \int_I \Delta t^n(x) R^n p^n v^n \, dx = \int_I cp^{n-1}v^{n-1,+} \, dx \\
&\quad + \int_I \Delta t^n(x) f^n v^n \, dx + \int_{J^n} g_0(t)v(0,t) \, dt - \int_{J^n} g_1(t)v(1,t) \, ,
\end{aligned} \tag{5.53}$$

where we recall that $\Delta t^n(x) = t^n - \check{t}(x)$ if $x < \hat{x}_0$ and $\Delta t^n = t^n - t^{n-1}$ otherwise. It follows from (5.53) that it suffices to define trial functions at

the discrete time level t^n only. In view of the test functions, a natural choice is continuous piecewise-linear functions in x. Thus we define

$$V_h = \{w : w \text{ is a continuous function on } I \\ \text{and } w|_{I_i} \text{ is linear, } i = 1, 2, \ldots, M\} .$$

Now, an ELLAM procedure is defined: For $n = 1, 2, \ldots$, find $p_h^n \in V_h$ such that

$$\begin{aligned} &\int_I c p_h^n v^n \, dx + \int_I \Delta t^n(x) a^n \frac{dp_h^n}{dx} \frac{dv^n}{dx} \, dx \\ &+ \int_I \Delta t^n(x) R^n p_h^n v^n \, dx = \int_I c p_h^{n-1} v^{n-1,+} \, dx \\ &+ \int_I \Delta t^n(x) f^n v^n \, dx + \int_{J^n} g_0(t) v(0,t) \, dt - \int_{J^n} g_1(t) v(1,t) , \end{aligned} \tag{5.54}$$

for all test functions v in the previous subsection. In the present flux case, the unknowns are $p_h^n(x_0)$, $p_h^n(x_1)$,..., $p_h^n(x_M)$. If desired, the unknowns $p_h(x_M, t_{M+1})$, $p_h(x_M, t_{M+2})$, ..., $p_h(x_M, t_{M+[Ku]-1})$ can be also obtained using equations at the right boundary in Sect. 5.3.1.3.

If a Dirichlet condition is exploited, a similar development can be done (cf. Exercise 5.8). A Dirichlet left boundary removes $p_h^n(x_0)$ as an unknown (cf. (5.42)). A Dirichlet right boundary replaces $p_h^n(x_M)$ with $\frac{dp_h^n}{dx}(x_M)$ as an unknown (cf. (5.48)). Again, if needed, the unknowns

$$\frac{dp_h}{dx}(x_M, t_{M+1}), \frac{dp_h}{dx}(x_M, t_{M+2}), \ldots, \frac{dp_h}{dx}(x_M, t_{M+[Ku]-1})$$

can be obtained using equations in Sect. 5.3.1.3. Since p_h is represented by piecewise-linear trial functions, we could consider linear ones for $\frac{dp_h}{dx}$ at the right boundary, but due to the expected loss of one order of accuracy in passing from p_h to $\frac{dp_h}{dx}$, piecewise constants are more suitable.

We end with a remark that test functions can be obtained from the trial functions. For any $w \in V_h$, we define $v(x,t)$ to be a constant extension of $w(x)$ into the space-time region $I \times J^n$ along characteristics (cf. (5.30) and (5.31)):

$$\begin{aligned} &v(\check{x}_n(x,t),t) = w(x), \qquad t \in [\check{t}(x), t^n], \ x \in I , \\ &v(\check{x}_n(1,\theta),\theta) = w(x), \qquad \theta \in [t^{n-1}, t] . \end{aligned} \tag{5.55}$$

The remarks made for (5.11) in Sect. 5.2.1 on the condition number of the stiffness matrix and the error estimate (5.14) apply to (5.54) (cf. Theorem 5.1 and Exercises 5.9 and 5.10).

5.3.2 Extension to Multi-Dimensional Problems

We now extend the ELLAM to a multi-dimensional problem:

$$\begin{aligned} &\frac{\partial(cp)}{\partial t} + \nabla \cdot \big(\mathbf{b}p - \mathbf{a}\nabla p\big) + Rp = f, && \mathbf{x} \in \Omega,\ t > 0\ , \\ &(\mathbf{b}p - \mathbf{a}\nabla p) \cdot \boldsymbol{\nu} = g, && \mathbf{x} \in \Gamma,\ t > 0\ , \\ &p(\mathbf{x}, 0) = p_0(\mathbf{x}), && \mathbf{x} \in \Omega\ , \end{aligned} \tag{5.56}$$

where $\Omega \subset \mathbb{R}^d$ ($d \leq 3$) is a bounded domain and $c = c(\mathbf{x}, t)$ and $\mathbf{b} = \mathbf{b}(\mathbf{x}, t)$ are now variable. We consider a flux boundary condition in (5.56), with $g(\cdot, t) \in H^{-1/2}(\Gamma)$, $t > 0$. An extension to a Dirichlet condition can be made as in the previous subsection.

For any $\mathbf{x} \in \Omega$ and two times $0 \leq t^{n-1} < t^n$, the hyperbolic part of problem (5.56), $c\partial p/\partial t + \mathbf{b} \cdot \nabla p$, defines the characteristic $\check{\mathbf{x}}_n(\mathbf{x}, t)$ along the interstitial velocity $\boldsymbol{\varphi} = \mathbf{b}/c$ (cf. Fig. 5.4):

$$\begin{aligned} &\frac{\partial}{\partial t}\check{\mathbf{x}}_n = \boldsymbol{\varphi}(\check{\mathbf{x}}_n, t), && t \in J^n\ , \\ &\check{\mathbf{x}}_n(\mathbf{x}, t^n) = \mathbf{x}\ . \end{aligned} \tag{5.57}$$

In general, the characteristics in (5.57) can be determined only approximately. There are many methods to solve this first-order ordinary differential equation for the approximate characteristics. We consider only the Euler method. Other methods, such as improved Euler and Runge-Kutta (cf. Sect. 10.3.2) methods, can be applied.

The Euler method to solve (5.57) for the approximate characteristics is given: For any $\mathbf{x} \in \Omega$, we define

$$\check{\mathbf{x}}_n(\mathbf{x}, t) = \mathbf{x} - \boldsymbol{\varphi}(\mathbf{x}, t^n)(t^n - t), \qquad t \in [\check{t}(\mathbf{x}), t^n]\ , \tag{5.58}$$

where $\check{t}(\mathbf{x}) = t^{n-1}$ if $\check{\mathbf{x}}_n(\mathbf{x}, t)$ does not backtrack to the boundary Γ for $t \in [t^{n-1}, t^n]$; $\check{t}(\mathbf{x}) \in (t^{n-1}, t^n]$ is the time instant when $\check{\mathbf{x}}_n(\mathbf{x}, t)$ intersects Γ, i.e., $\check{\mathbf{x}}_n(\mathbf{x}, \check{t}(\mathbf{x})) \in \Gamma$, otherwise. Let

$$\Gamma_+ = \{\mathbf{x} \in \Gamma :\ (\mathbf{b} \cdot \boldsymbol{\nu})(\mathbf{x}) \geq 0\}\ .$$

For $(\mathbf{x}, t) \in \Gamma_+ \times J^n$, the approximate characteristic emanating backward from $(\mathbf{x}, t)$ is given by

$$\check{\mathbf{x}}_n(\mathbf{x}, \theta) = \mathbf{x} - \boldsymbol{\varphi}(\mathbf{x}, t)(t - \theta), \qquad \theta \in [\check{t}(\mathbf{x}, t), t]\ , \tag{5.59}$$

where $\check{t}(\mathbf{x}, t) = t^{n-1}$ if $\check{\mathbf{x}}_n(\mathbf{x}, \theta)$ does not backtrack to the boundary Γ for $\theta \in [t^{n-1}, t]$; $\check{t}(\mathbf{x}, t) \in (t^{n-1}, t]$ is the time instant when $\check{\mathbf{x}}_n(\mathbf{x}, \theta)$ intersects Γ, otherwise. These characteristics are defined in the same way as in (5.30) and (5.31). We have exploited a single step Euler method to determine the approximate characteristics from (5.57); a multi-step version can be also employed.

If Δt^n is sufficiently small (depending upon the smoothness of $\boldsymbol{\varphi}$), the approximate characteristics do not cross each other, which is assumed. Then

$\check{\mathbf{x}}_n(\cdot,t)$ is a one-to-one mapping of $\mathbb{R}^d$ to $\mathbb{R}^d$ ($d \le 3$); we indicate its inverse by $\hat{\mathbf{x}}_n(\cdot,t)$.

For any $t \in (t^{n-1}, t^n]$, we define

$$\tilde{\boldsymbol{\varphi}}(\mathbf{x},t) = \boldsymbol{\varphi}(\hat{\mathbf{x}}_n(\mathbf{x},t), t^n), \quad \tilde{\mathbf{b}} = \tilde{\boldsymbol{\varphi}} c \,. \tag{5.60}$$

We assume that $\tilde{\mathbf{b}} \cdot \boldsymbol{\nu} \ge 0$ on Γ_+.

Let K_h be a partition of Ω into elements $\{K\}$. For each $K \in K_h$, let $\check{K}(t)$ represent the trace-back of K to time t, $t \in J^n$:

$$\check{K}(t) = \{\mathbf{x} \in \Omega : \mathbf{x} = \check{\mathbf{x}}_n(\mathbf{y},t) \text{ for some } \mathbf{y} \in K\} \,,$$

and $\mathcal{K}^n$ be the space-time region that follows the characteristics (cf. Fig. 5.13):

$$\mathcal{K}^n = \{(\mathbf{x},t) \in \Omega \times J : t \in J^n \text{ and } \mathbf{x} \in \check{K}(t)\} \,.$$

Also, we define $\mathcal{B}^n = \{(\mathbf{x},t) \in \partial\mathcal{K}^n : \mathbf{x} \in \partial\Omega\}$.

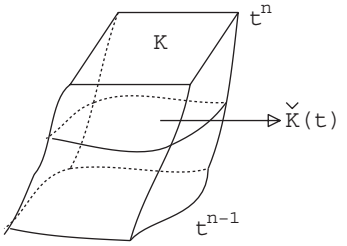

Fig. 5.13. An illustration of $\mathcal{K}^n$

We write the hyperbolic part of (5.56) as

$$\frac{\partial(cp)}{\partial t} + \nabla \cdot \big(\mathbf{b}p\big) = \frac{\partial(cp)}{\partial t} + \nabla \cdot \big(\tilde{\mathbf{b}}p\big) + \nabla \cdot \big([\mathbf{b} - \tilde{\mathbf{b}}]p\big) \,. \tag{5.61}$$

With $\boldsymbol{\tau}(x,t) = (\tilde{\mathbf{b}}, c)$ and a smooth test function $v(\mathbf{x},t)$, application of Green's formula in space and time gives

$$\begin{aligned} &\int_{\mathcal{K}^n} \left(\frac{\partial(cp)}{\partial t} + \nabla \cdot \big(\tilde{\mathbf{b}}p\big) \right) v \; d\mathbf{x} \; dt \\ &\quad = \int_K c^n p^n v^n \; d\mathbf{x} - \int_{\check{K}(t^{n-1})} c^{n-1} p^{n-1} v^{n-1,+} \; d\mathbf{x} \\ &\quad + \int_{\mathcal{B}^n} p\tilde{\mathbf{b}} \cdot \boldsymbol{\nu} v \; d\ell - \int_{\mathcal{K}^n} p\boldsymbol{\tau} \cdot \left(\nabla v, \frac{\partial v}{\partial t} \right) \; d\mathbf{x} \; dt, \end{aligned} \tag{5.62}$$

where we used the fact that $\boldsymbol{\tau} \cdot \boldsymbol{\nu}_{\mathcal{K}^n} = 0$ on the space-time edges $(\partial\mathcal{K}^n \cap (\check{K} \times J^n)) \setminus \mathcal{B}^n$. The establishment of (5.62) is analogous to that of (5.32); here we do not distinguish between interior and boundary elements.

Similarly, the diffusion part of (5.56) gives

$$\begin{aligned}&\int_{\mathcal{K}^n} \nabla\cdot\big(\mathbf{a}\nabla p\big)v\ d\mathbf{x}\ dt\\ &\quad=\int_{J^n}\left\{\int_{\partial\check{K}(t)}\mathbf{a}\nabla p\cdot\boldsymbol{\nu}_{\check{K}(t)}v\ d\ell-\int_{\check{K}(t)}(\mathbf{a}\nabla p)\cdot\nabla v\ d\mathbf{x}\right\}dt\ .\end{aligned}\tag{5.63}$$

We assume that the test function $v(\mathbf{x},t)$ is constant along the approximate characteristics. Then, combining (5.61)–(5.63) and using the same technique as for (5.52), the space-time variational form of (5.56) can be derived as follows:

$$\begin{aligned}&(c^np^n,v^n)-\big(c^{n-1}p^{n-1},v^{n-1,+}\big)\\ &+\int_{J^n}\{(\mathbf{a}\nabla p,\nabla v)+(Rp,v)\}\ dt=\int_{J^n}\{(f,v)-(g,v)_\Gamma\}\ dt\\ &+\int_{J^n}\left\{\left(\nabla\cdot\left[(\tilde{\mathbf{b}}-\mathbf{b})p\right],\hat{v}\right)-\left(p\left[\tilde{\mathbf{b}}-\mathbf{b}\right]\cdot\boldsymbol{\nu},v\right)_\Gamma\right\}dt\ ,\end{aligned}\tag{5.64}$$

where the inner product notation in space is used. If we apply backward Euler time integration along characteristics to the diffusion, reaction, and source term in (5.64), we see that

$$\begin{aligned}&(c^np^n,v^n)+(\Delta t^n\mathbf{a}^n\nabla p^n,\nabla v^n)+(\Delta t^nR^np^n,v^n)\\ &\quad=\big(c^{n-1}p^{n-1},v^{n-1,+}\big)+(\Delta t^nf^n,v^n)-\int_{J^n}(g,v)_\Gamma\ dt\\ &\qquad+\int_{J^n}\left\{\left(\nabla\cdot\left[(\tilde{\mathbf{b}}-\mathbf{b})p\right],\hat{v}\right)-\left(p\left[\tilde{\mathbf{b}}-\mathbf{b}\right]\cdot\boldsymbol{\nu},v\right)_\Gamma\right\}dt\ ,\end{aligned}\tag{5.65}$$

where $\Delta t^n(\mathbf{x})=t^n-\check{t}(\mathbf{x})$.

Let $V_h\subset H^1(\Omega)$ be a finite element space (cf. Chap. 1). For any $w\in V_h$, we define a test function $v(\mathbf{x},t)$ to be a constant extension of $w(\mathbf{x})$ into the space-time region $\Omega\times J^n$ along the approximate characteristics (refer to (5.58) and (5.59)):

$$\begin{aligned}&v(\check{\mathbf{x}}_n(\mathbf{x},t),t)=w(\mathbf{x}),\qquad t\in[\check{t}(\mathbf{x}),t^n],\ \mathbf{x}\in\Omega\ ,\\ &v(\check{\mathbf{x}}_n(\mathbf{x},\theta),\theta)=w(\mathbf{x}),\qquad \theta\in[\check{t}(\mathbf{x},t),t],\ (\mathbf{x},t)\in\Gamma_+\times J^n\ .\end{aligned}\tag{5.66}$$

Now, based on (5.65), an ELLAM procedure is defined: For $n=1,2,\ldots$, find $p_h^n\in V_h$ such that

$$\begin{aligned}&(c^np_h^n,v^n)+(\Delta t^n\mathbf{a}^n\nabla p_h^n,\nabla v^n)+(\Delta t^nR^np_h^n,v^n)\\ &\quad=\big(c^{n-1}p_h^{n-1},v^{n-1,+}\big)+(\Delta t^nf^n,v^n)-\int_{J^n}(g,v)_\Gamma\ dt\ .\end{aligned}\tag{5.67}$$

The remarks made on accuracy and efficiency of the MMOC apply to (5.67), too (cf. Exercise 5.11). In particular, when V_h is the space of piecewise linear functions defined on a regular triangulation K_h, the next theorem holds (Wang, 2000).

Theorem 5.1. *Assume that Ω is a convex polygonal domain or has a smooth boundary Γ, and the coefficients $\mathbf{a}$, $\mathbf{b}$, c, f, and R satisfy*

$$\begin{aligned} &\mathbf{a} \in \left(W^{1,\infty}(\Omega \times J)\right)^{d\times d}, \quad \mathbf{b} \in \left(W^{1,\infty}(\Omega \times J)\right)^{d}, \\ &c,\ f \in W^{1,\infty}(\Omega \times J), \qquad R \in L^{\infty}(J; W^{1,\infty}(\Omega)) . \end{aligned}$$

If the solution p to (5.56) satisfies $p \in L^{\infty}(J; W^{2,\infty}(\Omega))$ and $\partial p/\partial t \in L^2(J; H^2(\Omega))$, the initialization error satisfies

$$\|p_0 - p_h^0\|_{L^2(\Omega)} \le Ch^2 \|p_0\|_{H^2(\Omega)} ,$$

and Δt is sufficiently small, then

$$\begin{aligned} \max_{1\le n\le N} \|p^n - p_h^n\|_{L^2(\Omega)} \le C\Bigg\{ &\Delta t \Bigg(\left\| \frac{dp}{d\boldsymbol{\tau}} \right\|_{L^2(J;H^1(\Omega))} \\ &+ \|p\|_{L^\infty(J;W^{2,\infty}(\Omega))} + \left\| \frac{df}{d\boldsymbol{\tau}} \right\|_{L^2(\Omega\times J)} + \|f\|_{L^2(\Omega\times J)} \Bigg) \\ &+ h^2 \Bigg(\|p\|_{L^\infty(J;W^{2,\infty}(\Omega))} + \left\| \frac{\partial p}{\partial t} \right\|_{L^2(J;H^2(\Omega))} + \|p_0\|_{H^2(\Omega)} \Bigg) \Bigg\} , \end{aligned}$$

where p_h is the solution of (5.67), and for real numbers $q, r \ge 0$,

$$\begin{aligned} \|v\|_{L^2(J;W^{q,r}(\Omega))} &= \left\| \|v(\cdot,t)\|_{W^{q,r}(\Omega)} \right\|_{L^2(J)} , \\ \|v\|_{L^\infty(J;W^{q,r}(\Omega))} &= \max_{t\in J} \|v(\cdot,t)\|_{W^{q,r}(\Omega)} . \end{aligned}$$

We mention that with advection on the right-hand side of (5.67) only, the linear system arising from (5.67) is well suited for iterative linear solution algorithms in multiple space dimensions (cf. Sect. 1.10).

We end this section with an example.

Example 5.1. Consider the problem

$$c\frac{\partial p}{\partial t} + \nabla \cdot (\mathbf{b}p) + Rp = f, \quad \mathbf{x} \in \Omega,\ t \in J ,$$

where $\Omega = (-0.5, 0.5) \times (-0.5, 0.5)$. The initial condition is given by

$$p_0(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}_c\|^2}{2\sigma^2} \right) ,$$

where $\mathbf{x}_c$ and σ are the centered and standard deviations, respectively. The corresponding exact solution to this problem, with $f = 0$, is

$$p(\mathbf{x},t) = \exp\left(-\frac{\|\bar{\mathbf{x}}-\mathbf{x}_c\|^2}{2\sigma^2} - \int_0^t R\left(\mathbf{r}(\zeta,\bar{\mathbf{x}}),\zeta\right)\, d\zeta\right) ,$$

where

$$\bar{\mathbf{x}} = (\bar{x}_1,\bar{x}_2) = \big(x_1\cos(4t) + x_2\sin(4t), -x_1\sin(4t) + x_2\cos(4t)\big) ,$$
$$\mathbf{r}(\zeta,\bar{\mathbf{x}}) = \big(\bar{x}_1\cos(4\zeta) - \bar{x}_2\sin(4\zeta), \bar{x}_1\sin(4\zeta) + \bar{x}_2\cos(4\zeta)\big) .$$

This example can be viewed as an incompressible flow problem in a two-dimensional homogeneous medium with a known analytical solution, and has been widely utilized to test the performance of a numerical method. In the test here, the data are chosen as follows: $c = 1$, $R = f = 0$, $T = \pi/2$, $\mathbf{x}_c = (-0.25, 0)$, $\sigma = 0.0447$, and $\mathbf{b} = (-4x_2, 4x_1)$ (a rotating field). A uniform spatial grid is utilized, with the spatial steps in the x_1- and x_2-directions being $1/64$, and a fixed time step of length $\Delta t = \pi/32$ is used. A numerical result obtained using (5.67) is shown in Fig. 5.14. This figure shows that the peak of the solution is accurately captured by the ELLAM.

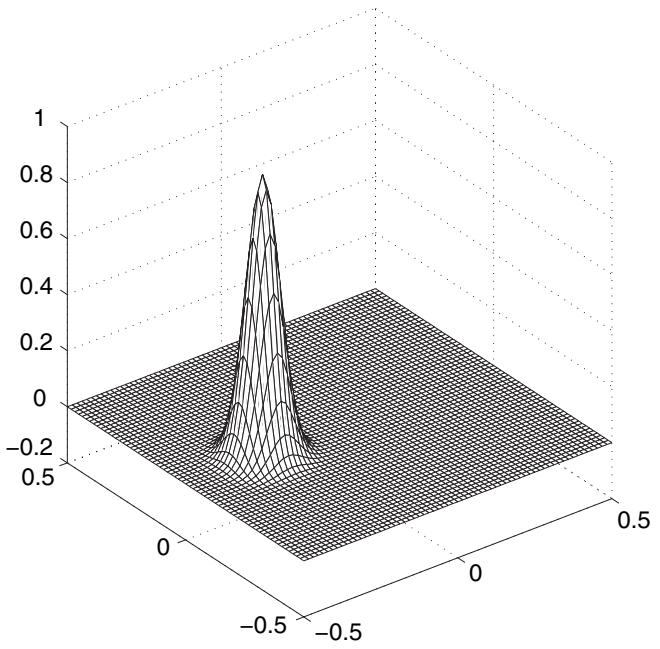

Fig. 5.14. A numerical result using the ELLAM

5.4 The Characteristic Mixed Method

In this section, we introduce the characteristic mixed method for the numerical solution of (5.56) (Yang, 1992; Arbogast-Wheeler, 1995). This method combines the ideas of the ELLAM and the mixed finite element method in Chap. 3; in time it adopts the ELLAM idea, and in space it is based on the mixed method. As in Chap. 3, introducing a new variable $\mathbf{u}$ in (5.56), this problem can be rewritten as

$$\begin{aligned}
&\frac{\partial(cp)}{\partial t} + \nabla \cdot \big(\mathbf{b}p - \mathbf{u}\big) + Rp = f && \text{in } \Omega \times J\ , \\
&\mathbf{u} = \mathbf{a}\nabla p && \text{in } \Omega \times J\ , \\
&(\mathbf{b}p - \mathbf{u}) \cdot \boldsymbol{\nu} = g_- && \text{on } \Gamma_- \times J\ , \\
&p = g_+ && \text{on } \Gamma_+ \times J\ , \\
&p(\mathbf{x}, 0) = p_0(\mathbf{x}) && \text{in } \Omega\ ,
\end{aligned} \tag{5.68}$$

where

$$\begin{aligned}
\Gamma_- &= \{\mathbf{x} \in \Gamma :\ (\mathbf{b} \cdot \boldsymbol{\nu})(\mathbf{x}) < 0\}\ , \\
\Gamma_+ &= \{\mathbf{x} \in \Gamma :\ (\mathbf{b} \cdot \boldsymbol{\nu})(\mathbf{x}) \geq 0\}\ ,
\end{aligned}$$

and $g_-(\cdot, t) \in H^{-1/2}(\Gamma_-)$ and $g_+(\cdot, t) \in H^{1/2}(\Gamma_+)$ $(t \in J)$ are given functions. Recall that Γ_- and Γ_+ are the *inflow and outflow boundaries* of Γ, respectively.

The spaces defined in Sect. 3.2 are used. In particular, we employ

$$W = L^2(\Omega) = \left\{ v :\ v \text{ is defined on } \Omega \text{ and } \int_\Omega v^2\ d\mathbf{x} < \infty \right\}\ ,$$

and, for $d \leq 3$,

$$\mathbf{V} = \mathbf{H}(\mathrm{div}, \Omega) = \left\{ \mathbf{v} \in (L^2(\Omega))^d : \nabla \cdot \mathbf{v} \in L^2(\Omega) \right\}\ .$$

The notation in the previous section is also used.

The space-time variational form of (5.68) is imposed in mixed form for $\mathbf{u} \in \mathbf{V}$ and $p \in W$. Replacing $\mathbf{b}$ by $\tilde{\mathbf{b}} + (\mathbf{b} - \tilde{\mathbf{b}})$ in (5.68) and following the argument for (5.64), with $\tilde{\mathbf{b}}$ given in (5.60), the first equation of (5.68) can be equivalently written as (cf. Exercise 5.13)

$$\begin{aligned}
&(c^n p^n, v^n) - \left(c^{n-1} p^{n-1}, v^{n-1,+}\right) - \int_{J^n} \{(\nabla \cdot \mathbf{u}, v) - (Rp, v)\}\ dt \\
&\quad = \int_{J^n} \{(f, v) - (g_- + \mathbf{u} \cdot \boldsymbol{\nu}, v)_{\Gamma_-} - (g_+ \tilde{\mathbf{b}} \cdot \boldsymbol{\nu}, v)_{\Gamma_+}\}\ dt \\
&\qquad + \int_{J^n} \{(\nabla \cdot [(\tilde{\mathbf{b}} - \mathbf{b})p], v) - (p[\tilde{\mathbf{b}} - \mathbf{b}] \cdot \boldsymbol{\nu}, v)_{\Gamma_-}\}\ dt\ ,
\end{aligned} \tag{5.69}$$

where the test function $v(x,t)$ is assumed to be constant along the approximate characteristics.

Invert $\mathbf{a}$ in the second equation of (5.68), multiply the resulting equation by $\mathbf{v} \in \mathbf{V}$, and use Green's formula (1.19) in space to see that

$$\left(\mathbf{a}^{-1}\mathbf{u}, \mathbf{v}\right) - \left(p, \mathbf{v}\cdot\boldsymbol{\nu}\right)_{\Gamma_-} + \left(p, \nabla\cdot\mathbf{v}\right) = \left(g_+, \mathbf{v}\cdot\boldsymbol{\nu}\right)_{\Gamma_+} . \tag{5.70}$$

Equations (5.69) and (5.70) are the characteristic mixed variational form of (5.68).

Note that it is difficult to approximate conservatively the inflow boundary conditions in these two equations since the unknown solution $\mathbf{u}, p$ appears in the integrals over Γ_-. To rectify this, let K_h be a partition of Ω into elements $\{K\}$. For each $K \in K_h$, let $\check{K}(t)$ represent the trace-back of K to time t, $t \in J^n$ (cf. Fig. 5.13):

$$\check{K}(t) = \{\mathbf{x} \in \Omega : \mathbf{x} = \check{\mathbf{x}}_n(\mathbf{y}, t) \text{ for some } \mathbf{y} \in K\} .$$

We apply Green's formula (1.19) in space on each K to the third term on the left-hand side of (5.69) to see that

$$\begin{aligned}
&(c^n p^n, v^n) - \left(c^{n-1}p^{n-1}, v^{n-1,+}\right) + \int_{J^n} (Rp, v)\ dt \\
&\quad - \int_{J^n} \sum_{K\in K_h} \left[\left(\mathbf{u}\cdot\boldsymbol{\nu}_{\check{K}(t)}, v\right)_{\partial\check{K}(t)\setminus\Gamma_-} - \left(\mathbf{u}, \nabla v\right)_{\check{K}(t)}\right] dt \\
&\quad = \int_{J^n} \left\{(f, v) - \left(g_-, v\right)_{\Gamma_-} - \left(g_+\tilde{\mathbf{b}}\cdot\boldsymbol{\nu}, v\right)_{\Gamma_+}\right\} dt \\
&\quad + \int_{J^n} \left\{\left(\nabla\cdot\left[(\tilde{\mathbf{b}} - \mathbf{b})p\right], v\right) - \left(p\left[\tilde{\mathbf{b}} - \mathbf{b}\right]\cdot\boldsymbol{\nu}, v\right)_{\Gamma_-}\right\} dt .
\end{aligned} \tag{5.71}$$

The same argument applied to (5.70) yields

$$\begin{aligned}
\left(\mathbf{a}^{-1}\mathbf{u}, \mathbf{v}\right) + \sum_{K\in K_h} &\left[\left(p, \mathbf{v}\cdot\boldsymbol{\nu}_K\right)_{K\setminus\Gamma_-} - \left(\nabla p, \mathbf{v}\right)_K\right] \\
&= \left(g_+, \mathbf{v}\cdot\boldsymbol{\nu}\right)_{\Gamma_+} , \qquad \mathbf{v} \in \mathbf{V} .
\end{aligned} \tag{5.72}$$

We apply backward Euler time integration along characteristics to the diffusion, reaction, and source term in (5.71) to obtain

$$\begin{aligned}
&(c^n p^n, v^n) - \left(c^{n-1}p^{n-1}, v^{n-1,+}\right) + \left(\Delta t^n R^n p^n, v^n\right) \\
&\quad - \sum_{K\in K_h} \left[\left(\Delta t^n \mathbf{u}^n\cdot\boldsymbol{\nu}_K, v^n\right)_{\partial K\setminus\Gamma_-} - \left(\Delta t^n \mathbf{u}^n, \nabla v^n\right)_K\right] \\
&\quad = \left(\Delta t^n f^n, v^n\right) - \int_{J^n} \left\{\left(g_-, v\right)_{\Gamma_-} + \left(g_+\tilde{\mathbf{b}}\cdot\boldsymbol{\nu}, v\right)_{\Gamma_+}\right\} dt \\
&\quad + \int_{J^n} \left\{\left(\nabla\cdot\left[(\tilde{\mathbf{b}} - \mathbf{b})p\right], v\right) - \left(p\left[\tilde{\mathbf{b}} - \mathbf{b}\right]\cdot\boldsymbol{\nu}, v\right)_{\Gamma_-}\right\} dt .
\end{aligned} \tag{5.73}$$

For $h > 0$, let K_h be a regular partition of Ω into triangles or rectangles if $d = 2$ (respectively, into tetrahedra, rectangular parallelepipeds, or prisms if $d = 3$) such that the inner and outer diameter of each element is comparable to h in size; refer to Chap. 3 on their definition. Furthermore, each exterior edge or face has imposed on it either the inflow or outflow conditions, but not both. Let $\mathbf{V}_h \times W_h \subset \mathbf{V} \times W$ be some pair of mixed finite element spaces (cf. Chap. 3). For any $w \in W_h$, a test function $v(\mathbf{x}, t)$ associated with $w(\mathbf{x})$ can be defined as in (5.66). Now, the characteristic mixed method is defined: For $n = 1, 2, \ldots$, find $\mathbf{u}_h^n \in \mathbf{V}_h$ and $p_h^n \in W_h$ such that

$$\begin{aligned}
&(c^n p_h^n, v^n) - \left(c^{n-1} p_h^{n-1}, v^{n-1,+}\right) + (\Delta t^n R^n p_h^n, v^n) \\
&\quad - \sum_{K \in K_h} \left[(\Delta t^n \mathbf{u}_h^n \cdot \boldsymbol{\nu}_K, v^n)_{\partial K \setminus \Gamma_-} - (\Delta t^n \mathbf{u}_h^n, \nabla v^n)_K \right] \\
&\quad = (\Delta t^n f^n, v^n) - \int_{J^n} \left\{ (g_-, v)_{\Gamma_-} + \left(g_+ \tilde{\mathbf{b}} \cdot \boldsymbol{\nu}, v \right)_{\Gamma_+} \right\} dt \,, \qquad (5.74) \\
&\left(\Delta t^n (\mathbf{a}^n)^{-1} \mathbf{u}_h^n, \mathbf{v}\right) + \sum_{K \in K_h} \Big[(\Delta t^n p_h^n, \mathbf{v} \cdot \boldsymbol{\nu}_K)_{K \setminus \Gamma_-} \\
&\qquad - (\Delta t^n \nabla p_h^n, \mathbf{v})_K \Big] = \int_{J^n} \left(g_+^n, \mathbf{v} \cdot \boldsymbol{\nu} \right)_{\Gamma_+} dt \,,
\end{aligned}$$

for $\mathbf{v} \in \mathbf{V}_h$ and $w \in W_h$. System (5.74) determines $\mathbf{u}_h^n$ and p_h^n uniquely in terms of the data f, g_-, g_+, and p_0 (cf. Exercise 5.14).

Note that the space W_h contains piecewise constants. If we take $w = 1$ on each element $K \in K_h$ in the first equation of (5.74), we see that mass is conserved locally up to the error in approximating the integrals involved. As a matter of fact, this equation expresses local conservation of mass where fluid is transported along the approximate characteristics. System (5.74) is a generalization of the original characteristic mixed method introduced by Arbogast-Wheeler (1995) where $\mathbf{V}_h \times W_h$ is the lowest-order Raviart-Thomas-Nedelec mixed finite element space (cf. Sect. 3.4). In this case, W_h is the space of piecewise constants, and the next theorem holds (Arbogast-Wheeler, 1995).

Theorem 5.2. *Assume that Ω is a convex polygonal domain or has a smooth boundary Γ, and the coefficients $\mathbf{a}$, $\mathbf{b}$, c, f, and R satisfy*

$$\begin{aligned}
&\mathbf{a} \in \left(W^{1,\infty}(\Omega \times J)\right)^{d \times d}, \quad \mathbf{b} \in \left(W^{1,\infty}(\Omega \times J)\right)^d \,, \\
&\nabla \cdot \mathbf{b}, \ c \in W^{1,\infty}(\Omega \times J), \quad f \in L^1(\Omega \times J) \,, \\
&R \in L^\infty(J; W^{1,\infty}(\Omega)).
\end{aligned}$$

If the solution $p, \mathbf{u}$ to (5.73) satisfies $p, \nabla \cdot \mathbf{u} \in C^1(J; H^1(\Omega))$ and $\mathbf{u} \in \left(C^1(J; H^1(\Omega))\right)^d$, the initialization error satisfies

$$\|p_0 - p_h^0\|_{L^2(\Omega)} \le Ch \|p_0\|_{H^1(\Omega)} \,,$$

and h and Δt are sufficiently small, then

$$\max_{1\le n\le N}\|p^n-p_h^n\|_{L^2(\Omega)} + \left(\sum_{n=1}^{N}\|\mathbf{u}^n-\mathbf{u}_h^n\|^2_{\mathbf{L}^2(\Omega)}\Delta t^n\right)^{1/2} \le C(p,\mathbf{u})\,(h+\Delta t)\ , \qquad (5.75)$$

where $C(p,\mathbf{u})>0$ is independent of h and Δt:

$$C(p,\mathbf{u}) = C\Big\{\|p\|_{L^2(J;H^1(\Omega))} + \left\|\frac{dp}{d\boldsymbol{\tau}}\right\|_{L^2(\Omega\times J)} + \left\|\frac{d}{d\boldsymbol{\tau}}\nabla\cdot\mathbf{u}\right\|_{L^2(\Omega\times J)} + \left\|\frac{\partial\mathbf{u}}{\partial t}\right\|_{\mathbf{L}^2(J;\mathbf{H}^1(\Omega))} + \left\|\nabla\cdot\frac{\partial\mathbf{u}}{\partial t}\right\|_{L^2(J;H^1(\Omega))} + \|\mathbf{u}\|_{\mathbf{L}^\infty(J;\mathbf{H}^1(\Omega))} + \|\nabla\cdot\mathbf{u}\|_{L^\infty(J;H^1(\Omega))} + \|p_0\|_{H^1(\Omega)}\Big\}\ .$$

The linear system arising from (5.74) is typically a *saddle point problem*, and thus it needs special solution techniques (cf. Sect. 3.7). Also, $\mathbf{V}_h\subset\mathbf{H}(\mathrm{div},\Omega)$ means that the normal components of elements in $\mathbf{V}_h$ are continuous across interior boundaries. To relax this continuity requirement, a mixed-hybrid approach (Arnold-Brezzi, 1985) can be applied, but this would introduce an additional unknown (cf. Sect. 3.7.5). For this reason, in the next section, we will describe another characteristic finite element method, which does not require continuity.

5.5 The Eulerian-Lagrangian Mixed Discontinuous Method

We discuss the recently developed *Eulerian-Lagrangian mixed discontinuous method* for the numerical solution of (5.68). This method combines the ideas of the ELLAM and the mixed discontinuous method in Sect. 4.3. For $h>0$, let K_h be a finite element partition of Ω. Unlike in the previous sections, adjacent elements in K_h here are not required to match; a vertex of one element can lie in the interior of the edge or face of another element, for example, as in Chap. 4. Let $\mathcal{E}_h^o$ denote the set of all interior edges (respectively, faces) e of K_h, $\mathcal{E}_h^b$ be the set of the edges (respectively, faces) e on Γ, and $\mathcal{E}_h=\mathcal{E}_h^o\cup\mathcal{E}_h^b$. We tacitly assume that $\mathcal{E}_h^o\neq\emptyset$.

For $l\ge 0$, define

$$H^l(K_h)=\left\{w\in L^2(\Omega): w|_K\in H^l(K), K\in K_h\right\}\ .$$

That is, functions in $H^l(K_h)$ are piecewise smooth. With each $e\in\mathcal{E}_h$, we associate a unit normal vector $\boldsymbol{\nu}$. For $e\in\mathcal{E}_h^b$, $\boldsymbol{\nu}$ is just the outward unit

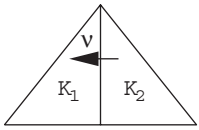

Fig. 5.15. An illustration of $\boldsymbol{\nu}$

normal to Γ. For $e \in \mathcal{E}_h^o$, with $e = \bar{K}_1 \cap \bar{K}_2$ and $K_1, K_2 \in K_h$, $\boldsymbol{\nu}$ is the unit normal exterior to K_2 with the corresponding jump definition (cf. Fig. 5.15): For $w \in H^l(K_h)$ with $l > 1/2$, we define the jump by

$$[w] = (w|_{K_2})|_e - (w|_{K_1})|_e \ .$$

For $e \in \mathcal{E}_h^o$, the average is defined by

$$\{\!\{w\}\!\} = \frac{1}{2}\big((w|_{K_1})|_e + (w|_{K_2})|_e\big) \ .$$

For $e \in \mathcal{E}_h^b$, we utilize the convention (from inside Ω)

$$\{\!\{w\}\!\} = w|_e \quad \text{and} \quad [w] = \begin{cases} w & \text{if } e \in \Gamma_+ \ , \\ 0 & \text{if } e \in \Gamma_- . \end{cases}$$

The trial and test functions in this section can be discontinuous in space. That is the reason that we utilize the averages and jumps (cf. Chap. 4).

The characteristic mixed variational form of (5.68) is the same as in (5.72) and (5.73) (Chen, 2002B). Let $\mathbf{V}_h \times W_h$ be any finite element spaces for the approximation of $\mathbf{u}$ and p, respectively. They are finite dimensional and defined locally on each element; neither continuity nor boundary data are imposed on $\mathbf{V}_h \times W_h$. For any $w \in W_h$, a test function $v(\mathbf{x}, t)$ associated with $w(\mathbf{x})$ is defined as in (5.66). The Eulerian-Lagrangian mixed discontinuous method is defined: For $n = 1, 2, \ldots$, find $\mathbf{u}_h^n \in \mathbf{V}_h$ and $p_h^n \in W_h$ such that

$$\begin{aligned}
&(c^n p_h^n, v^n) - \big(c^{n-1} p_h^{n-1}, v^{n-1,+}\big) + (\Delta t^n R^n p_h^n, v^n) \\
&\quad - \sum_{e \in \mathcal{E}_h} (\Delta t^n \{\!\{\mathbf{u}_h^n \cdot \boldsymbol{\nu}\}\!\}, [v^n])_e + \sum_{K \in K_h} (\Delta t^n \mathbf{u}_h^n, \nabla v^n)_K \\
&\quad = (\Delta t^n f^n, v^n) - \int_{J^n} \left\{ (g_-, v)_{\Gamma_-} + \Big(g_+ \tilde{\mathbf{b}} \cdot \boldsymbol{\nu}, v\Big)_{\Gamma_+} \right\} dt, \qquad (5.76) \\
&\big(\Delta t^n (\mathbf{a}^n)^{-1} \mathbf{u}_h^n, \mathbf{v}\big) + \sum_{e \in \mathcal{E}_h} (\Delta t^n [p_h^n], \{\!\{\mathbf{v} \cdot \boldsymbol{\nu}\}\!\})_e \\
&\qquad\qquad - \sum_{K \in K_h} (\Delta t^n \nabla p_h^n, \mathbf{v})_K = \int_{J^n} \big(g_+^n, \mathbf{v} \cdot \boldsymbol{\nu}\big)_{\Gamma_+} dt \ ,
\end{aligned}$$

for $\mathbf{v} \in \mathbf{V}_h$ and $w \in W_h$.

It can be checked that (5.76) has a unique solution (cf. Exercise 5.15). The next theorem (Chen, 2002B) yields a convergence result for (5.76) in the case where $\mathbf{V}_h$ and W_h are defined by

$$\mathbf{V}_h|_K = (P_r(K))^d\,, \quad W_h|_K = P_r(K), \quad r \geq 0\,, \tag{5.77}$$

where $P_r(T)$ is the set of polynomials of degree at most r on K. Note that in the Eulerian-Lagrangian mixed discontinuous method the set $P_r(K)$ can be even used on rectangles (respectively, on rectangular parallelepipeds or prisms).

Theorem 5.3. *Assume that Ω is a convex polygonal domain or has a smooth boundary Γ, and the coefficients $\mathbf{a}$, $\mathbf{b}$, c, f, and R satisfy*

$$\begin{aligned}
&\mathbf{a} \in \left(W^{1,\infty}(\Omega \times J)\right)^{d\times d}, \quad \mathbf{b} \in \left(W^{1,\infty}(\Omega \times J)\right)^d\,,\\
&\nabla \cdot \mathbf{b},\ c \in W^{1,\infty}(\Omega \times J), \quad f \in L^1(\Omega \times J)\,,\\
&R \in L^\infty(J; W^{1,\infty}(\Omega)).
\end{aligned}$$

If Δt is sufficiently small and the initialization error satisfies

$$\|p_0 - p_h^0\|_{L^2(\Omega)} \leq Ch^s \|p_0\|_{H^r(\Omega)}, \quad 1 \leq s \leq r+1\,,$$

then

$$\begin{aligned}
&\max_{1\leq n\leq N} \|p^n - p_h^n\|_{L^2(\Omega)}\\
&\quad + \left(\sum_{n=1}^N \|\mathbf{u}^n - \mathbf{u}_h^n\|^2_{\mathbf{L}^2(\Omega)} \Delta t^n\right)^{1/2} \leq C(p,\mathbf{u})\,(h^r + \Delta t)\,,
\end{aligned} \tag{5.78}$$

where

$$\begin{aligned}
C(p,\mathbf{u}) = C\Bigg\{ &\|p\|_{L^2(J;H^{r+1}(\Omega))} + \left\|\frac{dp}{d\boldsymbol{\tau}}\right\|_{L^2(J;H^r(\Omega))} + \left\|\frac{\partial p}{\partial t}\right\|_{L^2(J;H^r(\Omega))}\\
&+ \|\mathbf{u}\|_{\mathbf{L}^2(J;\mathbf{H}^{r+1}(\Omega))} + \|\mathbf{u}\|_{\mathbf{L}^2(J;\mathbf{H}(\mathrm{div},\Omega))} + \left\|\frac{d}{d\boldsymbol{\tau}}\nabla\cdot\mathbf{u}\right\|_{L^2(\Omega\times J)}\\
&+ \|p\|_{L^\infty(J;H^r(\Omega))} + \left\|\frac{d\mathbf{u}}{d\boldsymbol{\tau}}\right\|_{\mathbf{L}^2(J;\mathbf{H}^1(\Omega))} + \|p_0\|_{H^r(\Omega)}\Bigg\}\,.
\end{aligned}$$

Estimate (5.78) gives a suboptimal order of convergence in h (but optimal in Δt), but it is sharp in the general case (Chen et al., 2003B; also see Sect. 4.3). We consider the case where Ω is a rectangular domain, K_h is a Cartesian product of uniform grids in each of the coordinate directions, and

$$\mathbf{V}_h|_K = (Q_r(K))^d\,, \quad W_h|_K = Q_r(K), \quad r \geq 0\,, \tag{5.79}$$

where $Q_r(K)$ is the space of tensor products of one-dimensional polynomials of degree r on K. In this case, if r is even, it holds that (Chen, 2002B)

$$\left(\sum_{n=1}^{N} \|\mathbf{u}^n - \mathbf{u}_h^n\|^2_{\mathbf{L}^2(\Omega)} \Delta t^n\right)^{1/2} + \max_{1\le n\le N} \|p^n - p_h^n\|_{L^2(\Omega)} \le C(p,\mathbf{u})\left(h^{r+1} + \Delta t\right) , \tag{5.80}$$

where

$$C(p) = C\Bigg\{ \|p\|_{L^\infty(J;H^{r+1}(\Omega))} + \left\|\frac{dp}{d\boldsymbol{\tau}}\right\|_{L^2(J;H^{r+1}(\Omega))} + \left\|\frac{\partial p}{\partial t}\right\|_{L^2(J;H^{r+1}(\Omega))} + \|\mathbf{u}\|_{\mathbf{L}^2(J;\mathbf{H}^{r+1}(\Omega))} + \|\mathbf{u}\|_{\mathbf{L}^2(J;\mathbf{H}(\mathrm{div},\Omega))} + \left\|\frac{d}{d\boldsymbol{\tau}}\nabla\cdot\mathbf{u}\right\|_{L^2(\Omega\times J)} + \left\|\frac{d\mathbf{u}}{d\boldsymbol{\tau}}\right\|_{\mathbf{L}^2(J;\mathbf{H}^1(\Omega))} + \|p_0\|_{H^{r+1}(\Omega)} \Bigg\} .$$

Estimate (5.80) is optimal in both h and Δt. If r is odd, estimate (5.78) is sharp for (5.76), as noted.

5.6 Nonlinear Problems

We study an application of the characteristic finite element method to the nonlinear transient problem

$$\begin{aligned} &c(p)\frac{\partial p}{\partial t} + \mathbf{b}(p)\cdot\nabla p - \nabla\cdot\big(\mathbf{a}(p)\nabla p\big) = f(p) && \text{in } \Omega\times J , \\ &p(\cdot,0) = p_0 && \text{in } \Omega , \end{aligned} \tag{5.81}$$

where $c(p) = c(\mathbf{x},t,p)$, $\mathbf{b}(p) = \mathbf{b}(\mathbf{x},t,p)$, $\mathbf{a}(p) = \mathbf{a}(\mathbf{x},t,p)$, $f(p) = f(\mathbf{x},t,p)$, and $\Omega \subset \mathbb{R}^d$ ($d = 2$ or 3). This problem has been studied in the preceding four chapters. Here, as an example, we very briefly describe an application of the MMOC. Thus we assume that $\Omega \subset \mathbb{R}^d$ ($d \le 3$) is a rectangle (respectively, a rectangular parallelepiped) and (5.81) is Ω-periodic. We also assume that (5.81) admits a unique solution.

As for the linear problem (5.16), let c be a positive function and define

$$\psi(p) = \big(c^2(p) + \|\mathbf{b}(p)\|^2\big)^{1/2} .$$

The characteristic direction corresponding to the hyperbolic part of (5.81) is denoted by $\boldsymbol{\tau}$, so

$$\frac{\partial}{\partial\boldsymbol{\tau}} = \frac{c(p)}{\psi(p)}\frac{\partial}{\partial t} + \frac{1}{\psi(p)}\mathbf{b}(p)\cdot\nabla .$$

With this definition, (5.81) becomes

$$\begin{aligned} &\psi(p)\frac{\partial p}{\partial \boldsymbol{\tau}} - \nabla \cdot \big(\mathbf{a}(p)\nabla p\big) = f(p) && \text{in } \Omega \times J \ , \\ &p(\cdot, 0) = p_0 && \text{in } \Omega \ . \end{aligned} \tag{5.82}$$

Set

$$V = \{v \in H^1(\Omega) : v \text{ is } \Omega\text{-periodic}\} \ .$$

Then, using Green's formula (1.19) in space and the periodic boundary condition, problem (5.82) is recast in the variational form: Find $p : J \to V$ such that

$$\begin{aligned} &\left(\psi(p)\frac{\partial p}{\partial \boldsymbol{\tau}}, v\right) + \big(\mathbf{a}(p)\nabla p, \nabla v\big) = \big(f(p), v\big) && \forall v \in V,\ t \in J \ , \\ &p(\mathbf{x}, 0) = p_0(\mathbf{x}) && \forall \mathbf{x} \in \Omega \ . \end{aligned} \tag{5.83}$$

Because the coefficients c and $\mathbf{b}$ depend on the solution p itself, the characteristics now depend on p. This nonlinearity can be overcome by lagging one time step behind in the solution, for example. A more accurate approach is to use *extrapolations* of earlier values of the solution (cf. Sect. 1.8.1). For $n \geq 2$, take the linear extrapolation of p_h^{n-2} and p_h^{n-1} determined by

$$Ep_h^n = \left(1 + \frac{\Delta t^n}{\Delta t^{n-1}}\right) p_h^{n-1} - \frac{\Delta t^n}{\Delta t^{n-1}} p_h^{n-2} \ .$$

For $n = 1$, define

$$Ep_h^1 = p_h^0 \ .$$

Note that Ep_h^n is first-order accurate in time during the first step and second-order accurate during later steps. The characteristic derivative is now approximated by

$$\check{\mathbf{x}}_n = \mathbf{x} - \frac{\Delta t^n}{c(Ep_h^n)} \mathbf{b}(Ep_h^n). \tag{5.84}$$

With the same notation as in Sect. 5.2.3, the characteristic finite element method for (5.81) is defined: Find $p_h^n \in V_h$, $n = 1, 2, \ldots, N$, such that

$$\begin{aligned} &\left(c\left(p_h^n\right) \frac{p_h^n - \check{p}_h^{n-1}}{\Delta t^n}, v\right) + \left(\mathbf{a}\left(p_h^n\right) \nabla p_h^n, \nabla v\right) \\ &\qquad = \left(f\left(p_h^n\right), v\right) \qquad \forall v \in V_h \ , \end{aligned} \tag{5.85}$$

where

$$\check{p}_h^{n-1} = p_h\left(\check{\mathbf{x}}_n, t^{n-1}\right) \ .$$

Note that (5.85) produces a nonlinear system of algebraic equations. The solution techniques (e.g., the linearization, implicit time approximation, and explicit time approximation) discussed in Sect. 1.8 apply to it. For an analysis of method (5.85), refer to Sect. 9.5.2.

5.7 Remarks on Characteristic Finite Elements

In this chapter, we have developed the characteristic finite element method for numerically solving the reaction-diffusion-advection problem (5.1). The classical method of characteristics is a finite difference method that is based on the *forward tracking* of particles in cells or elements (Garder et al., 1984). It is known that the forward tracked characteristic method gives rise to distorted grids. The MMOC is defined in terms of a backward tracking of characteristics. It has many advantages and one fundamental flaw, the failure to preserve as an algebraic identity a desired conservation law associated with (5.1) (cf. Sect. 5.2.4). It also has an inherent difficulty in the treatment of boundary conditions. The ELLAM conserves this algebraic identity globally and can handle general boundary conditions. The characteristic mixed and Eulerian-Lagrangian mixed discontinuous methods conserve this identity locally. The ELLAM can also conserve locally if discontinuous finite elements are used (Chen, 2002B). The Eulerian-Lagrangian mixed discontinuous method relaxes a continuity requirement on normal components (across interior boundaries) of functions in the vector space in the characteristic mixed method. The MMOC is based on a nondivergence form of (5.1), while others are based on a divergence form. To see the relationships among all the existing characteristic methods, refer to Chen (2002C). Although the characteristic finite element method has been mainly developed for linear problems in this chapter, it can be also generalized to nonlinear problems where the coefficients in (5.1) depend on the solution itself (cf. Sect. 5.6); also see (Dahle et al., 1995), Douglas et al. (2000), and Chen et al. (2002, 2003C). Finally, *purely hyperbolic problems* can be directly handled by the ELLAM and Eulerian-Lagrangian mixed discontinuous method.

5.8 Theoretical Considerations

In this section, as an example, we present a theoretical analysis for the MMOC (Douglas-Russell, 1982). The reader may refer to Wang (2000), Arbogast-Wheeler (1995), and Chen (2002B) for theoretical studies, respectively, for the ELLAM, characteristic mixed method, and Eulerian-Lagrangian mixed discontinuous method. As in the preceding chapters, the reader who is not interested in the theory may skip this section.

We consider (5.6) on the whole line or (5.6) with the periodic boundary condition (5.15). Let $a(\cdot,\cdot) : W^{1,2}(\mathbb{R}) \times W^{1,2}(\mathbb{R}) \to \mathbb{R}$ be the bilinear form

$$a(v,w) = \left(a\frac{dv}{dx}, \frac{dw}{dx}\right), \qquad v,\, w \in V = W^{1,2}(\mathbb{R}) ,$$

where, for the simplicity of analysis, a is assumed to be independent of t. We recall (5.8) as

$$\left(\psi\frac{\partial p}{\partial\tau},v\right)+a(p,v)+(Rp,v)=(f,v),\quad v\in W^{1,2}(\mathbb{R}),\ t>0\ . \tag{5.86}$$

Let $V_h\subset V\cap W^{1,\infty}(\mathbb{R})$ be a finite element space such that the following approximation property holds:

$$\begin{aligned}\inf_{v_h\in V_h}\left(\|v-v_h\|_{L^2(\mathbb{R})}+h\|v-v_h\|_{W^{1,2}(\mathbb{R})}\right)\\ \le Ch^{r+1}|v|_{W^{r+1,2}(\mathbb{R})}\ ,\end{aligned} \tag{5.87}$$

where $r\ge 1$. We also recall (5.11) as

$$\left(c\frac{p_h^n-\check{p}_h^{n-1}}{\Delta t^n},v\right)+a(p_h^n,v)+(R^np_h^n,v)=(f^n,v)\qquad\forall v\in V_h\ , \tag{5.88}$$

for $n=1,2,\ldots,N$. We define the initial approximation $p_h^0\in V_h$ by

$$a(p_h^0-p_0,v)=0\qquad\forall v\in V_h\ . \tag{5.89}$$

Assume that the coefficients a, b, c, and R are bounded and satisfy

$$a_*\le a(x),\quad \left|\frac{b(x)}{c(x)}\right|+\left|\frac{d}{dx}\left(\frac{b(x)}{c(x)}\right)\right|\le C,\qquad x\in\mathbb{R}\ , \tag{5.90}$$

where a_* is a positive constant. Also, assume that the solution p of (5.6) satisfies

$$\begin{aligned}&p\in L^\infty(J;W^{r+1,2}(\mathbb{R})),\quad \frac{\partial^2 p}{\partial\tau^2}\in L^2(\mathbb{R}\times J)\ ,\\ &\frac{\partial p}{\partial t}\in L^2(J;W^{r+\zeta,2}(\mathbb{R})),\quad \zeta=1 \text{ if } r=1 \text{ and } \zeta=0 \text{ if } r>1\ .\end{aligned}$$

Let $w_h:J\to V_h$ satisfy

$$a(p-w_h,v)=0\qquad\forall v\in V_h,\quad t\in J\ . \tag{5.91}$$

Set

$$\eta=p-w_h,\quad \xi=p_h-w_h\ .$$

It follows from Sect. 1.9 that, for $q=2$ or ∞ and $1\le s\le r+1$,

$$\|\eta\|_{L^q(J;L^2(\mathbb{R}))}+h\|\eta\|_{L^q(J;W^{1,2}(\mathbb{R}))}\le Ch^s\|p\|_{L^q(J;W^{s,2}(\mathbb{R}))}\ . \tag{5.92}$$

Because the bilinear form $a(\cdot,\cdot)$ is independent of time, it follows that, for $r\ge 2$ and $1\le s\le r+1$,

$$\begin{aligned}\left\|\frac{\partial\eta}{\partial t}\right\|_{L^2(J;W^{-1,2}(\mathbb{R}))}+h\left\|\frac{\partial\eta}{\partial t}\right\|_{L^2(\mathbb{R}\times J)}\\ \le Ch^s\left\|\frac{\partial p}{\partial t}\right\|_{L^2(J;W^{s-1,2}(\mathbb{R}))}\ ;\end{aligned} \tag{5.93}$$

in the case $r = 1$, there is no gain of a factor h in the $W^{-1,2}$ estimate.

From (5.92) and (5.93), to obtain error bounds for $p - p_h$, it suffices to estimate ξ. To that end, we need the next lemma. For simplicity of exposition, let $\Delta t = \Delta t^n$, $n = 1, 2, \ldots, N$.

Lemma 5.4. *If $\eta \in L^2(\mathbb{R})$ and $\check{\eta} = \eta(x - \varphi(x)\Delta t)$, where φ and $d\varphi/dx$ are bounded, then*

$$\|\eta - \check{\eta}\|_{W^{-1,2}(\mathbb{R})} \le C\Delta t \|\eta\|_{L^2(\mathbb{R})} .$$

Proof. Set $z = F(x) = x - \varphi(x)\Delta t$. Then it is easy to see that F is invertible if Δt is sufficiently small, and that $d\varphi/dx$ and $d\varphi^{-1}/dx$ are both of order $1 + \mathcal{O}(\Delta t)$. Thus we see that

$$\begin{aligned}
&\|\eta - \check{\eta}\|_{W^{-1,2}(\mathbb{R})} \\
&= \sup_{v \in W^{1,2}(\mathbb{R})} \left(\|v\|^{-1}_{W^{1,2}(\mathbb{R})} \int_{\mathbb{R}} [\eta(x) - \eta(z)]\, v(x)\, dx \right) \\
&= \sup_{v \in W^{1,2}(\mathbb{R})} \left(\|v\|^{-1}_{W^{1,2}(\mathbb{R})} \left[\int_{\mathbb{R}} \eta(x) v(x)\, dx \right.\right. \\
&\qquad \left.\left. - \int_{\mathbb{R}} \eta(z) v\left(F^{-1}(z)\right) \left(1 + \mathcal{O}(\Delta t)\right)\, dz \right] \right) \\
&\le \sup_{v \in W^{1,2}(\mathbb{R})} \left(\|v\|^{-1}_{W^{1,2}(\mathbb{R})} \int_{\mathbb{R}} \eta(x) \left[v(x) - v\left(F^{-1}(x)\right) \right]\, dx \right) \\
&\quad + C\Delta t \sup_{v \in W^{1,2}(\mathbb{R})} \left(\|v\|^{-1}_{W^{1,2}(\mathbb{R})} \int_{\mathbb{R}} \eta(x) v\left(F^{-1}(x)\right)\, dx \right) .
\end{aligned} \tag{5.94}$$

Let $G(x) = x - F^{-1}(x)$; then $|G(x)| \le C\Delta t$, and

$$\begin{aligned}
\left\| v(x) - v\left(F^{-1}(x)\right) \right\|_{L^2(\mathbb{R})} &\le \int_{\mathbb{R}} \left(\int_{F^{-1}(x)}^{x} \left| \frac{dv}{dx} \right|\, dy \right)^2 dx \\
&\le C\left(\Delta t\right)^2 \int_{\mathbb{R}} \int_0^1 \left| \frac{dv}{dx}(x - G(x)y) \right|^2 dy\, dx \\
&\le C\left(\Delta t\right)^2 \|v\|^2_{W^{1,2}(\mathbb{R})} ,
\end{aligned} \tag{5.95}$$

where the last step uses the change of variables $\tilde{x} = x - G(x)y$, which induces a factor of $1 + \mathcal{O}(\Delta t)$. A similar change of variables yields

$$\left\| v \circ F^{-1} \right\|^2 = (1 + \gamma C \Delta t) \|v\|^2_{L^2(\mathbb{R})}, \quad |\gamma| \le 1 , \tag{5.96}$$

where C is the constant in (5.90). The same result is true for $v \circ F$. Combining (5.94)–(5.96) implies the desired result. □

We also need the following discrete Gronwall lemma:

Lemma 5.5. *Assume that $B(n)$, $D(n)$, and $w(n)$ are three sequences of real nonnegative numbers such that*

$$B(n) \le D(n) + \sum_{k=0}^{n-1} w(k)B(k), \quad n = 1, 2, \dots .$$

Furthermore, assume that $D(n)$ is nondecreasing. Then

$$B(n) \le D(n)\exp\left(\sum_{\mathrm{k=0}}^{\mathrm{n-1}} \mathrm{w(k)}\right) .$$

Proof. Let $v(m) = D(n) + \sum_{k=0}^{m-1} w(k)B(k)$ for $m \le n$. Then

$$\begin{aligned} v(m) &= v(m-1) + w(m-1)B(m-1) \\ &\le (1 + w(m-1))\, v(m-1) \le e^{w(m-1)} v(m-1) . \end{aligned}$$

Because $v(0) = D(n)$, the desired result follows. □

Remark 5.6. In the special case where $w(n) = C\Delta t$ for $n \ge 0$ and $T = n\Delta t$, with C and T fixed, it holds that

$$B(n) \le D(n)e^{CT} .$$

Theorem 5.7. *Let p and p_h be the respective solutions of* (5.86) *and* (5.88). *Then, for Δt sufficiently small, we have*

$$\begin{aligned} \max_{1\le n\le N} \|p^n - p_h^n\|_{L^2(\mathbb{R})} \le C\Bigg\{ &\Delta t \left\| \frac{\partial^2 p}{\partial \tau^2} \right\|_{L^2(\mathbb{R}\times J)} \\ &+ h^{r+1}\left(\|p\|_{L^\infty(J;W^{r+1,2}(\mathbb{R}))} + \left\| \frac{\partial p}{\partial t} \right\|_{L^2(J;W^{r+\zeta,2}(\mathbb{R}))} \right) \Bigg\} , \end{aligned}$$

where $\zeta = 1$ if $r = 1$ and $\zeta = 0$ if $r > 1$.

Proof. Subtract (5.86) from (5.88) to give

$$\begin{aligned} &\left(c\frac{\xi^n - \check{\xi}^{n-1}}{\Delta t}, v \right) + a(\xi^n, v) \\ &\quad = \left(\psi \frac{\partial p^n}{\partial \tau} - c\frac{p^n - \check{p}_h^{n-1}}{\Delta t}, v \right) + \left(c\frac{\eta^n - \check{\eta}^{n-1}}{\Delta t}, v \right) \\ &\qquad + (R^n(p^n - p_h^n), v) \qquad \forall v \in V_h . \end{aligned} \tag{5.97}$$

We take $v = \xi^n$ in (5.97) to give

$$\begin{aligned}
&\left(c\frac{\xi^n - \check{\xi}^{n-1}}{\Delta t}, \xi^n\right) + a(\xi^n, \xi^n) \\
&\quad = \left(\psi\frac{\partial p^n}{\partial \tau} - c\frac{p^n - \check{p}_h^{n-1}}{\Delta t}, \xi^n\right) + \left(c\frac{\eta^n - \check{\eta}^{n-1}}{\Delta t}, \xi^n\right) \\
&\qquad + (R^n(p^n - p_h^n), \xi^n) \ .
\end{aligned} \tag{5.98}$$

Denote by $\big(x(\tau), t(\tau)\big)$ the coordinates of the point on the segment of the tangent to the characteristic from $(\check{x}, t^{n-1})$ to (x, t^n). Then the backward difference quotient error along this tangent to the characteristic is given by

$$\begin{aligned}
&\psi\frac{\partial p^n}{\partial \tau} - c\frac{p^n - \check{p}_h^{n-1}}{\Delta t} \\
&\quad = \frac{c}{\Delta t}\int_{(\check{x}, t^{n-1})}^{(x,t^n)} \big((x(\tau) - \check{x})^2 + (t(\tau) - t^{n-1})^2\big)^{1/2}\,\frac{\partial^2 p}{\partial \tau^2}\ d\tau \ .
\end{aligned}$$

Taking the $L^2(\mathbb{R})$-norm of this error, we see that

$$\begin{aligned}
&\left\|\psi\frac{\partial p^n}{\partial \tau} - c\frac{p^n - \check{p}_h^{n-1}}{\Delta t}\right\|_{L^2(\mathbb{R})}^2 \\
&\quad \le \int_{\mathbb{R}} \left(\frac{c}{\Delta t}\right)^2 \left(\frac{\psi}{c}\Delta t\right)^2 \left|\int_{(\check{x}, t^{n-1})}^{(x,t^n)} \frac{\partial^2 p}{\partial \tau^2}\ d\tau\right|^2 dx \\
&\quad \le \Delta t \left\|\frac{\psi^3}{c}\right\|_{L^\infty(\mathbb{R})} \int_{\mathbb{R}} \int_{(\check{x}, t^{n-1})}^{(x,t^n)} \left|\frac{\partial^2 p}{\partial \tau^2}\right|^2 d\tau\ dx \\
&\quad \le \Delta t \left\|\frac{\psi^4}{c^2}\right\|_{L^\infty(\mathbb{R})} \int_{\mathbb{R}} \int_{J^n} \left|\frac{\partial^2 p}{\partial \tau^2}\left(\frac{t^n - t}{\Delta t}\check{x} + \frac{t - t^{n-1}}{\Delta t}x, t\right)\right|^2 dt\ dx \ .
\end{aligned}$$

To relate this to a standard norm of $\partial^2 p/\partial \tau^2$, we introduce the transformation

$$\begin{aligned}
\mathcal{F} : (x, t) \to (z, t) &= \left(\frac{t^n - t}{\Delta t}\check{x} + \frac{t - t^{n-1}}{\Delta t}x, t\right) \\
&= (\theta(t)\check{x} + (1 - \theta(t))x, t) \ .
\end{aligned}$$

The Jacobian of this map is given by

$$J(\mathcal{F}) = \begin{pmatrix} 1 - \theta(t)\Delta t\dfrac{d}{dx}\left(\dfrac{b}{c}\right) & \dfrac{b(x)}{c(x)} \\ 0 & 1 \end{pmatrix} .$$

Using (5.90), $\mathcal{F}$ is invertible for Δt small enough and the determinant of J is $1 + \mathcal{O}(\Delta t)$. For any fixed t, $\mathcal{F}$ obviously maps $\mathbb{R} \times \{t\}$ onto itself, so the same is true for $\mathbb{R} \times J^n$. Thus it follows that

$$\left\| \psi \frac{\partial p^n}{\partial \tau} - c \frac{p^n - \check{p}_h^{n-1}}{\Delta t} \right\|^2_{L^2(\mathbb{R})} \leq 2\Delta t \left\| \frac{\psi^4}{c^2} \right\|_{L^\infty(\mathbb{R})} \left\| \frac{\partial^2 p}{\partial \tau^2} \right\|^2_{L^2(\mathbb{R} \times J^n)} ,$$

and the first term on the right-hand side of (5.98) is bounded by

$$C \left(\Delta t \left\| \frac{\partial^2 p}{\partial \tau^2} \right\|^2_{L^2(\mathbb{R} \times J^n)} + \|\xi^n\|^2_{L^2(\mathbb{R})} \right). \tag{5.99}$$

Next, we write $\eta^n - \check{\eta}^{n-1}$ as the sum of $(\eta^n - \eta^{n-1}) + (\eta^{n-1} - \check{\eta}^{n-1})$. Then we see that

$$\begin{aligned} \left| \left(c \frac{\eta^n - \eta^{n-1}}{\Delta t}, \xi^n \right) \right| &\leq \frac{C}{\Delta t} \|\xi^n\|_{W^{1,2}(\mathbb{R})} \int_{J^n} \left\| \frac{\partial \eta}{\partial t} \right\|_{W^{-1,2}(\mathbb{R})} dt \\ &\leq \epsilon \|\xi^n\|^2_{W^{1,2}(\mathbb{R})} + \frac{C}{\Delta t} \left\| \frac{\partial \eta}{\partial t} \right\|^2_{L^2(J^n; W^{-1,2}(\mathbb{R}))} , \end{aligned} \tag{5.100}$$

where ϵ is a positive constant, as small as we please. Also, by Lemma 5.4, we have

$$\begin{aligned} &\left| \left(c \frac{\eta^{n-1} - \check{\eta}^{n-1}}{\Delta t}, \xi^n \right) \right| \\ &\quad \leq C \|\xi^n\|_{W^{1,2}(\mathbb{R})} \left\| \frac{\eta^{n-1} - \check{\eta}^{n-1}}{\Delta t} \right\|_{W^{-1,2}(\mathbb{R})} \\ &\quad \leq \epsilon \|\xi^n\|^2_{W^{1,2}(\mathbb{R})} + C \|\eta^{n-1}\|^2_{L^2(\mathbb{R})} . \end{aligned} \tag{5.101}$$

It is obvious that

$$|(R^n(p^n - p_h^n), \xi^n)| \leq C \left(\|\xi^n\|^2_{L^2(\mathbb{R})} + \|\eta^n\|^2_{L^2(\mathbb{R})} \right). \tag{5.102}$$

This completes the treatment of the right-hand side of (5.98).

The left-hand side is bounded below:

$$\begin{aligned} &\left(c \frac{\xi^n - \check{\xi}^{n-1}}{\Delta t}, \xi^n \right) + a(\xi^n, \xi^n) \\ &\quad \geq \frac{1}{2\Delta t} \left[(c\xi^n, \xi^n) - (c\check{\xi}^{n-1}, \check{\xi}^{n-1}) \right] + a(\xi^n, \xi^n) \\ &\quad = \frac{1}{2\Delta t} \left[(c\xi^n, \xi^n) - (c\xi^{n-1}, \xi^{n-1})(1 + \gamma^n C \Delta t) \right] \\ &\qquad + a(\xi^n, \xi^n), \quad |\gamma^n| \leq 1 , \end{aligned} \tag{5.103}$$

where (5.96) has been used. Inequalities (5.95)–(5.103) can be combined with (5.98) to give the recursion relation

$$\begin{aligned}
&\frac{1}{2\Delta t}\left[(c\xi^n,\xi^n)-(c\xi^{n-1},\xi^{n-1})\right]+\frac{a_0}{2}\|\xi^n\|^2_{W^{1,2}(\mathbb{R})}\\
&\quad\le C\Bigg\{\|\xi^n\|^2_{L^2(\mathbb{R})}+\|\xi^{n-1}\|^2_{L^2(\mathbb{R})}+\Delta t\left\|\frac{\partial^2 p}{\partial\tau^2}\right\|^2_{L^2(\mathbb{R}\times J^n)}\\
&+\frac{1}{\Delta t}\left\|\frac{\partial\eta}{\partial t}\right\|^2_{L^2(J^n;W^{-1,2}(\mathbb{R}))}+\|\eta^{n-1}\|^2_{L^2(\mathbb{R})}+\|\eta^n\|^2_{L^2(\mathbb{R})}\Bigg\}.
\end{aligned}\tag{5.104}$$

It follows from (5.89) and (5.91) that $\xi^0=0$. If we multiply (5.104) by $2\Delta t$, sum over n, and apply Lemma 5.5 and Remark 5.6, it follows that

$$\begin{aligned}
\max_{1\le n\le N}\|\xi^n\|_{L^2(\mathbb{R})}+\left(\sum_{n=1}^N\|\xi^n\|^2_{W^{1,2}(\mathbb{R})}\Delta t\right)^{1/2}&\le C\Bigg\{\Delta t\left\|\frac{\partial^2 p}{\partial\tau^2}\right\|_{L^2(\mathbb{R}\times J)}\\
&+\left\|\frac{\partial\eta}{\partial t}\right\|_{L^2(J;W^{-1,2}(\mathbb{R}))}+\|\eta\|_{L^\infty(J;L^2(\mathbb{R}))}\Bigg\},
\end{aligned}$$

which, together with (5.92) and (5.93), yields the desired result. □

Theorem 5.8. *Let p and p_h be the respective solutions of* (5.86) *and* (5.88). *Then, for Δt sufficiently small, we have*

$$\begin{aligned}
\max_{1\le n\le N}\|p^n-p_h^n\|_{W^{1,2}(\mathbb{R})}&\le C\Bigg\{\Delta t\left\|\frac{\partial^2 p}{\partial\tau^2}\right\|_{L^2(\mathbb{R}\times J)}\\
&+h^r\left(\|p\|_{L^\infty(J;W^{r+1,2}(\mathbb{R}))}+\left\|\frac{\partial p}{\partial t}\right\|_{L^2(J;W^{r,2}(\mathbb{R}))}\right)\Bigg\}.
\end{aligned}$$

Proof. Taking $v=(\xi^n-\xi^{n-1})/\Delta t$ in (5.97), we see that

$$\begin{aligned}
&\left(c\frac{\xi^n-\check{\xi}^{n-1}}{\Delta t},\frac{\xi^n-\xi^{n-1}}{\Delta t}\right)+a\left(\xi^n,\frac{\xi^n-\xi^{n-1}}{\Delta t}\right)\\
&\quad=\left(\psi\frac{\partial p^n}{\partial\tau}-c\frac{p^n-\check{p}_h^{n-1}}{\Delta t},\frac{\xi^n-\xi^{n-1}}{\Delta t}\right)\\
&\qquad+\left(c\frac{\eta^n-\check{\eta}^{n-1}}{\Delta t},\frac{\xi^n-\xi^{n-1}}{\Delta t}\right)\\
&\qquad+\left(R^n(p^n-p_h^n),\frac{\xi^n-\xi^{n-1}}{\Delta t}\right).
\end{aligned}\tag{5.105}$$

Because

$$\left|\left(c\frac{\xi^{n-1}-\check{\xi}^{n-1}}{\Delta t},\frac{\xi^n-\xi^{n-1}}{\Delta t}\right)\right|\le C\left\|\frac{\partial\xi^{n-1}}{\partial x}\right\|_{L^2(\mathbb{R})}\left\|\frac{\xi^n-\xi^{n-1}}{\Delta t}\right\|_{L^2(\mathbb{R})},$$

the left-hand side of (5.105) is bounded below:

$$\begin{aligned}&\left(c\frac{\xi^n-\check{\xi}^{n-1}}{\Delta t},\frac{\xi^n-\xi^{n-1}}{\Delta t}\right)+a\left(\xi^n,\frac{\xi^n-\xi^{n-1}}{\Delta t}\right)\\&\quad\geq\left(c\frac{\xi^n-\xi^{n-1}}{\Delta t},\frac{\xi^n-\xi^{n-1}}{\Delta t}\right)-\epsilon\left\|\frac{\xi^n-\xi^{n-1}}{\Delta t}\right\|^2_{L^2(\mathbb{R})}\\&+\frac{1}{2\Delta t}\left[a(\xi^n,\xi^n)-a(\xi^{n-1},\xi^{n-1})\right]-C\|\xi^{n-1}\|^2_{W^{1,2}(\mathbb{R})}\,.\end{aligned}\tag{5.106}$$

The right-hand side can be bounded as follows. First, we see that

$$\begin{aligned}&\left|\left(\psi\frac{\partial p^n}{\partial\tau}-c\frac{p^n-\check{p}^{n-1}_h}{\Delta t},\frac{\xi^n-\xi^{n-1}}{\Delta t}\right)\right|\\&\quad\leq\epsilon\left\|\frac{\xi^n-\xi^{n-1}}{\Delta t}\right\|^2_{L^2(\mathbb{R})}+C\Delta t\left\|\frac{\partial^2 p}{\partial\tau^2}\right\|_{L^2(\mathbb{R}\times J^n)}\,.\end{aligned}\tag{5.107}$$

Second, we observe that

$$\begin{aligned}&\left|\left(c\frac{\eta^n-\check{\eta}^{n-1}}{\Delta t},\frac{\xi^n-\xi^{n-1}}{\Delta t}\right)\right|\leq\epsilon\left\|\frac{\xi^n-\xi^{n-1}}{\Delta t}\right\|^2_{L^2(\mathbb{R})}\\&\quad+C\left(\|\eta^{n-1}\|^2_{W^{1,2}(\mathbb{R})}+(\Delta t)^{-1}\left\|\frac{\partial\eta}{\partial t}\right\|^2_{L^2(\mathbb{R}\times J^n)}\right)\,.\end{aligned}\tag{5.108}$$

Establishing (5.108) does not use Lemma 5.4. Third, we have

$$\begin{aligned}&\left|\left(R^n(p^n-p^n_h),\frac{\xi^n-\xi^{n-1}}{\Delta t}\right)\right|\leq\epsilon\left\|\frac{\xi^n-\xi^{n-1}}{\Delta t}\right\|^2_{L^2(\mathbb{R})}\\&\quad+C\left(\|\xi^n\|^2_{L^2(\mathbb{R})}+\|\eta^n\|^2_{L^2(\mathbb{R})}\right)\,.\end{aligned}\tag{5.109}$$

Finally, we combine (5.105)–(5.109) to obtain

$$\begin{aligned}\max_{1\leq n\leq N}\|\xi^n\|_{W^{1,2}(\mathbb{R})}&\leq C\Bigg\{\Delta t\left\|\frac{\partial^2 p}{\partial\tau^2}\right\|_{L^2(\mathbb{R}\times J)}\\&+\|\xi\|_{L^\infty(J;L^2(\mathbb{R}))}+\|\eta\|_{L^\infty(J;W^{1,2}(\mathbb{R}))}+\left\|\frac{\partial\eta}{\partial t}\right\|_{L^2(\mathbb{R}\times J)}\Bigg\}\,,\end{aligned}$$

which, together with (5.92), (5.93), and Theorem 5.7, implies the desired result. □

Note that Theorems 5.7 and 5.8 imply estimate (5.14). As noted in Sect. 5.2.1, the term $\|\partial^2 p/\partial\tau^2\|_{L^2(\mathbb{R}\times J)}$ appears in the error estimates in these two theorems, instead of the term $\|\partial^2 p/\partial t^2\|_{L^2(\mathbb{R}\times J)}$. The former is

much smaller than the later for an advection-dominated problem. Also, note that we have studied the one-dimensional problem (5.6). A similar analysis can be carried out for its multi-dimensional counterpart (5.16) (Douglas-Russell, 1982; Ewing et al., 1984; Dawson et al., 1989; Chen et al., 2003C).

5.9 Bibliographical Remarks

The original definition of the MMOC, ELLAM, characteristic mixed method, and Eulerian-Lagrangian mixed discontinuous method presented in Sects. 5.2–5.5 can be found in Douglas-Russell (1982), Celia et al. (1990), Arbogast-Wheeler (1995), and Chen (2002B), respectively. The definition of the ELLAM in one dimension given in Sect. 5.3.1 follows Russell (1990). Finally, the content of Sect. 5.8 is chosen from Douglas-Russell (1982).

5.10 Exercises

5.1. Show that after multiplying both sides of (5.11) by Δt^n, the condition number of the stiffness matrix corresponding to the left-hand side of (5.11) is of order (cf. Sect. 1.10)

$$\mathcal{O}\left(1 + \max_{x\in\mathbb{R},\ t\geq 0} |a(x,t)|h^{-2}\Delta t\right), \quad \Delta t = \max_{n=1,2,\ldots} \Delta t^n \ .$$

5.2. Let $v \in C^1(\mathbb{R})$ be a $(0,1)$-periodic function. Show that the condition $v(0) = v(1)$ implies

$$\frac{\partial v(0)}{\partial x} = \frac{\partial v(1)}{\partial x} \ .$$

5.3. Let $\mathbf{a}$ be positive semi-definite, c be uniformly positive with respect to x and t, and R be nonnegative. Show that (5.21) has a unique solution $p_h^n \in V_h$ for each n.

5.4. Prove relation (5.27).

5.5. Equation (5.36) is written for the first type of element at the left boundary in Fig. 5.7. Write down the equation corresponding to the second, third, and fourth type, respectively, in Fig. 5.7.

5.6. Equation (5.44) is written for the first type of element at the right boundary in Fig. 5.8. Write down the equation corresponding to the second type in Fig. 5.8.

5.7. In (5.51), a linear test function, which is constant along characteristics, is defined. Define a quadratic test function which satisfies the same property.

5.8. The ELLAM procedure (5.54) is established with the boundaries of the flux type at both $x = 0$ and $x = 1$. Develop an ELLAM procedure with the left boundary of the flux type and the right boundary of the Dirichlet type.

5.9. Let the coefficients a and c be uniformly positive with respect to x and t and R be nonnegative. Show that (5.54) possesses a unique solution $p_h^n \in V_h$ for each n.

5.10. Derive an error estimate (similar to (5.14)) for (5.54) with the linear test functions. (If necessary, refer to Wang (2000).)

5.11. Let $\mathbf{a}$ be positive semi-definite, c be uniformly positive with respect to x and t, and R be nonnegative. Show that (5.67) has a unique solution $p_h^n \in V_h$ for each n.

5.12. The ELLAM procedure (5.67) is defined for the flux boundary condition in (5.56). Extend (5.67) to a Dirichlet boundary condition for (5.56).

5.13. Derive (5.69) from the first equation and the boundary conditions in (5.68).

5.14. Let $\mathbf{a}$ be positive-definite, c be uniformly positive with respect to x and t, and R be nonnegative. Show that (5.74) has a unique solution $\mathbf{u}_h^n \in \mathbf{V}_h$ and $p_h^n \in W_h$ for each n.

5.15. Let $\mathbf{a}$ be positive-definite, c be uniformly positive with respect to x and t, and R be nonnegative. Show that (5.76) has a unique solution $\mathbf{u}_h^n \in \mathbf{V}_h$ and $p_h^n \in W_h$ for each n.

5.16. Extend the error analysis in Sect. 5.8 to a multi-dimensional ($d = 2$ or 3) case.

5.17. Extend the error analysis in Sect. 5.8 for the linear problem (5.6) to method (5.85) for the nonlinear problem (5.81).

6 Adaptive Finite Elements

In real applications, many important physical and chemical phenomena are sufficiently localized and transient that *adaptive numerical methods* are necessary to resolve them. Adaptive numerical methods have become increasingly important because researchers have realized the great potential of the concepts underlying these methods. They are numerical schemes that automatically adjust themselves to improve approximate solutions. These methods are not exactly new in the computational area, even in the finite element literature. The adaptive adjustment of time steps in the numerical solution of ordinary differential equations, particularly non-stiff equations, has been the subject of research for many decades. Furthermore, the search for optimal finite element grids dates back to the early 1970's (Oliveira, 1971). But modern interest in this subject began in the late 1970's, mainly thanks to important contributions by Babuška-Rheinboldt (1978A,B) and many others.

In the numerical solution of practical problems in engineering and physics such as in solid and fluid mechanics (cf. Chaps. 7 and 8) and in porous media flow and semiconductor device simulation (cf. Chaps. 9 and 10), the overall accuracy of numerical approximations often deteriorates due to local singularities like those arising from re-entrant corners of domains, interior or boundary layers, and sharp moving fronts. An obvious strategy is to refine the grids near these critical regions, i.e., to insert more grid points where the singularities occur. The question is then how we identify those regions, refine them, and obtain a good balance between the refined and unrefined regions such that the overall accuracy is optimal. To answer this question, we need to utilize adaptivity. That is, we need somehow to *restructure a numerical scheme* to improve the quality of its approximate solutions. This puts a great demand on the choice of numerical methods. Restructuring a numerical scheme includes changing the number of elements, refining local grids, increasing the local order of approximation, moving nodal points, and modifying algorithm structures.

Another closely related question is how to obtain reliable estimates of the accuracy of computed approximate solutions. A-priori error estimates, as obtained in the preceding five chapters, are often insufficient because they produce information only on the asymptotic behavior of errors and they require a solution regularity that is not satisfied in the presence of the above

mentioned singularities. To answer this question, we need to assess the quality of approximate solutions a-posteriori, i.e., after an initial approximation is obtained. This requires that we compute a-posteriori error estimates. Of course, the computation of the a-posteriori estimates should be far less expensive than that of the approximate solutions. Moreover, it must be possible to compute dynamically local error indicators that lead to some estimate of the local quality of the solution.

The aim of this chapter is to present a brief introduction of some of basic topics on the two components for the adaptive finite element method: the *adaptive strategy* and a-posteriori *error estimation*. We focus on these two components for the standard finite element method considered in Chap. 1. Research of how to combine them with other methods in Chaps. 2–5 is mentioned at the end of this chapter (cf. Sect. 6.7). In Sect. 6.1, we introduce the concept of local grid refinement in space. For large-scale problems, the choice of data structures that permit efficient and accurate solution is important. In Sect. 6.2, we briefly discuss a data structure that efficiently supports adaptive refinement and unrefinement. In Sect. 6.3, we discuss a-posteriori error estimates for stationary problems, and, in Sect. 6.4, extend them to transient problems. In Sect. 6.5, we briefly consider their application to nonlinear problems. In Sect. 6.6, we present theoretical considerations. Finally, in Sect. 6.7, we make a few remarks on adaptive finite elements.

6.1 Local Grid Refinement in Space

There are three basic types of adaptive strategies: (1) local refinement of a fixed grid, (2) addition of more degrees of freedom locally by utilizing higher-order basis functions in certain elements, and (3) adaptively moving a computational grid to achieve better local resolution.

Local grid refinement of a fixed grid is called an *h-scheme*. In this scheme, the mesh is automatically refined or unrefined depending upon a local error indicator. Such a scheme leads to a very complex data management problem because it involves the dynamic regeneration of a grid, renumbering of nodal points and elements, and element connectivity. However, the h-scheme can be very effective in generating near-optimal grids for a given error tolerance. Efficient h-schemes with fast data management procedures have been developed for complex problems (Diaz et al., 1984; Ewing, 1986; Bank, 1990). Moreover, the h-scheme can be also employed to *unrefine* a grid (or *coarsen* a grid) when a local error indicator becomes smaller than a preassigned tolerance.

Addition of more degrees of freedom locally by utilizing higher-order basis functions in certain elements is referred to as a *p-scheme* (Babuška et al., 1983; Szabo, 1986). As discussed in Chap. 1, the finite element method for a given problem attempts to approximate a solution by functions in a finite-dimensional space of polynomials. The p-scheme generally utilizes a fixed

grid and a fixed number of grid elements. If the error indicator in any element exceeds a given tolerance, the local order of the polynomial degree is increased to reduce the error. This scheme can be very effective in modeling thin boundary layers around bodies moving in a flow field, where the use of very fine grids is impractical and costly. However, the data management problem associated with the p-scheme, especially for regions of complex geometry, can be very difficult.

Adaptively moving a computational grid to achieve better local resolution is usually termed a *r-scheme* (Miller-Miller, 1981). It employs a fixed number of grid points and attempts to move them dynamically to areas where the error indicator exceeds a preassigned tolerance. The r-scheme can be easily implemented, and does not have the difficult data management problem associated with the h- and p-schemes. On the other hand, it suffers from several deficiencies. Without special care in its implementation, it can be unstable and result in grid tangling and local degradation of approximate solutions. It can never reduce the error below a fixed limit since it is not capable of handling the migration of regions where the solution is singular. However, by an appropriate combination with other adaptive strategies, the r-scheme can lead to a useful scheme for controlling solution errors.

Combinations of these three basic strategies such as the *hr-, hp-*, and *hpr-schemes* are also possible (Babuška-Dorr, 1981; Oden et al., 1989). In this chapter, as an example, we study the widely applied h-scheme.

6.1.1 Regular H-Schemes

We focus on a two-dimensional domain. An extension of the concept in this section to three dimensions is simple to visualize. However, the modification of the supporting algorithms in the next section is not straightforward.

In the two-dimensional case, a grid can be triangular, quadrilateral, or of mixed-type (i.e., consisting of both triangles and quadrilaterals); see Chap. 1. A vertex is *regular* if it is a vertex of each of its neighboring elements, and a grid is *regular* if its every vertex is regular. All other vertices are said to be *irregular* (*slave nodes* or *hanging nodes*); see Fig. 6.1. The *irregularity index* of a grid is the maximum number of irregular vertices belonging to the same edge of an element.

If all elements in a grid are subdivided into an equal number (usually four) of smaller elements simultaneously, the refinement is referred to as *global*. For example, a refinement is global by connecting the opposite midpoints of the edges of each triangle or quadrilateral in the grid. Global refinement does not introduce irregular vertices. In the preceding five chapters, all the refinements were global and regular. In contrast, in the case of a *local refinement* where only some of the elements in a grid are subdivided into smaller elements, irregular vertices may appear; refer to Fig. 6.1.

In this subsection, we study a regular local refinement. The following *refinement rule* can be used to convert irregular vertices to regular ones (Bank,

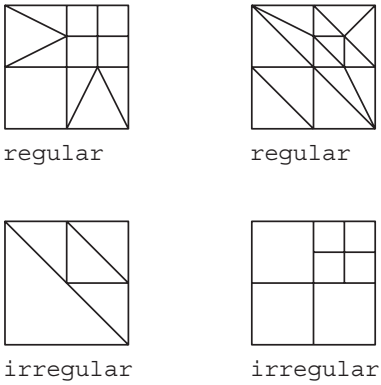

Fig. 6.1. Examples of regular and irregular vertices

1990; Braess, 1997). This rule is designed for a triangular grid and guarantees that each of the angles in the original grid is bisected at most once. We may think of starting with a triangulation as in Fig. 6.2. It contains several irregular vertices, which need to be converted to regular vertices.

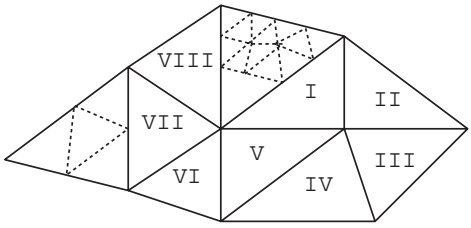

Fig. 6.2. A coarse grid (*solid lines*) and a refinement (*dotted lines*)

A refinement rule for a triangulation is defined as follows:

1. If an edge of a triangle contains two or more vertices of other triangles (not counting its own vertices), then this triangle is subdivided into four equal smaller triangles. This procedure is repeated until such triangles no longer exist.
2. If the midpoint of an edge of a triangle contains a vertex of another triangle, this triangle is subdivided into two parts. The new edge is called a *green edge*.
3. If a further refinement is needed, the green edges are first eliminated before the next iteration.

For the triangulation in Fig. 6.2, we apply the first step to triangles I and VIII. This requires the use of the refinement rule twice on triangle VII. Next,

we construct green edges on triangles II, V, and VI and on three subtriangles (cf. Exercise 6.1).

Despite its recursive nature, this procedure stops after a finite number of iterations. Let k be the maximum number of levels in the underlying refinement, where the maximum is taken over all elements ($k = 2$ in Fig. 6.2). Then every element is subdivided at most k times, which presents an upper bound on the number of times step 1 is to be used. We emphasize that this procedure is purely two-dimensional. A generalization to three dimensions is not straightforward. For a triangulation of Ω into tetrahedra, see a technique due to Rivara (1984A).

6.1.2 Irregular H-Schemes

Irregular grids leave more freedom for local refinement. In the general case of arbitrary irregular grids, an element may be refined locally without any interference with its neighbors. As for regular local refinement, some desirable properties should be preserved for irregular refinement as well.

First, in the process of consecutive refinements no distorted elements should be generated. That is, the minimal angle of every element should be bounded away from zero by a common bound that probably depends only on the initial grid (cf. (1.52)).

Second, a new grid resulting from a local refinement should contain all the nodes of the old grid. In particular, if continuous finite element spaces $\{V_{h_k}\}$ are exploited for a second-order partial differential problem in all levels, consecutive refinements should lead to a *nested* sequence of these spaces:

$$V_{h_1} \subset V_{h_2} \subset \cdots \subset V_{h_k} \subset V_{h_{k+1}} \subset \cdots ,$$

where $h_{k+1} < h_k$ and recall that h_k is the mesh size at the kth grid level. In the case of irregular local refinements, to preserve continuity of functions in these spaces the function values at the irregular nodes of a new grid are obtained by *polynomial interpolation* of the values at the old grid nodes.

Third, as defined before, the *irregularity index* of a grid is the maximum number of irregular vertices belonging to an edge of an element. There are reasons to restrict ourselves to 1-irregular grids. In practice, it seems to be very unlikely that grids with a higher irregularity index can be useful for a local h-scheme. Also, in general, the stiffness matrix arising from the finite element discretization of a problem should be sparse; see Sect. 1.10. It turns out that the sparsity cannot be guaranteed for a general irregular grid (Bank et al., 1983). To produce 1-irregular grids, we can employ the *1-irregular rule*: Refine any unrefined element for which any of the edges contains more than one irregular node.

6.1.3 Unrefinements

As noted, an h-scheme can be also employed to *unrefine* a grid. There are two factors that decide if an element needs to be unrefined: (1) a local error indicator and (2) a structural condition imposed on the grid resulting from the regularity or 1-irregularity requirement. These two factors must be examined before an element is unrefined.

When an element is refined, it produces a number of new smaller elements; the old element is called a *father* and the smaller ones are termed its *sons*. A *tree structure* (or *family structure*) consists of remembering for each element its father (if there is one) and its sons. Figure 6.3 shows a typical tree structure, together with a corresponding current grid generated by consecutive refinements of a single square. The *root* of the tree originates at the initial element and the *leaves* are those elements being not refined.

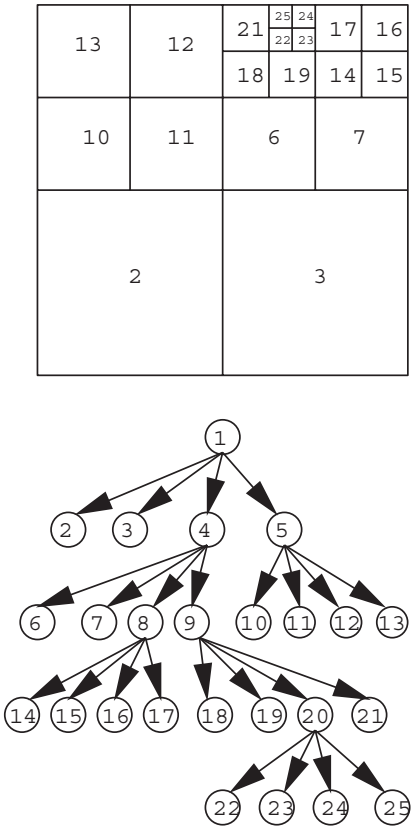

Fig. 6.3. A local refinement and the corresponding tree structure

The tree structure provides for easy and fast unrefinements. When the tree information is stored, a local unrefinement can be done by simply "cutting the corresponding branch" of the tree, i.e., unrefining previously refined elements and restoring locally the previous grid. This tree structure will be further discussed in the data management of local refinements in the next section.

6.2 Data Structures

In the finite element method developed in Chap. 1, all elements and nodes are usually numbered in a consecutive fashion so that a minimal band in the stiffness matrix of a finite element system can be produced (cf. Sect. 1.10). When a computational code identifies an element to evaluate its contribution to this matrix, the minimal information required is the set of node numbers corresponding to this element (cf. Sect. 1.1.4).

Adaptive local refinements and unrefinements require much more complex data structures than the classical global ones in Chap. 1. Because elements and nodes are added and deleted adaptively, it is often impossible to number them in a consecutive fashion. Hence we need to establish some kind of *natural order of elements*. In particular, all elements must be placed in an order and a code must recognize, for a given element, the next element (or the previous element if necessary) in the sequence. Therefore, for an element, the following information should be stored:

- nodes,
- neighbors,
- father,
- sons,
- level of refinement.

For a given node, its coordinates are also needed. The logic of a data structure corresponding to a particular local refinement may need additional information. However, the above listed information seems to be the minimal requirement for all existing data structures.

There are several data structures available for adaptive local grid refinements and unrefinements (Rheinboldt-Mesztenyi, 1980; Bank et al., 1983; Rivara, 1984B). As an example, we discuss the Rheinboldt-Mesztenyi tree-like data structure. This data structure has been designed to treat arbitrary irregular grids resulting from a refinement of an element, contrary to the discussion on 1-irregular grids in Sect. 6.1.2. A number of connected elements form an initial grid. Each element has its own data structure and some additional information (in the above tree-like manner) is stored to handle the initial element interfaces. For simplicity, we focus our discussion on the data structure supporting local refinements of a single square element (the shaded square $(0,1) \times (0,1)$ in Fig. 6.4, where this square is regarded as a son of the bigger square $(-1,3) \times (-1,3)$).

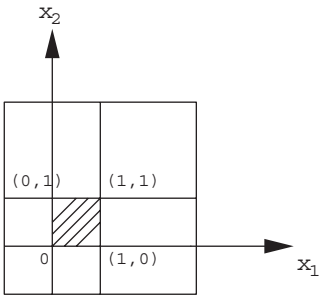

Fig. 6.4. An initial grid

A typical local refinement is shown in Fig. 6.5. If further refinements are developed, the number of neighbors for a typical element is theoretically unlimited, and it is clear that storing the neighbors for this element in an explicit form is practically impossible. The presence of irregular nodes makes it more difficult to use a static structure where we remember the nodes for this element and the elements adjacent to each node. Because the tree structure must be maintained anyway, we modify it in such a way that all the necessary information in the refinement and/or solution process can be reconstructed from the tree. Such a tree associated with the grid in Fig. 6.5 is presented in Fig. 6.6.

Several observations necessary to understand the tree in Fig. 6.6 are (Oden-Demkowicz, 1988):

- The element numbers are identified with the central node numbers. Specifically, when an element (e.g., element 11) is to be refined, a new (central) node is created which takes on the number of the element (in this case,

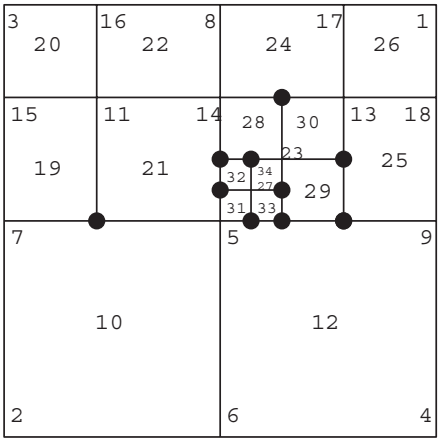

Fig. 6.5. An example of local refinement

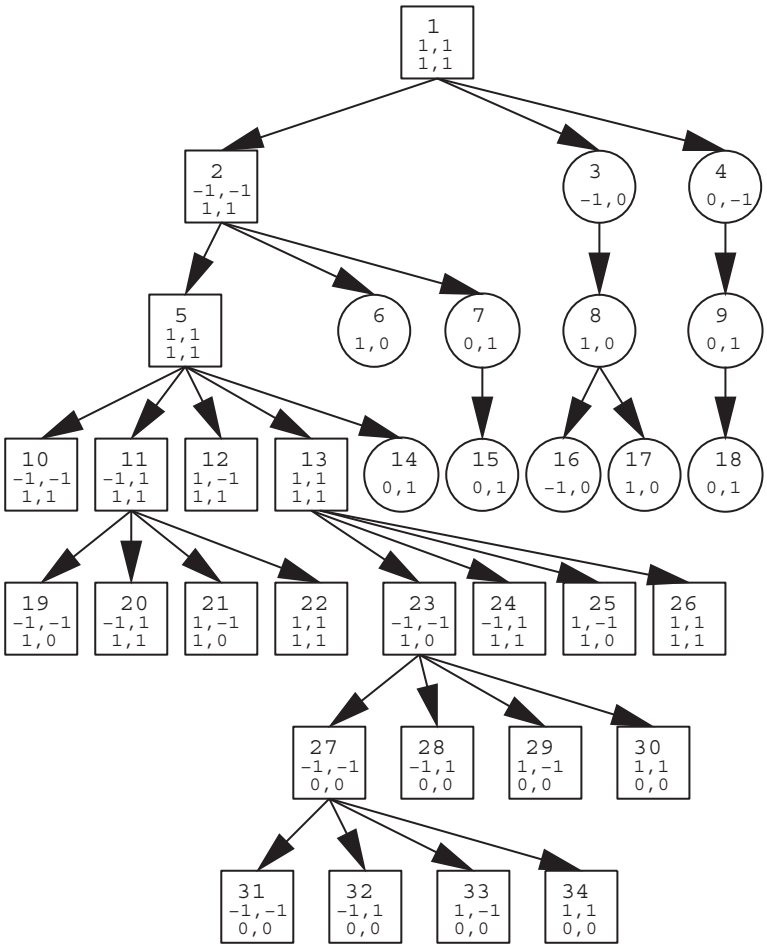

Fig. 6.6. The data structure

node 11). Hence, while consecutive numbers are used to enumerate nodes and elements, some of the numbers represent elements and others denote nodes.

- When an element is refined, it gives rise to four sons (except for elements 1 and 2 which are artificially introduced to handle the initial information about the initial square $(0,1) \times (0,1)$). The sons are assigned the first available numbers which must be remembered by their father. Referring to Fig. 6.6, the sole son of element 1 (i.e., square $(-1,3) \times (-1,3)$) is element 2. Analogously, element 2 has the only son, element 5, which has sons numbered 10, 11, 12, and 13. Element 11 gives rise to sons 19, 20, 21, and 22.

- When an element is refined, it may give rise not only to new elements, but also to new *active* nodes. For example, suppose that element 11 has been just subdivided, resulting in the creation of the new regular node 14. According to the rule, this node is associated with node 5. Roughly speaking, when a father-element is refined, it gives rise to four son-elements. If two neighboring sons are refined and a new common node is created, the node is assigned to their father. Thus every element has up to four son-elements and four *daughter-nodes*. The daughter-nodes must be remembered. For example, in Figs. 6.5 and 6.6, element 2 has two daughter-nodes numbered 6 and 7.
- When two neighboring elements that are not sons of the same father are refined (e.g., elements 22 and 24 in Fig. 6.5), a new node is created, which is assigned to the center node of a line half of which constitutes the common boundary of the elements (in this case, to node 14). Node 14 is not identified with any element; it is the daughter-node of element 5. It is clear that every daughter-node may have in turn two daughter-nodes.
- Each node is assigned a label indicating a relationship with its father-element or mother-node. Every daughter-node has only one label representing the direction to its parent; the label $(1, 0)$ for node 8 indicates that one must have right from its parent-node 3 to reach node 8, for example. Sons are assigned two labels. The first indicates the direction from their father, while the second represents a regularity tag describing which of the four nodes are irregular.

Virtually every essential piece of information about an element and its nodes can be reconstructed from such a modified tree structure. The principal idea is to travel up and down the tree making use of the labels and to collect all the necessary information (Rheinboldt-Mesztenyi, 1980).

In summary, precise information on the storage requirements is difficult to obtain. Theoretically, because we do not distinguish between nodes and elements, for every node one must remember

- the number of its parent,
- the numbers of its up to four sons,
- the numbers of its up to four daughters, and
- the labels.

6.3 A-Posteriori Error Estimates for Stationary Problems

We now study the second component of the adaptive finite element method: a-posteriori error estimation. *A-posteriori error estimators* and *indicators* can be utilized to give a specific assessment of errors and to form a solid basis for local refinements and unrefinements.

A-posteriori error estimators can be roughly classified as follows (Verfürth, 1996):

1. *Residual estimators.* These estimators bound the error of the computed approximate solution by a suitable norm of its residual with respect to the strong form of a differential equation (Babuška-Rheinboldt, 1978a).
2. *Local problem-based estimators.* This approach solves locally discrete problems, which are similar to, but simpler than, the original problem, and uses appropriate norms of the local solutions for error estimation (Babuška-Rheinboldt, 1978b; Bank-Weiser, 1985).
3. *Averaging-based estimators.* This approach utilizes a local extrapolation or averaging technique in error estimation (Zienkiewicz-Zhu, 1987).
4. *Hierarchical basis estimators.* This approach calculates the residual of the computed approximate solution with respect to another finite element space of higher-order polynomials or with respect to a refined grid (Deuflhard et al., 1989).

Following Verfürth (1996), we briefly study these four different approaches.

6.3.1 Residual Estimators

For the purpose of introduction, we consider the model problem in two dimensions

$$\begin{aligned} -\Delta p &= f && \text{in } \Omega\ , \\ p &= 0 && \text{on } \Gamma_D\ , \\ \frac{\partial p}{\partial \nu} &= g && \text{on } \Gamma_N, \end{aligned} \tag{6.1}$$

where Ω is a bounded domain in the plane with boundary $\bar{\Gamma} = \bar{\Gamma}_D \cup \bar{\Gamma}_N$, $\Gamma_D \cap \Gamma_N = \emptyset$, $f \in L^2(\Omega)$ and $g \in L^2(\Gamma_N)$ are given functions, and the *Laplacian operator* Δ is defined as in Sect. 1.1.2. We only study this simple problem; for generalizations to more general problems, refer to Chap. 1 or the references cited in this chapter.

Assume that Γ_D is closed relative to Γ and has a positive length. Define (cf. Sect. 1.2)

$$V = \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_D\}\ .$$

Also, introduce the notation

$$a(p, v) = \int_\Omega \nabla p \cdot \nabla v \ d\mathbf{x}, \quad L(v) = \int_\Omega f v \ d\mathbf{x} + \int_{\Gamma_N} g v \ d\ell, \qquad v \in V\ .$$

As in (1.20), problem (6.1) can be recast in the variational form:

$$\text{Find } p \in V \text{ such that } a(p, v) = L(v) \qquad \forall v \in V. \tag{6.2}$$

Let Ω be a convex polygonal domain (or its boundary Γ is smooth), and let K_h be a triangulation of Ω into triangles K of diameter h_K, as in Sect. 1.1.2. To the triangulation K_h, associate a grid function $h(\mathbf{x})$ such that, for some positive constant C_1,

$$C_1 h_K \le h(\mathbf{x}) \le h_K \qquad \forall \mathbf{x} \in K, \quad K \in K_h \; . \tag{6.3}$$

Moreover, assume that there exists a positive constant C_2 such that

$$C_2 h_K^2 \le |K| \qquad \forall K \in K_h \; , \tag{6.4}$$

where $|K|$ is the area of K. Recall that (6.4) is the *minimum angle* condition stating that the angles of triangles in K_h are bounded below by C_2 (cf. (1.52)).

To keep the notation to a minimum, let $V_h \subset V$ be defined by

$$V_h = \{v \in V : v|_K \in P_1(K), \; K \in K_h\} \; .$$

An extension to finite element spaces of higher-order polynomials will be noted at the end of this subsection. The finite element method for (6.1) is formulated:

$$\text{Find } p_h \in V_h \text{ such that } a(p_h, v) = L(v) \qquad \forall v \in V_h \; . \tag{6.5}$$

It follows from (6.2) and (6.5) that

$$a(p - p_h, v) = L(v) - a(p_h, v) \qquad \forall v \in V. \tag{6.6}$$

The right-hand side of (6.6) implicitly defines the *residual* of p_h as an element in the dual space of V. Because Γ_D has a positive length, Poincaré's inequality holds (cf. (1.36)):

$$\|v\|_{L^2(\Omega)} \le C(\Omega) \, \|\nabla v\|_{\mathbf{L}^2(\Omega)} \qquad \forall v \in V, \tag{6.7}$$

where we recall that C depends on Ω and the length of Γ_D. Using (6.7) and Cauchy's inequality (1.10), we have

$$\begin{aligned} \frac{1}{1 + C^2(\Omega)} \|v\|_{H^1(\Omega)} &\le \sup\{a(v, w) : w \in V, \; \|w\|_{H^1(\Omega)} = 1\} \\ &\le \|v\|_{H^1(\Omega)} \; . \end{aligned} \tag{6.8}$$

Consequently, it follows from (6.6) and (6.8) that

$$\begin{aligned} &\sup \big\{ L(v) - a(p_h, v) : v \in V, \; \|v\|_{H^1(\Omega)} = 1 \big\} \\ &\le \|p - p_h\|_{H^1(\Omega)} \\ &\le \big(1 + C^2(\Omega)\big) \sup \big\{ L(v) - a(p_h, v) : v \in V, \; \|v\|_{H^1(\Omega)} = 1 \big\} \; . \end{aligned} \tag{6.9}$$

Since the supremum term in (6.9) is equivalent to the norm of the residual in the dual space of V, this inequality implies that the norm in V of the error is,

up to multiplicative constants, bounded from above and below by the norm of the residual in the dual space of V. Most a-posteriori error estimators attempt to bound this dual norm of the residual by quantities that can be more easily evaluated from f, g, and p_h.

Let $\mathcal{E}_h^o$ denote the set of all interior edges e in K_h, $\mathcal{E}_h^b$ the set of the edges e on Γ, and $\mathcal{E}_h = \mathcal{E}_h^o \cup \mathcal{E}_h^b$. Furthermore, let $\mathcal{E}_h^D$ and $\mathcal{E}_h^N$ be the sets of edges e on Γ_D and Γ_N, respectively.

With each $e \in \mathcal{E}_h$, we associate a unit normal vector $\boldsymbol{\nu}$. For $e \in \mathcal{E}_h^b$, $\boldsymbol{\nu}$ is just the outward unit normal to Γ. For $e \in \mathcal{E}_h^o$, with $e = \bar{K}_1 \cap \bar{K}_2$, $K_1, K_2 \in K_h$, the direction of $\boldsymbol{\nu}$ is associated with the definition of jumps across e; for $v \in H^l(T_h)$ with $l > 1/2$, if the jump of v across e is defined by

$$[\![v]\!] = (v|_{K_2})|_e - (v|_{K_1})|_e \ , \tag{6.10}$$

then $\boldsymbol{\nu}$ is defined as the unit normal exterior to K_2 (cf. Fig. 4.11 or Fig. 5.15).

We recall the scalar product notation

$$(v, w)_{\mathcal{S}} = \int_{\mathcal{S}} v(\mathbf{x}) w(\mathbf{x}) \, d\mathbf{x}, \qquad v,\, w \in L^2(\mathcal{S}) \ .$$

If $\mathcal{S} = \Omega$, we omit it in this notation. Note that, by Green's formula (1.19), the definition of $L(\cdot)$, and the fact that $\Delta p_h = 0$ on all $K \in K_h$,

$$\begin{aligned} L(v) - a(p_h, v) &= L(v) - \sum_{K \in K_h} (\nabla p_h, \nabla v)_K \\ &= L(v) - \sum_{K \in K_h} \left[(\nabla p_h \cdot \boldsymbol{\nu}_K, v)_{\partial K} - (\Delta p_h, v)_K \right] \\ &= (f, v) + \sum_{e \in \mathcal{E}_h^N} (g - \nabla p_h \cdot \boldsymbol{\nu}, v)_e - \sum_{e \in \mathcal{E}_h^o} ([\![\nabla p_h \cdot \boldsymbol{\nu}]\!], v)_e \ . \end{aligned} \tag{6.11}$$

Applying (6.9) and (6.11), one can show that (cf. Sect. 6.7)

$$\begin{aligned} \|p - p_h\|_{H^1(\Omega)} \le C \Bigg\{ & \sum_{K \in K_h} h_K^2 \|f\|_{L^2(K)}^2 \\ & + \sum_{e \in \mathcal{E}_h^N} h_e \|g - \nabla p_h \cdot \boldsymbol{\nu}\|_{L^2(e)}^2 + \sum_{e \in \mathcal{E}_h^o} h_e \|[\![\nabla p_h \cdot \boldsymbol{\nu}]\!]\|_{L^2(e)}^2 \Bigg\}^{1/2} , \end{aligned} \tag{6.12}$$

where C depends on C_2 in (6.2) and $C(\Omega)$ in (6.7), and h_K and h_e represent the diameter and length, respectively, of K and e.

The right-hand side in (6.12) can be utilized as an a-posteriori error estimator because it involves only the known data f and g, the approximate solution p_h, and the geometrical data of the triangulation K_h. For general functions f and g, the exact computation of the integrals in the first and second terms of the right-hand side of (6.12) is often impossible. These integrals must be approximated by appropriate quadrature formulas (cf. Sect. 1.6).

On the other hand, it is also possible to approximate f and g by polynomials in suitable finite element spaces. Both approaches, numerical quadrature and approximation by simpler functions combined with exact integration of the latter functions, are often equivalent and generate analogous a-posteriori estimators. We restrict ourselves to the simpler function approximation approach. In particular, let f_h and g_h be the L^2-projections of f and g into the spaces of piecewise constants with respect to K_h and $\mathcal{E}_h^N$, respectively; i.e., on each $K \in K_h$ and $e \in \mathcal{E}_h^N$, $f_K = f_h|_K$ and $g_e = g_h|_e$ are given by the local mean values

$$f_K = \frac{1}{|K|} \int_K f \, d\mathbf{x}, \quad g_e = \frac{1}{h_e} \int_e g \, d\ell \, . \tag{6.13}$$

Then we define a *residual a-posteriori error estimator*:

$$\begin{aligned} \mathcal{R}_K = \Big\{ & h_K^2 \|f_K\|^2_{L^2(K)} + \sum_{e \in \partial K \cap \mathcal{E}_h^N} h_e \|g_e - \nabla p_h \cdot \boldsymbol{\nu}\|^2_{L^2(e)} \\ & + \frac{1}{2} \sum_{e \in \partial K \cap \mathcal{E}_h^o} h_e \|[\nabla p_h \cdot \boldsymbol{\nu}]\|^2_{L^2(e)} \Big\}^{1/2} . \end{aligned} \tag{6.14}$$

The first term in $\mathcal{R}_K$ is related to the residual of p_h with respect to the strong form of the differential equation. The second and third terms reflect the facts that p_h does not exactly satisfy the Neumann boundary condition and that $p_h \notin H^2(\Omega)$. Since interior edges are counted twice, combining (6.12), (6.14), and the triangle inequality, we obtain (cf. Exercise 6.3)

$$\begin{aligned} \|p - p_h\|_{H^1(\Omega)} \leq C \Big\{ & \sum_{K \in K_h} \left(\mathcal{R}_K^2 + h_K^2 \|f - f_K\|^2_{L^2(K)} \right) \\ & + \sum_{e \in \mathcal{E}_h^N} h_e \|g - g_e\|^2_{L^2(e)} \Big\}^{1/2} . \end{aligned} \tag{6.15}$$

Based on (6.15), with a given tolerance $\epsilon > 0$, an *adaptive algorithm* (termed Algorithm I) can be defined as follows (below RHS denotes the right-hand side of (6.15)):

- Choose an initial grid K_{h_0} with grid size h_0, and find a finite element solution p_{h_0} using (6.5) with $V_h = V_{h_0}$;
- Given a solution p_{h_k} in V_{h_k} with grid size h_k, stop if the following stopping criterion is satisfied:

$$RHS \leq \epsilon \, ; \tag{6.16}$$

- If (6.16) is violated, find a new grid K_{h_k} with grid size h_k such that the following equation is satisfied:

$$RHS = \epsilon \, , \tag{6.17}$$

and continue.

Inequality (6.16) is the stopping criterion and equation (6.17) defines the adaptive strategy. It follows from (6.15) that the estimate $\|p - p_h\|_{H^1(\Omega)}$ is bounded by ϵ if (6.16) is reached with $p_h = p_{h_k}$. Equation (6.17) determines a new grid size h_k by maximality. Namely, we seek a grid function h_k as large as possible (to maintain efficiency) such that (6.17) is satisfied. The maximality is generally determined by *equidistribution* of an error such that the error contributions from the individual elements K are approximately equal. Let M_{h_k} be the number of elements in K_{h_k}; equidistribution means that

$$(RHS|_K)^2 = \frac{\epsilon^2}{M_{h_k}}, \qquad K \in K_{h_k} \ .$$

Since the solution p_{h_k} depends on K_{h_k}, this is a nonlinear equation. The nonlinearity can be simplified by replacing M_{h_k} by $M_{h_{k-1}}$ (the previous level), for example.

The following inequality implies, in a sense, that the converse of (6.15) also holds (cf. Sect. 6.7): for $K \in K_h$,

$$\begin{aligned} \mathcal{R}_K \le C \Bigg\{ & \sum_{K' \in \Omega_K} \Big(\|p - p_h\|^2_{H^1(K')} + h^2_{K'} \|f - f_{K'}\|^2_{L^2(K')} \Big) \\ & + \sum_{e \in \partial K \cap \mathcal{E}_h^N} h_e \|g - g_e\|^2_{L^2(e)} \Bigg\}^{1/2} , \end{aligned} \tag{6.18}$$

where (cf. Fig. 6.7)

$$\Omega_K = \bigcup \{ K' \in K_h \ : \ \partial K' \cap \partial K \neq \emptyset \} \ .$$

Estimate (6.18) indicates that Algorithm I is *efficient* in the sense that the computational grid produced by this algorithm is not overly refined for a given accuracy, while (6.16) implies that this algorithm is *reliable* in the sense that the H^1-error is guaranteed to be within a given tolerance. The efficiency of error estimators will be further discussed in Sect. 6.3.5.

We end this subsection with a couple of remarks. First, it is also possible to control the error in norms other than the H^1-norm; we can control the

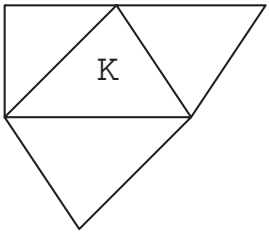

Fig. 6.7. An illustration of Ω_K

gradient error in the maximum norm (the L^∞-norm; cf. Johnson, 1994), for example. Second, the results in this section carry over to finite element spaces of polynomials of degree $r \geq 2$. In this case, f_h and g_h are the L^2-projections of f and g into the spaces of piecewise polynomials of degree $r-1$ with respect to K_h and $\mathcal{E}_h^N$, respectively, and f_K in the first term of $\mathcal{R}_K$ is replaced by $\Delta p_h|_K + f_K$ (cf. Exercise 6.4).

Example 6.1. This example follows Verfürth (1996). Consider problem (6.1) on a circular segment centered at the origin, with radius one and angle $3\pi/2$ (cf. Fig. 6.8). The function f is zero, and the solution p vanishes on the straight parts of the boundary Γ and has a normal derivative $\frac{2}{3}\cos\left(\frac{2}{3}\theta\right)$ on the curved part of Γ. In terms of polar coordinates, the exact solution p to (6.1) is $p = r^{2/3}\sin\left(\frac{2}{3}\theta\right)$. We calculate the finite element solution p_h using (6.5) with the space of piecewise linear functions V_h associated with the two triangulations shown in Fig. 6.8. The left triangulation is constructed by five uniform refinements of an initial triangulation K_{h_0}, which is composed of three right-angled isosceles triangles with short edges of unit length. In each refinement step, every triangle is divided into four smaller triangles by connecting the midpoints of its edges, as mentioned in Sect. 6.1.1. The midpoint of an edge having its two endpoints on Γ is projected onto Γ. The right triangulation in Fig. 6.8 is obtained from K_{h_0} by using Algorithm I based on the error estimator in (6.14). A triangle $K \in K_h$ is divided into four smaller triangles if $\mathcal{R}_K \geq 0.5 \max_{K' \in K_h} \mathcal{R}_{K'}$. Again, the midpoint of an edge having its two endpoints on Γ is projected onto Γ. For both triangulations, Table 6.1 lists the number of triangles (NT), the number of unknowns (NN), the relative error $e_r = \|p - p_h\|_{H^1(\Omega)} / \|p\|_{H^1(\Omega)}$, and the measurement $m_q = (\sum_{K \in K_h} \mathcal{R}_K^2)^{1/2} / \|p - p_h\|_{H^1(\Omega)}$ of the quality of the error estimator. From this table we clearly see the advantage of the adaptive method and the reliability of the error estimator.

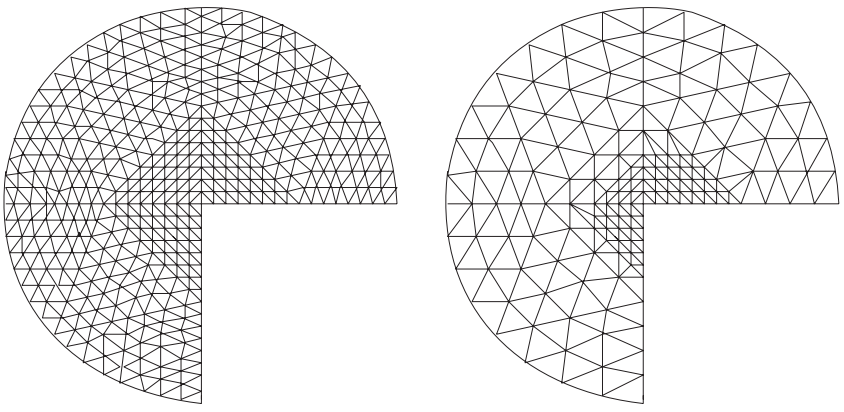

Fig. 6.8. Uniform (*left*) and adaptive (*right*) triangulations

Table 6.1. A comparison of uniform and adaptive refinements

Refinement	NT	NN	e_r	m_q
uniform	3072	1552	3.8%	0.7
adaptive	298	143	2.8%	0.6

6.3.2 Local Problem-Based Estimators

The results in the previous subsection show that we must accurately and reliably estimate the norm of the residual as an element in the dual space of V. This can be accomplished by lifting the residual to a suitable subspace of V through solving auxiliary problems analogous to, but simpler than, the original discrete problem (6.5). Practical consideration and the results in the previous subsection suggest that these auxiliary problems should possess the following properties:

- To obtain information about the local behavior of the error $p - p_h$, they should involve a small number of elements in K_h.
- To produce an accurate error, they should utilize finite element spaces of higher order than the original space.
- To minimize computation, they should involve as few degrees of freedom as possible.

There are many possible ways to make the auxiliary problems satisfy these properties. In this subsection, we present three of them.

6.3.2.1 Local Dirichlet Problem Estimators I

For a triangle $K \in K_h$, let $\lambda_{K,1}$, $\lambda_{K,2}$, and $\lambda_{K,3}$ denote the *barycentric coordinates* of K (cf. Example 1.6). We define the *triangle bubble function* b_K by

$$b_K = \begin{cases} 27\lambda_{K,1}\lambda_{K,2}\lambda_{K,3} & \text{in } K\ , \\ 0 & \text{in } \Omega \setminus K\ . \end{cases} \tag{6.19}$$

Also, for $e \in \mathcal{E}_h^o$ with $e = \bar{K}_1 \cap \bar{K}_2$, we enumerate the vertices of K_1 and K_2 such that the vertices of e are numbered first. Then we define the *edge bubble function* b_e by

$$b_e = \begin{cases} 4\lambda_{K_i,1}\lambda_{K_i,2} & \text{in } K_i,\ i = 1, 2\ , \\ 0 & \text{in } \Omega \setminus \{K_1 \cup K_2\}\ . \end{cases} \tag{6.20}$$

For $e \in \mathcal{E}_h^b$, b_e can be defined similarly. Finally, for a node $\mathbf{m}$, let $\Omega_{\mathbf{m}}$ indicate the union of triangles with the common vertex $\mathbf{m}$ (cf. Fig. 6.9); i.e.,

$$\Omega_{\mathbf{m}} = \bigcup_{K \in K_h\ :\ \mathbf{m} \in \partial K} K\ .$$

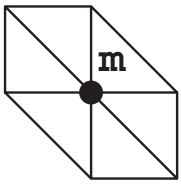

Fig. 6.9. An illustration of $\Omega_{\mathbf{m}}$

Now, for all nodes $\mathbf{m}$ in $\mathcal{E}_h^o \cup \mathcal{E}_h^N$, we introduce the space

$$V_{\mathbf{m}} = \text{span}\,\{b_K,\ K \in \Omega_{\mathbf{m}};\ b_e,\ \mathbf{m} \in e \in \mathcal{E}_h^o;\ b_e,\ e \in \partial\Omega_{\mathbf{m}} \cap \Gamma_N\}\ ,$$

and the error estimator

$$\mathcal{R}_{D,\mathbf{m}} = \|\nabla p_{\mathbf{m}}\|_{\mathbf{L}^2(\Omega_{\mathbf{m}})}\ , \tag{6.21}$$

where $p_{\mathbf{m}} \in V_{\mathbf{m}}$ is the solution of the discrete problem

$$\begin{aligned}(\nabla p_{\mathbf{m}}, \nabla v)_{\Omega_{\mathbf{m}}} = &\sum_{K \in \Omega_{\mathbf{m}}} (f_K, v)_K + \sum_{e \in \partial\Omega_{\mathbf{m}} \cap \Gamma_N} (g_e, v)_e \\ &-(\nabla p_h, \nabla v)_{\Omega_{\mathbf{m}}} \qquad \forall v \in V_{\mathbf{m}}\ .\end{aligned} \tag{6.22}$$

To see a different interpretation of (6.22), set

$$w_{\mathbf{m}} = p_h + p_{\mathbf{m}}\ .$$

Then we have

$$\mathcal{R}_{D,\mathbf{m}} = \|\nabla(p_h - w_{\mathbf{m}})\|_{\mathbf{L}^2(\Omega_{\mathbf{m}})}\ ,$$

and $w_{\mathbf{m}} \in p_h + V_{\mathbf{m}}$ is the solution of the problem

$$\begin{aligned}(\nabla w_{\mathbf{m}}, \nabla v)_{\Omega_{\mathbf{m}}} = &\sum_{K \in \Omega_{\mathbf{m}}} (f_K, v)_K + \sum_{e \in \partial\Omega_{\mathbf{m}} \cap \Gamma_N} (g_e, v)_e \\ &\forall v \in V_{\mathbf{m}}\ ,\end{aligned} \tag{6.23}$$

which is a discrete analogue of the Dirichlet problem

$$\begin{aligned}-\Delta w &= f && \text{in } \Omega_{\mathbf{m}}\ , \\ w &= p_h && \text{on } \partial\Omega_{\mathbf{m}} \setminus \Gamma_N\ , \\ \frac{\partial w}{\partial \boldsymbol{\nu}} &= g && \text{on } \partial\Omega_{\mathbf{m}} \cap \Gamma_N\ .\end{aligned}$$

Hence (6.22) can be thought of solving a local analogue of the residual equation (6.6) by a higher-order finite element approximation and of using a suitable norm of the solution as an error estimator, while (6.23) can be thought of solving a local discrete analogue of problem (6.1) in a higher-order finite

element space and of comparing the solution of this local problem with that of problem (6.5).

The following estimates imply that the error estimator $\mathcal{R}_{D,\mathbf{m}}$ is comparable to $\mathcal{R}_K$ (Verfürth, 1996):

$$\begin{aligned} \mathcal{R}_{D,\mathbf{m}}^2 &\le C \sum_{K\in\Omega_\mathbf{m}} \mathcal{R}_K^2, \qquad & \mathbf{m} &\in \mathcal{E}_h^o \cup \mathcal{E}_h^N \ , \\ \mathcal{R}_K^2 &\le C \sum_{\mathbf{m}\in K\setminus\mathcal{E}_h^D} \mathcal{R}_{D,\mathbf{m}}^2, \qquad & K &\in K_h \ , \end{aligned} \tag{6.24}$$

where the constants C depend only on C_2 in (6.4). It can be also shown that this error estimator provides upper and lower bounds on the error $p - p_h$:

$$\begin{aligned} \|p-p_h\|_{H^1(\Omega)} &\le C\Bigg\{ \sum_{\mathbf{m}\in\mathcal{E}_h^o\cup\mathcal{E}_h^N} \mathcal{R}_{D,\mathbf{m}}^2 \\ &\quad + \sum_{K\in K_h} h_K^2 \|f-f_K\|_{L^2(K)}^2 + \sum_{e\in\mathcal{E}_h^N} h_e \|g-g_e\|_{L^2(e)}^2 \Bigg\}^{1/2} , \\ \mathcal{R}_{D,\mathbf{m}} &\le C\Bigg\{ \|p-p_h\|_{H^1(\Omega_\mathbf{m})}^2 \\ &\quad + \sum_{K\in\Omega_\mathbf{m}} h_K^2 \|f-f_K\|_{L^2(K)}^2 + \sum_{e\in\partial\Omega_\mathbf{m}\cap\Gamma_N} h_e \|g-g_e\|_{L^2(e)}^2 \Bigg\}^{1/2} . \end{aligned} \tag{6.25}$$

6.3.2.2 Local Dirichlet Problem Estimators II

We now introduce an error estimator that is a minor modification of the previous one. Instead of starting with nodes $\mathbf{m} \in \mathcal{E}_h^o \cup \mathcal{E}_h^N$ and the corresponding sets $\Omega_\mathbf{m}$, we begin with elements $K \in K_h$ and the associated sets Ω_K. For all $K \in K_h$, we introduce the space

$$\tilde{V}_K = \operatorname{span}\{ b_{K'},\ K' \in \Omega_K;\ b_e,\ e \in \partial K \cap \mathcal{E}_h^o;\ b_e,\ e \in \partial\Omega_K \cap \Gamma_N \} \ ,$$

and the error estimator

$$\mathcal{R}_{D,K} = \|\nabla \tilde{p}_K\|_{\mathbf{L}^2(\Omega_K)} \ , \tag{6.26}$$

where $\tilde{p}_K \in \tilde{V}_K$ is the solution of the discrete problem

$$\begin{aligned} (\nabla\tilde{p}_K, \nabla v)_{\Omega_K} = &\sum_{K'\in\Omega_K} (f_{K'}, v)_{K'} + \sum_{e\in\partial\Omega_K\cap\Gamma_N} (g_e, v)_e \\ &-(\nabla p_h, \nabla v)_{\Omega_K} \qquad \forall v \in \tilde{V}_K \ . \end{aligned} \tag{6.27}$$

As in the previous subsection, $\tilde{w}_K = p_h + \tilde{p}_K$ can be thought of as an approximate solution of the Dirichlet problem

$$
\begin{aligned}
-\Delta w &= f && \text{in } \Omega_K \ , \\
w &= p_h && \text{on } \partial\Omega_K \setminus \Gamma_N \ , \\
\frac{\partial w}{\partial \boldsymbol{\nu}} &= g && \text{on } \partial\Omega_K \cap \Gamma_N \ .
\end{aligned}
$$

Furthermore, (6.24) and (6.25) remain true (Verfürth, 1996):

$$
\begin{aligned}
\mathcal{R}_{D,K}^2 &\le C \sum_{K' \in \Omega_K} \mathcal{R}_{K'}^2, && K \in K_h \ , \\
\mathcal{R}_K^2 &\le C \sum_{K' \in \Omega_K} \mathcal{R}_{D,K'}^2, && K \in K_h \ ,
\end{aligned}
\tag{6.28}
$$

and

$$
\begin{aligned}
\|p - p_h\|_{H^1(\Omega)} &\le C \Bigg\{ \sum_{K \in K_h} \mathcal{R}_{D,K}^2 \\
&\quad + \sum_{K \in K_h} h_K^2 \|f - f_K\|_{L^2(K)}^2 + \sum_{e \in \mathcal{E}_h^N} h_e \|g - g_e\|_{L^2(e)}^2 \Bigg\}^{1/2} \ , \\
\mathcal{R}_{D,K} &\le C \Bigg\{ \|p - p_h\|_{H^1(\Omega_K)}^2 + \sum_{K' \in \Omega_K} h_{K'}^2 \|f - f_{K'}\|_{L^2(K')}^2 \\
&\qquad + \sum_{e \in \partial\Omega_K \cap \Gamma_N} h_e \|g - g_e\|_{L^2(e)}^2 \Bigg\}^{1/2} \ .
\end{aligned}
\tag{6.29}
$$

6.3.2.3 Local Neumann Problem Estimators

In the previous two subsections, the Dirichlet boundary conditions are used for the auxiliary problems. We now impose Neumann conditions. For each $K \in K_h$, we define the space

$$
V_K = \operatorname{span} \left\{ b_K;\ b_e,\ e \in \partial K \setminus \mathcal{E}_h^D \right\} \ ,
$$

and the error estimator

$$
\mathcal{R}_{N,K} = \|\nabla p_K\|_{\mathbf{L}^2(K)} \ , \tag{6.30}
$$

where $p_K \in V_K$ is the solution of the discrete problem

$$
\begin{aligned}
(\nabla p_K, \nabla v)_K = (f_K, v)_K &- \frac{1}{2} \sum_{e \in \partial K \cap \mathcal{E}_h^o} ([\nabla p_h \cdot \boldsymbol{\nu}], v)_e \\
&+ \sum_{e \in \partial K \cap \mathcal{E}_h^N} (g_e - \nabla p_h \cdot \boldsymbol{\nu}, v)_e \qquad \forall v \in V_K \ .
\end{aligned}
\tag{6.31}
$$

Problem (6.31) can be interpreted as a discrete analogue of the Neumann problem

$$
\begin{aligned}
-\Delta w &= f && \text{in } K\ , \\
w &= 0 && \text{on } \partial K \cap \Gamma_D\ , \\
\frac{\partial w}{\partial \boldsymbol{\nu}} &= \eta(p_h) && \text{on } \partial K \setminus \Gamma_D\ ,
\end{aligned}
$$

where

$$
\eta(p_h)|_e = \begin{cases} -[\nabla p_h \cdot \boldsymbol{\nu}]_e/2 & \text{if } e \in \mathcal{E}_h^o\ , \\ g_e - \nabla p_h \cdot \boldsymbol{\nu}_e & \text{if } e \in \mathcal{E}_h^N\ . \end{cases}
$$

Moreover, (6.24) and (6.25) hold for the present estimator (Verfürth, 1996); i.e.,

$$
\begin{aligned}
&\mathcal{R}_{N,K} \le C\mathcal{R}_K, && K \in K_h\ , \\
&\mathcal{R}_K^2 \le C \sum_{K' \in \Omega_K} \mathcal{R}_{N,K'}^2, && K \in K_h\ ,
\end{aligned}
\tag{6.32}
$$

and

$$
\begin{aligned}
\|p - p_h\|_{H^1(\Omega)} &\le C \Bigg\{ \sum_{K \in K_h} \mathcal{R}_{N,K}^2 \\
&\quad + \sum_{K \in K_h} h_K^2 \|f - f_K\|_{L^2(K)}^2 + \sum_{e \in \mathcal{E}_h^N} h_e \|g - g_e\|_{L^2(e)}^2 \Bigg\}^{1/2}\ , \\
\mathcal{R}_{N,K} &\le C \Bigg\{ \|p - p_h\|_{H^1(\Omega_K)}^2 + \sum_{K' \in \Omega_K} h_{K'}^2 \|f - f_{K'}\|_{L^2(K')}^2 \\
&\quad + \sum_{e \in \partial K \cap \Gamma_N} h_e \|g - g_e\|_{L^2(e)}^2 \Bigg\}^{1/2}\ .
\end{aligned}
\tag{6.33}
$$

Figures 6.7 and 6.9 show the generic forms of Ω_K, $K \in K_h$, and $\Omega_{\mathbf{m}}$, $\mathbf{m} \in \mathcal{E}_h^o \cup \mathcal{E}_h^N$, respectively. From these two figures we see that the respective dimensions of the discrete problems (6.22), (6.27), and (6.31) are 12, 7, and 4 (cf. Exercise 6.6). In general, the evaluation of $\mathcal{R}_{D,\mathbf{m}}$, $\mathcal{R}_{D,K}$, and $\mathcal{R}_{N,K}$ each is more expensive than that of $\mathcal{R}_K$ because of their construction.

6.3.3 Averaging-Based Estimators

In this subsection, we assume that $\Gamma_N = \emptyset$; i.e., we only consider a Dirichlet boundary value problem. Suppose that we can find an easily computable approximation $\mathbf{R}(p_h)$ of ∇p_h such that

$$
\|\nabla p - \mathbf{R}(p_h)\|_{\mathbf{L}^2(\Omega)} \le \gamma \|\nabla p - \nabla p_h\|_{\mathbf{L}^2(\Omega)}\ , \tag{6.34}
$$

where p and p_h are the respective solutions of (6.1) and (6.5), and the constant γ satisfies $0 \le \gamma < 1$. Then we see that

$$\begin{aligned}\frac{1}{1+\gamma}\|\mathbf{R}(p_h)-\nabla p_h\|_{\mathbf{L}^2(\Omega)} &\le \|\nabla p-\nabla p_h\|_{\mathbf{L}^2(\Omega)}\\ &\le \frac{1}{1-\gamma}\|\mathbf{R}(p_h)-\nabla p_h\|_{\mathbf{L}^2(\Omega)}\ .\end{aligned} \tag{6.35}$$

This inequality suggests that we may use $\|\mathbf{R}(p_h)-\nabla p_h\|_{L^2(\Omega)}$ as an error estimator. Because ∇p_h is a piecewise constant vector, its L^2-projection into the space of continuous piecewise linear vectors may satisfy (6.34). The evaluation of this projection, however, is as expensive as the solution of (6.5). Thus the idea is to replace the L^2-scalar product by an approximation that can be more easily computed.

We define the spaces

$$\begin{aligned}\mathbf{W}_h^{-1} &= \{\mathbf{v}\in(L^2(\Omega))^2 : \mathbf{v}|_K\in(P_1(K))^2,\ K\in K_h\}\ ,\\ \mathbf{W}_h &= \{\mathbf{v}\in(C(\bar\Omega))^2 : \mathbf{v}|_K\in(P_1(K))^2,\ K\in K_h\}\ .\end{aligned}$$

Note that $\nabla V_h\subset\mathbf{W}_h^{-1}$. Also, we introduce a mesh-dependent scalar product on $\mathbf{W}_h^{-1}$ by

$$\begin{aligned}(\mathbf{v},\mathbf{w})_h = \sum_{K\in K_h}\frac{|K|}{3}\Big(&\mathbf{v}(\mathbf{m}_{K,1})\cdot\mathbf{w}(\mathbf{m}_{K,1})\\ &+\mathbf{v}(\mathbf{m}_{K,2})\cdot\mathbf{w}(\mathbf{m}_{K,2})+\mathbf{v}(\mathbf{m}_{K,3})\cdot\mathbf{w}(\mathbf{m}_{K,3})\Big)\ ,\end{aligned} \tag{6.36}$$

where $\mathbf{m}_{K,i}$ are the vertices of the triangle K. Since the quadrature formula

$$\int_K v\ d\mathbf{x}\approx\frac{|K|}{3}\left(v(\mathbf{m}_{K,1})+v(\mathbf{m}_{K,2})+v(\mathbf{m}_{K,3})\right)$$

is exact for all linear functions (cf. Sect. 1.6), we see that

$$(\mathbf{v},\mathbf{w})_h=(\mathbf{v},\mathbf{w})\ , \tag{6.37}$$

if $\mathbf{v},\ \mathbf{w}\in\mathbf{W}_h^{-1}$ and one of them is piecewise constant. Moreover, it can be shown that (cf. Exercise 6.7)

$$\frac{1}{4}\|\mathbf{v}\|^2_{\mathbf{L}^2(\Omega)}\le(\mathbf{v},\mathbf{v})_h\le\|\mathbf{v}\|^2_{\mathbf{L}^2(\Omega)},\qquad \mathbf{v}\in\mathbf{W}_h^{-1}\ , \tag{6.38}$$

and

$$(\mathbf{v},\mathbf{w})_h=\frac{1}{3}\sum_{\mathbf{m}\in\mathcal{N}_h}|\Omega_{\mathbf{m}}|\mathbf{v}(\mathbf{m})\cdot\mathbf{w}(\mathbf{m}),\qquad \mathbf{v},\ \mathbf{w}\in\mathbf{W}_h\ , \tag{6.39}$$

where $\mathcal{N}_h$ is the set of vertices in K_h and $\Omega_{\mathbf{m}}$ is defined as in Sect. 6.3.2.1 (cf. Fig. 6.9).

We now define $\mathbf{R}(p_h) \in \mathbf{W}_h$ to be the L^2-projection of ∇p_h with respect to the discrete scalar product $(\cdot,\cdot)_h$; i.e.,

$$(\mathbf{R}(p_h), \mathbf{v})_h = (\nabla p_h, \mathbf{v})_h \qquad \forall \mathbf{v} \in \mathbf{W}_h \ . \tag{6.40}$$

Equations (6.37) and (6.39) imply that

$$\mathbf{R}(p_h)(\mathbf{m}) = \sum_{K \in \Omega_{\mathbf{m}}} \frac{|K|}{|\Omega_{\mathbf{m}}|} \nabla p_h|_K, \qquad \mathbf{m} \in \mathcal{N}_h \ . \tag{6.41}$$

Therefore, $\mathbf{R}(p_h)$ can be easily computed by a local averaging of ∇p_h. We define the error estimator

$$\mathcal{R}_Z = \left(\sum_{K \in K_h} \mathcal{R}_{Z,K} \right)^{1/2}, \quad \mathcal{R}_{Z,K} = \|\mathbf{R}(p_h) - \nabla p_h\|_{\mathbf{L}^2(K)} \ . \tag{6.42}$$

To see that $\mathcal{R}_Z$ is indeed an estimator, we compare $\mathcal{R}_{Z,K}$ with a modification $\tilde{\mathcal{R}}_K$ of the residual error estimator in Sect. 6.3.1:

$$\tilde{\mathcal{R}}_K = \left(\frac{1}{2} \sum_{e \in \partial K \cap \mathcal{E}_h^o} h_e \|[\nabla p_h \cdot \boldsymbol{\nu}]\|^2_{L^2(e)} \right)^{1/2} . \tag{6.43}$$

Set

$$\tilde{\mathcal{R}} = \left(\sum_{K \in K_h} \tilde{\mathcal{R}}_K^2 \right)^{1/2} . \tag{6.44}$$

Then one sees that (Rodriguez, 1994)

$$\begin{aligned} &\|p - p_h\|_{H^1(\Omega)} \le C \left(\tilde{\mathcal{R}}^2 + \sum_{K \in K_h} h_K^2 \|f\|^2_{L^2(K)} \right)^{1/2}, \\ &\tilde{\mathcal{R}}_K \le C \left(\|p - p_h\|^2_{H^1(\Omega_K)} + \sum_{K' \in \Omega_K} h_{K'}^2 \|f\|^2_{L^2(K')} \right)^{1/2}, \\ &C_3 \tilde{\mathcal{R}} \le \mathcal{R}_Z \le C_4 \tilde{\mathcal{R}} \ . \end{aligned} \tag{6.45}$$

6.3.4 Hierarchical Basis Estimators

The general idea of a hierarchical basis estimator is to estimate the error $p - p_h$ by computing the residual of p_h with respect to certain basis functions of another finite element space B_h that satisfies $V_h \subset B_h \subset V$ and either consists of higher-order finite elements or corresponds to a refinement of K_h. When these basis functions are appropriately scaled, (6.6) and Cauchy's inequality (1.10) imply that lower bounds on the error can be obtained. Furthermore, with a suitable choice of B_h, upper bounds can be also obtained.

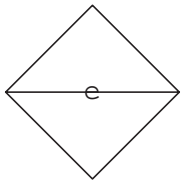

Fig. 6.10. An illustration of Ω_e

Let p and p_h be the respective solutions of (6.1) and (6.5), and B_h be a finite element space associated with K_h satisfying $V_h \subset B_h \subset V$. Suppose that to each edge $e \in \mathcal{E}_h^o \cup \mathcal{E}_h^N$, there corresponds a function $\varphi_e \in B_h \cap C_0(\Omega_e)$ such that

$$\begin{aligned} &0 \le \varphi_e \le 1, \quad Ch_e \le \int_e \varphi_e \, d\ell \,, \\ &Ch_e \|\varphi_e\|_{H^1(\Omega_e)} \le \|\varphi_e\|_{L^2(\Omega_e)} \,, \end{aligned} \tag{6.46}$$

where $\Omega_e = \bigcup\{K \in K_h : e \in \partial K\}$ (cf. Fig. 6.10) and $C_0(\Omega_e)$ is the set of continuous functions on Ω_e whose trace on $\partial\Omega_e$ is zero. Set

$$R_e = (f, \varphi_e) + (g, \varphi_e)_{\Gamma_N} - (\nabla p_h, \nabla \varphi_e) \,,$$

which is the residual of p_h with respect to φ_e, $e \in \mathcal{E}_h^o \cup \mathcal{E}_h^N$. Then the following a-posteriori error estimates hold:

$$\begin{aligned} &|R_e| \le C\|p - p_h\|_{H^1(\Omega_e)}, \qquad e \in \mathcal{E}_h^o \cup \mathcal{E}_h^N \,, \\ &\|p - p_h\|_{H^1(\Omega)} \le C\Big\{ \sum_{e \in \mathcal{E}_h^o \cup \mathcal{E}_h^N} R_e^2 + \sum_{K \in K_h} h_K^2 \|f\|_{L^2(K)}^2 \\ &\qquad\qquad + \sum_{e \in \mathcal{E}_h^N} h_e \|g - g_e\|_{L^2(e)}^2 \Big\}^{1/2} . \end{aligned} \tag{6.47}$$

Instead of using edges $e \in \mathcal{E}_h^o \cup \mathcal{E}_h^N$, we can utilize elements $K \in K_h$ as well. Suppose that to each $K \in K_h$, there be associated a function $\varphi_K \in B_h \cap C_0(K)$ such that

$$\begin{aligned} &0 \le \varphi_K \le 1, \quad Ch_K^2 \le \int_K \varphi_K \, d\mathbf{x} \,, \\ &Ch_K \|\varphi_K\|_{H^1(K)} \le \|\varphi_K\|_{L^2(K)} \,. \end{aligned} \tag{6.48}$$

Let

$$R_K = (f, \varphi_K) + (g, \varphi_K)_{\Gamma_N} - (\nabla p_h, \nabla \varphi_K)$$

be the residual of p_h with respect to φ_K, $K \in K_h$. Then we have the following a-posteriori error estimates:

$$
\begin{aligned}
&|R_K| \le C\|p - p_h\|_{H^1(K)}, \qquad K \in K_h \ , \\
&\|p - p_h\|_{H^1(\Omega)} \le C\Big\{ \sum_{e \in \mathcal{E}_h^o \cup \mathcal{E}_h^N} R_e^2 + \sum_{K \in K_h} R_K^2 \\
&\qquad + \sum_{K \in K_h} h_K^2 \|f - f_K\|_{L^2(K)}^2 + \sum_{e \in \mathcal{E}_h^N} h_e \|g - g_e\|_{L^2(e)}^2 \Big\}^{1/2} .
\end{aligned}
\tag{6.49}
$$

To see examples for these two results, we recall the definition of bubble functions. For $K \in K_h$ and $e \in \mathcal{E}_h$, b_K and b_e are defined as in (6.19) and (6.20). These bubble functions satisfy (cf. Exercise 6.8)

$$
\begin{aligned}
&\operatorname{supp}(b_K) \subset K, \quad 0 \le b_K \le 1, \quad \max_{\mathbf{x} \in K} b_K(\mathbf{x}) = 1 \ , \\
&Ch_K^2 \le \int_K b_K \, d\mathbf{x} = \frac{9}{20}|K| \ , \\
&Ch_K \|\nabla b_K\|_{\mathbf{L}^2(K)} \le \|b_K\|_{L^2(K)} \ ,
\end{aligned}
\tag{6.50}
$$

and

$$
\begin{aligned}
&\operatorname{supp}(b_e) \subset \Omega_e, \quad 0 \le b_e \le 1, \quad \max_{\mathbf{x} \in e} b_e(\mathbf{x}) = 1 \ , \\
&\int_e b_e \, d\ell = \frac{2}{3} h_e, \quad Ch_e^2 \le \int_K b_e \, d\mathbf{x} = \frac{1}{3}|K| \ , \quad K \in \Omega_e \ , \\
&Ch_e \|\nabla b_e\|_{\mathbf{L}^2(K)} \le \|b_e\|_{L^2(K)} \ , \quad K \in \Omega_e \ ,
\end{aligned}
\tag{6.51}
$$

where $\operatorname{supp}(b_K)$ denotes the support of b_K.

Using these properties of the bubbles, we now state several examples for B_h and the corresponding functions φ_e and φ_K that satisfy (6.46) and (6.48).

Example 6.2. Define

$$
B_h = \{ v \in V : v|_K \in P_2(K), \ K \in K_h \} \ ,
$$

and $\varphi_e = b_e$, $e \in \mathcal{E}_h^o \cup \mathcal{E}_h^N$.

Example 6.3. Set

$$
B_h = \{ v \in V : v|_K \in P_3(K), \ K \in K_h \} \ ,
$$

$\varphi_e = b_e$, $e \in \mathcal{E}_h^o \cup \mathcal{E}_h^N$, and $\varphi_K = b_K$, $K \in K_h$.

Example 6.4. Denote by $K_{h/2}$ the triangulation obtained by a uniform refinement of K_h; that is, each $K \in K_h$ is divided into four equal smaller triangles by joining the midpoints of the edges of K. Let

$$
B_h = \{ v \in V : v|_K \in P_1(K), \ K \in K_{h/2} \} \ ,
$$

and choose φ_e, $e \in \mathcal{E}_h^o \cup \mathcal{E}_h^N$, to be the nodal basis functions of B_h associated with the midpoint of e (cf. Fig. 6.11).

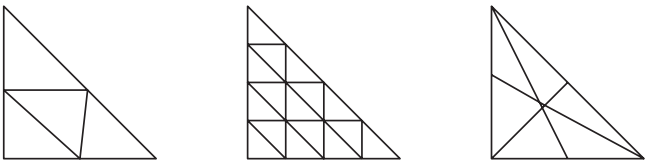

Fig. 6.11. Partition of K into 4, 16, and 6 smaller triangles

Example 6.5. Indicate by $K_{h/4}$ the triangulation obtained by a uniform refinement of $K_{h/2}$, define

$$B_h = \{v \in V : v|_K \in P_1(K),\ K \in K_{h/4}\}\ ,$$

and choose φ_e, $e \in \mathcal{E}_h^o \cup \mathcal{E}_h^N$, as in Example 6.4. For any $K \in K_h$, let $K' \in K_{h/4}$ be the triangle that has all its vertices in the interior of K (cf. Fig. 6.11). Then take φ_K to be the function that identically equals one on K' and zero on ∂K.

Example 6.6. Divide each triangle $K \in K_h$ into six smaller triangles by joining every vertex of K with the midpoint of the edge opposite to it, and denote by $\mathcal{K}_{h/2}$ the resulting triangulation. Then define

$$B_h = \{v \in V : v|_K \in P_1(K),\ K \in \mathcal{K}_{h/2}\}\ ,$$

and choose φ_K, $K \in K_h$, and φ_e, $e \in \mathcal{E}_h^o \cup \mathcal{E}_h^N$, to be the nodal basis functions of B_h corresponding to the centroids of triangles in K_h and to the midpoints of edges in $\mathcal{E}_h^o \cup \mathcal{E}_h^N$ (cf. Fig. 6.11), respectively.

We remark that the usual hierarchical basis estimators are presented in a slightly different form. We split B_h by

$$B_h = V_h \oplus S_h\ ,$$

and assume that V_h and S_h satisfy a *strengthened Cauchy-Schwarz inequality*

$$(\nabla v, \nabla w) \le \gamma_1 \|\nabla v\|_{\mathbf{L}^2(\Omega)} \|\nabla w\|_{\mathbf{L}^2(\Omega)}, \qquad v \in V_h,\ w \in S_h\ , \tag{6.52}$$

where the constant γ_1 satisfies $0 \le \gamma_1 < 1$. Also, we assume that there exists a symmetric bilinear form $b(\cdot,\cdot) : S_h \times S_h \to \mathbb{R}$ satisfying

$$\|\nabla v\|^2_{\mathbf{L}^2(\Omega)} \le b(v,v) \le \frac{1}{\gamma_2} \|\nabla v\|^2_{\mathbf{L}^2(\Omega)} \qquad \forall v \in S_h\ , \tag{6.53}$$

where $0 < \gamma_2 \le 1$. Let $q_h \in S_h$ be the solution of the problem

$$b(q_h, v) = (f, v) + (g, v)_{\Gamma_N} - (\nabla p_h, \nabla v) \qquad \forall v \in S_h\ , \tag{6.54}$$

where p_h is the solution of (6.5). The crucial point of introducing the bilinear form $b(\cdot,\cdot)$ is that (6.54) should be much easier to solve than the original problem (6.5). Set

$$\mathcal{R}_H = \sqrt{b(q_h, q_h)}\ . \tag{6.55}$$

If the space B_h satisfies a *saturation assumption* with respect to V_h, i.e., there exists a constant $0 \le \gamma_3 < 1$ such that

$$\min_{w \in B_h} \|\nabla(v-w)\|_{\mathbf{L}^2(\Omega)} \le \gamma_3 \min_{w \in V_h} \|\nabla(v-w)\|_{\mathbf{L}^2(\Omega)}, \quad w \in V, \tag{6.56}$$

then the following a-posteriori error estimate holds (Verfürth, 1996):

$$(1-\gamma_3)\sqrt{\frac{(1-\gamma_1)\gamma_2}{1+C^2(\Omega)}}\|p-p_h\|_{H^1(\Omega)} \le \mathcal{R}_H \le \frac{1}{\sqrt{\gamma_2}}\|p-p_h\|_{H^1(\Omega)}, \tag{6.57}$$

where the constant $C(\Omega)$ is defined in (6.7).

Example 6.7. Let B_h be defined as in Examples 6.2–6.6, and define

$$S_h = \{v \in B_h : v(\mathbf{x}) = 0 \quad \forall \mathbf{x} \in \mathcal{N}_h\}\ ,$$

where we recall that $\mathcal{N}_h$ is the set of vertices in K_h. The strengthened Cauchy-Schwarz inequality can be shown for these choices (Eijkhout-Vassilevski, 1991). The saturation assumption also holds except in the trivial case where

$$\lim_{h \to 0} \frac{1}{h}\|p-p_h\|_{H^1(\Omega)} = 0. \tag{6.58}$$

To define $b(\cdot,\cdot)$, denote by $\mathcal{N}_h'$ the set of nodes corresponding to B_h, e.g., $\mathcal{N}_h' = \mathcal{N}_{h/2}$ in Example 6.4 and $\mathcal{N}_h' = \mathcal{N}_{h/4}$ in Example 6.5. Then define

$$b(v,w) = \sum_{\mathbf{x} \in \mathcal{N}_h' \setminus \mathcal{N}_h} v(\mathbf{x})w(\mathbf{x}), \qquad v,\ w \in S_h\ .$$

It can be proven that (6.53) holds for this bilinear form on S_h (cf. Exercise 6.9). Also, with this choice of $b(\cdot,\cdot)$, the matrix corresponding to the left-hand side of (6.54) is diagonal with diagonal entries of order one. Thus, up to multiplicative constants of order one, $\mathcal{R}_H$ is equivalent to

$$\left(\sum_{K \in K_h} R_K^2 + \sum_{e \in \mathcal{E}_h^o \cup \mathcal{E}_h^N} R_e^2\right)^{1/2},$$

which recovers the error estimators in (6.47) and (6.49).

6.3.5 Efficiency of Error Estimators

The quality of an a-posteriori error estimator can be measured in terms of its *efficiency index*, i.e., the ratio of the estimated error and of the true error. An estimator is referred to as *efficient* if its efficiency index is bounded from above

and below for all grids, and *asymptotically exact* if this index approaches one when the grid size goes to zero.

In general, we have

$$\left\{ \sum_{K\in\Omega_{\mathbf{m}}} h_K^2 \|f - f_K\|^2_{L^2(K)} + \sum_{e\in\partial\Omega_{\mathbf{m}}\cap\Gamma_N} h_e \|g - g_e\|^2_{L^2(e)} \right\}^{1/2} = o(h) \ ,$$

where $h = \max\limits_{K\in K_h} h_K$. On the other hand, the solutions to (6.2) and (6.5) satisfy

$$\|p - p_h\|_{H^1(\Omega)} \geq Ch \ ,$$

except in the trivial case (6.58). Thus the a-posteriori error estimates in Sects. 6.3.1–6.3.4 show that the corresponding error estimators are efficient (cf. Exercise 6.10). Their efficiency indices can be in principle estimated explicitly since the constants in these a-posteriori estimates depend only on C_2 in (6.4) and $C(\Omega)$ in (6.7). Furthermore, applying *superconvergence* results, it can be proven that the error estimates in Sects. 6.3.2–6.3.4 are asymptotically exact on some special grids.

As an example, we consider the estimator $\mathcal{R}_{N,K}$ defined in (6.30). The triangulation K_h is called *parallel* in a subdomain $\Omega_1 \subset \Omega$ if Ω_e is a parallelogram for each edge $e \in \Omega_1$ (cf. Fig. 6.10). Then, under the conditions

$$\begin{aligned} &K_h \text{ is parallel in } \Omega_1 \ , \\ &p|_{\Omega_1} \in H^3(\Omega_1) \ , \\ &\|\nabla(p - p_h)\|_{\mathbf{L}^2(\Omega_1)} \geq Ch \ , \\ &\|p - p_h\|_{L^2(\Omega_1)} \leq Ch^{1+\epsilon_1} \text{ for some } \epsilon_1 > 0 \ , \end{aligned}$$

it can be proven (cf. Exercise 6.11) that the error estimator $\mathcal{R}_{N,K}$ is asymptotically exact in the subdomain Ω_1:

$$\left\{ \sum_{K\subset\Omega_1, K\in K_h} \mathcal{R}^2_{N,K} \right\}^{1/2} \Bigg/ \|\nabla(p - p_h)\|_{\mathbf{L}^2(\Omega_1)} \to 1 \text{ as } h \to 0 \ .$$

Example 6.8. Consider problem (6.1) with $\Omega = (0,1)\times(0,1)$, $\Gamma_N = (0,1)\times\{0\} \cup (0,1)\times\{1\}$, $g = 0$, and $f = 1$. The analytical solution is $p(x_1, x_2) = x_1(x_1 - 1)/2$. The partition K_h is constructed as follows: First, Ω is cut into m^2 squares of length $h = 1/m$, and each square is then divided into four triangles by drawing the diagonals (cf. Fig. 6.12). This triangulation is often termed a *criss-cross grid*, and is not parallel. Because the solution p is quadratic and the homogeneous Neumann boundary condition is used, the approximate solution p_h to (6.5) is determined by (cf. Exercise 6.12)

$$p_h(\mathbf{m}) = \begin{cases} p(\mathbf{m}) & \text{if } \mathbf{m} \text{ is a vertex of a square} \ , \\ p(\mathbf{m}) - \dfrac{h^2}{24} & \text{if } \mathbf{m} \text{ is the centroid of a square} \ . \end{cases}$$

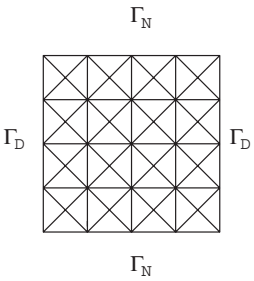

Fig. 6.12. A criss-cross grid

With this, we can explicitly compute the error $\|\nabla(p - p_h)\|_{\mathbf{L}^2(Q)}$ and the error estimator defined in (6.30): For any square Q disjoint with Γ_N, it can be checked that

$$\left\{ \sum_{K \subset Q, K \in K_h} \mathcal{R}_{N,K}^2 \right\}^{1/2} / \|\nabla(p - p_h)\|_{\mathbf{L}^2(Q)} = \sqrt{\frac{17}{6}} \, .$$

This example shows that the asymptotical exactness may not hold on all grids even if they are strongly structured.

6.4 A-Posteriori Error Estimates for Transient Problems

In this section, we briefly discuss the adaptive finite element method for a transient problem in a bounded domain $\Omega \subset \mathbb{R}^2$:

$$\begin{aligned} &\frac{\partial p}{\partial t} - \Delta p = f && \text{in } \Omega \times J \, , \\ &p = 0 && \text{on } \Gamma \times J \, , \\ &p(\cdot, 0) = p_0 && \text{in } \Omega \, , \end{aligned} \tag{6.59}$$

where $J = (0, T]$ $(T > 0)$ and f and p_0 are given functions. In the numerical computation of many transient problems (e.g., those with a moving front), it is necessary to change a grid dynamically during a solution process to maintain reasonable accuracy (cf. Sect. 1.7.1). Such a method is usually called a moving grid method, i.e., a r-scheme as described in Sect. 6.1. As discussed there, the pure r-scheme can suffer from several deficiencies. In particular, the r-scheme couples the determination of the solution with the grid selection, which can significantly increase the size of a discrete system, especially in

multidimensional cases when a separate grid is utilized for each component of the solution vector.

To overcome this difficulty, a number of different methods have been developed such as the finite element method of lines for transient problems (Bieterman-Babuška, 1982; Adjerid et al., 1993). In this section, we briefly review an adaptive finite element method for (6.59) developed by Eriksson-Johnson (1991). This method is an extension to transient problems of the adaptive algorithm presented in Sect. 6.3.1 for stationary problems.

Let $0 = t^0 < t^1 < \ldots < t^N = T$ be a partition of J into subintervals $J^n = (t^{n-1}, t^n)$, with length $\Delta t^n = t^n - t^{n-1}$. For a generic function v of time, set $v^n = v(t^n)$.

For each n $(n = 0, 1, \ldots, N)$, let (h_n, K_{h_n}, V_{h_n}) be a finite element triple where $h_n = h_n(\mathbf{x})$ is the grid function, K_{h_n} is a regular triangulation of Ω into triangles, and $V_{h_n} \subset V$ is the finite element space of piecewise linear functions associated with K_{h_n} at time level n. Assume that h_n satisfies (6.3) and (6.4) with C_1 and C_2 independent of n. Now, the finite element method for (6.59) is defined: Find $p_h^n \in V_{h_n}$, $n = 1, 2, \ldots, N$, such that

$$\begin{aligned} &\left(\frac{p_h^n - p_h^{n-1}}{\Delta t^n}, v\right) + a\,(p_h^n, v) = (f^n, v) \qquad \forall v \in V_{h_n}\ , \\ &\left(p_h^0, v\right) = (p_0, v) \qquad\qquad\qquad\qquad\quad \forall v \in V_{h_0}\ , \end{aligned} \tag{6.60}$$

where a backward Euler scheme is used in the discretization of the time differentiation term in (6.59) (cf. Sect. 1.7). Note that (6.60) resembles (1.80) in form. The difference between them is that in (6.60) the space and time steps can vary in time, and the space steps can vary in space. Observe that the time steps in (6.60) are kept constant in space. Obviously, optimal computational grid design requires that the time steps vary in space as well. While there are adaptive computational codes available that are established on grids variable in both space and time, there is no theory for them. Thus we concentrate on the discretization (6.60) in this section.

Let $\mathcal{E}_{h_n}^o$ denote the set of all interior edges e of K_{h_n}, and for each n $(n = 1, 2, \ldots, N)$, define

$$L^n = \max_{1 \le k \le n} \left(\log\left(\frac{t^k}{\Delta t^k}\right) + 1\right)^{1/2} .$$

Now, introduce the error estimator

$$\begin{aligned} \mathcal{R}(p_h^n) = CL^n \Bigg\{ &\sum_{K \in K_{h_n}} \left(\left\| h_n^2 f_K^n \right\|_{L^2(K)}^2 + \left\| h_n^2 (f^n - f_K^n) \right\|_{L^2(K)}^2 \right) \\ &+ \sum_{e \in \mathcal{E}_{h_n}^o} \left\| h_n^{3/2} [\nabla p_h \cdot \boldsymbol{\nu}] \right\|_{L^2(e)}^2 \\ &+ \left\| p_h^n - p_h^{n-1} \right\|^2 + \left(\left\| \frac{(h_n)^2}{\Delta t^n} (p_h^n - p_h^{n-1}) \right\|^* \right)^2 \Bigg\}^{1/2} , \end{aligned}$$

where the constant C depends only on C_1 and C_2, the superscript $*$ indicates that the corresponding term is present only if $V_{h_{n-1}} \not\subset V_{h_n}$, and $\|\cdot\|$ is the L^2-norm. Since the norm $\|\cdot\|$ is used in the definition of the estimator $\mathcal{R}(p_h^n)$ for the transient problem (6.59), f_h^n is now the L^2-projection of f^n into V_{h_n} (not into the space of piecewise constants).

Let p and p_h be the respective solutions of (6.59) and (6.60). Then we have the following a-posteriori estimate (Eriksson-Johnson, 1991): for each n, $n = 1, 2, \ldots, N$,

$$\|p(t^n) - p_h^n\| \leq \max_{1 \leq k \leq n} \mathcal{R}(p_h^k) \, . \tag{6.61}$$

As in Sect. 6.3.1, an adaptive algorithm can be designed based on (6.61). For a given tolerance $\epsilon > 0$, we seek (h_n, K_{h_n}, V_{h_n}) and Δt^n, $n = 1, 2, \ldots, N$, such that

$$\|p(t^n) - p_h^n\| \leq \epsilon \, , \tag{6.62}$$

and the number of degrees of freedom is minimal. It follows from (6.61) that if the following estimate holds, so does (6.62):

$$\mathcal{R}(p_h^n) \leq \epsilon, \qquad n = 1, 2, \ldots, N \, . \tag{6.63}$$

We now define an adaptive algorithm (termed Algorithm II) for choosing (h_n, K_{h_n}, V_{h_n}) and Δt^n, $n = 1, 2, \ldots, N$, as follows:

1. Choose an initial space grid $K_{h_{n,0}}$, with grid size $h_{n,0}$ and an initial time step $\Delta t^{n,0}$;
2. Determine grids $K_{h_{n,k+1}}$, with $M_{h_{n,k+1}}$ elements of size $h_{n,k+1}$, time steps $\Delta t^{n,k+1}$, and the corresponding solutions $p_h^{n,k+1}$ such that, for $k = 0, 1, \ldots, \hat{n} - 1$ and $K \in K_{h_{n,k}}$,

$$\begin{aligned} CL^{n,k} \Bigg\{ & \left\| h_{n,k+1}^2 f_K^n \right\|_{L^2(K)} + \left\| h_{n,k+1}^2 (f^n - f_K^n) \right\|_{L^2(K)} \\ & + \left(\frac{1}{2} \sum_{e \in \partial K} \left\| h_{n,k+1}^{3/2} R_e(p_h^{n,k}) \right\|_{L^2(e)}^2 \right)^{1/2} \\ & + \left\| \frac{h_{n,k+1}^2}{\Delta t^{n,k}} (p_h^{n,k} - p_h^{n-1}) \right\|_{L^2(K)}^* \Bigg\} = \frac{\epsilon}{2\sqrt{M_{h_{n,k}}}} \, , \end{aligned}$$

$$\Delta t^{n,k+1} C L^{n,k} \left\| p_h^{n,k} - p_h^{n-1} \right\| = \frac{\epsilon \Delta t^{n,k}}{2} \, ;$$

3. Define $K_{h_n} = K_{h_{n,\hat{n}}}$ with grid size $h_n = h_{n,\hat{n}}$, and time step $\Delta t^n = \Delta t^{n,\hat{n}}$.

For each n, the number of "trials" $\hat{n}$ is the smallest integer such that for $k = \hat{n}$, (6.63) holds with p_h^n replaced by $p_h^{n,\hat{n}}$. In applications, we can choose $K_{h_{n,0}} = K_{h_{n-1}}$, $n = 2, 3, \ldots, N - 1$, and it is usually the case that $\hat{n} = 1$.

To see the efficiency of the adaptive algorithm based on (6.61), we can establish a lower bound for $\mathcal{R}(p_h^n)$, as in the previous section. For example, in the case $f = 0$, it can be shown (Eriksson-Johnson, 1991) that there is a constant C such that

$$\mathcal{R}(p_h^n) \le C\,(L^n)^2 \max_{0\le k\le n} \left(\max_{t\in J^k} \|h_k^2 D^2 p(t)\| + \Delta t^k \max_{t\in J^k} \left\| \frac{\partial p}{\partial t} \right\| \right) ,$$

for $n = 1, 2, \ldots, N$, provided $\max_{\mathbf{x}\in\bar{\Omega}} h_n(\mathbf{x}) \le C_1\sqrt{\Delta t^n}$ whenever $V_{h_{n-1}} \not\subset V_{h_n}$. Recall that

$$D^2 p = \left(\sum_{|\alpha|=2} \left| \frac{\partial^{|\alpha|} p}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2}} \right|^2 \right)^{1/2} ,$$

with $\alpha = (\alpha_1, \alpha_2)$ and $|\alpha| = \alpha_1 + \alpha_2$ (cf. Sect. 1.2).

The result (6.61) carries over to finite element spaces of polynomials of degree $r \ge 2$. In this case, f_h^n is the L^2-projection of f^n into the space of piecewise polynomials of degree r with respect to K_{h_n}, and f_K^n in the first term of $\mathcal{R}(p_h^n)$ is replaced by $\Delta p_h^n|_K + f_K^n$.

6.5 A-Posteriori Error Estimates for Nonlinear Problems

We present an application of the adaptive finite element method to the nonlinear transient problem

$$\begin{aligned} & c(p)\frac{\partial p}{\partial t} - \nabla\cdot\big(a(p)\nabla p\big) = f(p) && \text{in } \Omega\times J\,, \\ & p = 0 && \text{on } \Gamma\times J\,, \\ & p(\cdot,0) = p_0 && \text{in } \Omega\,, \end{aligned} \tag{6.64}$$

where $c(p) = c(\mathbf{x}, t, p)$, $a(p) = a(\mathbf{x}, t, p)$, $f(p) = f(\mathbf{x}, t, p)$, and $\Omega \subset \mathbb{R}^2$. This problem has been studied in the preceding five chapters. Here we very briefly describe an application of the adaptive method (6.60) to it. We assume that (6.64) admits a unique solution.

With $V = H_0^1(\Omega)$, problem (6.64) is recast in the variational form: Find $p : J \to V$ such that

$$\begin{aligned} & \left(c(p)\frac{\partial p}{\partial t}, v \right) + \big(a(p)\nabla p, \nabla v\big) = \big(f(p), v\big) && \forall v\in V,\ t\in J\,, \\ & p(\mathbf{x},0) = p_0(\mathbf{x}) && \forall \mathbf{x}\in\Omega\,. \end{aligned} \tag{6.65}$$

Different solution approaches (e.g., the linearization, implicit time approximation, and explicit time approximation) have been developed in Sect. 1.8

for (6.65). As an example, we describe the implicit time approximation approach. With the same notation as in the previous section, the adaptive finite element method for (6.64) is defined: Find $p_h^n \in V_{h_n}$, $n = 1, 2, \ldots, N$, such that

$$\left(c\left(p_h^n\right) \frac{p_h^n - p_h^{n-1}}{\Delta t^n}, v \right) + \left(a\left(p_h^n\right) \nabla p_h^n, \nabla v \right) = \left(f\left(p_h^n\right), v \right) \qquad \forall v \in V_{h_n} , \tag{6.66}$$

$$\left(p_h^0, v\right) = (p_0, v) \qquad \forall v \in V_{h_0} .$$

As in (6.60), the space and time steps in (6.66) can vary in time, the space steps can vary in space, and the time steps are kept constant in space. With appropriate assumptions on the coefficients c, a, and f, it can be seen that (6.66) admits a solution for all t; see Sect. 1.8.

An a-posteriori estimate for the general form (6.66) is not available in the literature. However, for a simplified version of (6.64), such an estimate was shown by Eriksson-Johnson (1995). In particular, we assume that $c(p) = 1$ and $a : \mathbb{R} \to [1, a^*]$. For each n, we define

$$R(p_h^n)(\mathbf{x}) = h_K^{-1} \max_{\mathbf{y} \in \partial K} \left| a\left(p_h^n\right) \left[\nabla p_h^n \cdot \boldsymbol{\nu}\right](\mathbf{y}) \right| + \left| \frac{da}{dp}\left(p_h^n(\mathbf{x})\right) \right| \left| \nabla p_h^n(\mathbf{x}) \right|^2 , \qquad \mathbf{x} \in K,\ K \in K_{h_n} .$$

Also, we introduce the estimator, $n = 1, 2, \ldots, N$,

$$\mathcal{R}(p_h^n) = CL^n \left\{ \left\| h_n^2 R(p_h^n) \right\| + \left\| h_n^2 \left(f(p_h^n) - P_{h_n} f(p_h^n) \right) \right\| + \left\| p_h^n - p_h^{n-1} \right\| + \left\| \frac{(h_n)^2}{\Delta t^n} (p_h^n - p_h^{n-1}) \right\|^* \right\} ,$$

where the superscript $*$ indicates that the corresponding term is present only if $V_{h_{n-1}} \not\subset V_{h_n}$ and $P_{h_n} : L^2(\Omega) \to V_{h_n}$ is the L^2-projection. With all these choices, it can be proven that (6.61) remains valid; i.e.,

$$\| p(t^n) - p_h^n \| \le \max_{1 \le k \le n} \mathcal{R}(p_h^k) , \tag{6.67}$$

where p and p_h are the solutions to (6.65) and (6.66), respectively. Based on (6.67), an adaptive algorithm similar to Algorithm II can be designed for the nonlinear problem (6.64).

6.6 Theoretical Considerations

As in the preceding chapters, we present an abstract theory for a-posteriori error estimators. The theory follows Verfürth (1996).

6.6.1 An Abstract Theory

Let $\mathcal{X}$ and $\mathcal{Y}$ be two Banach spaces, with norms $\|\cdot\|_{\mathcal{X}}$ and $\|\cdot\|_{\mathcal{Y}}$, respectively. Denote by $\mathcal{L}(\mathcal{X},\mathcal{Y})$ and $\mathrm{Hom}(\mathcal{X},\mathcal{Y})$ the spaces of continuous operators of $\mathcal{X}$ into $\mathcal{Y}$ and of linear homeomorphisms of $\mathcal{X}$ onto $\mathcal{Y}$, respectively. $\|\cdot\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})}$ represents the norm in $\mathcal{L}(\mathcal{X},\mathcal{Y})$. The dual space of $\mathcal{X}$ is $\mathcal{X}' = \mathcal{L}(\mathcal{X},\mathbb{R})$, and the corresponding duality pairing is $\langle\cdot,\cdot\rangle_{\mathcal{X}'\times\mathcal{X}}$ (cf. Sect. 1.2.5). The adjoint of a linear operator $\mathcal{F} \in \mathcal{L}(\mathcal{X},\mathcal{Y})$ is denoted $\mathcal{F}' \in \mathcal{L}(\mathcal{Y}',\mathcal{X}')$. Finally, for a given real number R and $v \in \mathcal{X}$, we define the ball centered at v:

$$B(v,R) = \{w \in \mathcal{X} : \|w - v\|_{\mathcal{X}} < R\} .$$

Let $F \in C^1(\mathcal{X},\mathcal{Y}')$ be a given continuously *Fréchet differentiable* mapping, and consider an equation of the form

$$F(p) = 0 . \tag{6.68}$$

Below the notation DF will indicate the *Fréchet derivative* of F. We recall that $F \in C^1(\mathcal{X},\mathcal{Y}')$ is Fréchet differentiable at $p_0 \in \mathcal{X}$ if there is a linear operator $DF \in \mathcal{L}(\mathcal{X},\mathcal{Y}')$ such that in a neighborhood of p_0,

$$\|F(p) - F(p_0) - DF(p - p_0)\|_{\mathcal{L}(\mathcal{X},\mathcal{Y}')} = o\left(\|p - p_0\|_{\mathcal{X}}\right) ,$$

and $DF(p_0)$ is called the Fréchet derivative of F at p_0.

The next theorem provides a-posteriori error estimates for the elements in a neighborhood of a solution to (6.68).

Theorem 6.1. *Assume that $p_0 \in \mathcal{X}$ is a regular solution of* (6.68)*, i.e., $DF(p_0) \in \mathrm{Hom}(\mathcal{X},\mathcal{Y}')$, and that DF is Lipschitz continuous at p_0, i.e., there exists $R_0 > 0$ satisfying*

$$\gamma \equiv \sup_{p \in B(p_0,R_0)} \frac{\|DF(p) - DF(p_0)\|_{\mathcal{L}(\mathcal{X},\mathcal{Y}')}}{\|p - p_0\|_{\mathcal{X}}} < \infty .$$

Then, with

$$R = \min\left\{R_0, \gamma^{-1}\|DF^{-1}(p_0)\|^{-1}_{\mathcal{L}(\mathcal{Y}',\mathcal{X})}, 2\gamma^{-1}\|DF(p_0)\|_{\mathcal{L}(\mathcal{X},\mathcal{Y}')}\right\} ,$$

we have the error estimate, for all $p \in B(p_0,R)$,

$$\begin{aligned} \frac{1}{2}\|DF(p_0)\|^{-1}_{\mathcal{L}(\mathcal{X},\mathcal{Y}')}\|F(p)\|_{\mathcal{Y}'} &\le \|p - p_0\|_{\mathcal{X}} \\ &\le 2\|DF^{-1}(p_0)\|_{\mathcal{L}(\mathcal{Y}',\mathcal{X})}\|F(p)\|_{\mathcal{Y}'} . \end{aligned} \tag{6.69}$$

Proof. For $p \in B(p_0, R)$, we see that

$$p - p_0 = DF^{-1}(p_0)\Bigg(F(p) + \int_0^1 \left[DF(p_0) - DF(p_0 + \ell(p - p_0)) \right](p - p_0)\, d\ell \Bigg) ,$$

which implies

$$\begin{aligned} \|p - p_0\|_{\mathcal{X}} &\le \|DF^{-1}(p_0)\|_{\mathcal{L}(\mathcal{Y}',\mathcal{X})} \Bigg(\|F(p)\|_{\mathcal{Y}'} \\ &\quad + \int_0^1 \|DF(p_0) - DF(p_0 + \ell(p - p_0))\|_{\mathcal{L}(\mathcal{X},\mathcal{Y}')} \|p - p_0\|_{\mathcal{X}}\, d\ell \Bigg) \\ &\le \|DF^{-1}(p_0)\|_{\mathcal{L}(\mathcal{Y}',\mathcal{X})} \left(\|F(p)\|_{\mathcal{Y}'} + \frac{\gamma}{2} \|p - p_0\|_{\mathcal{X}}^2 \right) \\ &\le \|DF^{-1}(p_0)\|_{\mathcal{L}(\mathcal{Y}',\mathcal{X})} \|F(p)\|_{\mathcal{Y}'} + \frac{1}{2} \|p - p_0\|_{\mathcal{X}} . \end{aligned}$$

Consequently, the right-hand inequality in (6.69) follows. Also, for all $v \in \mathcal{Y}$ with $\|v\|_{\mathcal{Y}} = 1$, we have

$$\begin{aligned} \langle F(p), v \rangle_{\mathcal{Y}' \times \mathcal{Y}} &= \langle DF(p_0)(p - p_0), v \rangle_{\mathcal{Y}' \times \mathcal{Y}} \\ &\quad + \left\langle \int_0^1 \left[DF(p_0 + \ell(p - p_0)) - DF(p_0) \right](p - p_0)\, d\ell, v \right\rangle_{\mathcal{Y}' \times \mathcal{Y}} , \end{aligned} \tag{6.70}$$

which yields

$$\begin{aligned} \|F(p)\|_{\mathcal{Y}'} &\le \|DF(p_0)\|_{\mathcal{L}(\mathcal{X},\mathcal{Y}')} \|p - p_0\|_{\mathcal{X}} \\ &\quad + \int_0^1 \|DF(p_0 + \ell(p - p_0)) - DF(p_0)\|_{\mathcal{L}(\mathcal{X},\mathcal{Y}')} \|p - p_0\|_{\mathcal{X}}\, d\ell \\ &\le \|DF(p_0)\|_{\mathcal{L}(\mathcal{X},\mathcal{Y}')} \|p - p_0\|_{\mathcal{X}} + \frac{\gamma}{2} \|p - p_0\|_{\mathcal{X}}^2 \\ &\le 2 \|DF(p_0)\|_{\mathcal{L}(\mathcal{X},\mathcal{Y}')} \|p - p_0\|_{\mathcal{X}} . \end{aligned}$$

Thus the left-hand inequality in (6.69) also follows. □

The assumptions on F can be weakened. For example, we can assume that $F \in C(\mathcal{X}, \mathcal{Y}')$, $F(p_0) = 0$, and there are $R > 0$ and two monotonically increasing homeomorphisms σ_1, σ_2 of $[0, \infty)$ onto itself such that

$$\begin{aligned} \sigma_1 \left(\|p - p_0\|_{\mathcal{X}} \right) &\le \|F(p)\|_{\mathcal{Y}'} \\ &\le \sigma_2 \left(\|p - p_0\|_{\mathcal{X}} \right) \qquad \forall p \in B(p_0, R) . \end{aligned} \tag{6.71}$$

This inequality then implies

$$\sigma_2^{-1} \left(\|F(p)\|_{\mathcal{Y}'} \right) \le \|p - p_0\|_{\mathcal{X}} \le \sigma_1^{-1} \left(\|F(p)\|_{\mathcal{Y}'} \right) \qquad \forall p \in B(p_0, R) .$$

The left-hand inequality in (6.71) is true if F is strongly monotone in a neighborhood of p_0, and the right-hand one is satisfied if F is Hölder continuous at p_0, for example.

Let $\mathcal{X}_h \subset \mathcal{X}$ and $\mathcal{Y}_h \subset \mathcal{Y}$ be two finite-dimensional subspaces and $F_h \in C(\mathcal{X}_h, \mathcal{Y}_h')$ be an approximation of F. We now estimate $\|F(p_h)\|_{\mathcal{Y}'}$, where $p_h \in \mathcal{X}_h$ is an approximate solution of the problem

$$F_h(p_{0h}) = 0\,; \tag{6.72}$$

i.e., $\|F_h(p_h)\|_{\mathcal{Y}'}$ is "small".

Theorem 6.2. *Assume that $p_h \in \mathcal{X}_h$ is an approximate solution of* (6.72) *and that there are a restriction operator $Q_h \in \mathcal{L}(\mathcal{Y}, \mathcal{Y}_h)$, a finite-dimensional subspace $\tilde{\mathcal{Y}}_h \subset \mathcal{Y}$, and an approximation $\tilde{F}_h : \mathcal{X}_h \to \mathcal{Y}'$ of F at p_h such that*

$$\|(I - Q_h)'\tilde{F}_h(p_h)\|_{\mathcal{Y}'} \le C\|\tilde{F}_h(p_h)\|_{\tilde{\mathcal{Y}}_h'}, \tag{6.73}$$

where I is the identity operator. Then the following estimate holds:

$$\begin{aligned}\|F(p_h)\|_{\mathcal{Y}'} &\le C\|\tilde{F}_h(p_h)\|_{\tilde{\mathcal{Y}}_h'}\\ &\quad +\|(I - Q_h)'[F(p_h) - \tilde{F}_h(p_h)]\|_{\mathcal{Y}'}\\ &\quad +\|Q_h\|_{\mathcal{L}(\mathcal{Y},\mathcal{Y}_h)}\left(\|F(p_h) - F_h(p_h)\|_{\mathcal{Y}_h'} + \|F_h(p_h)\|_{\mathcal{Y}_h'}\right).\end{aligned} \tag{6.74}$$

Proof. Note that, for $v \in \mathcal{Y}$ with $\|v\|_{\mathcal{Y}} = 1$,

$$\begin{aligned}&\langle F(p_h), v\rangle_{\mathcal{Y}'\times\mathcal{Y}}\\ &\quad= \left\langle \tilde{F}_h(p_h), v - Q_h v\right\rangle_{\mathcal{Y}'\times\mathcal{Y}} + \left\langle F(p_h) - \tilde{F}_h(p_h), v - Q_h v\right\rangle_{\mathcal{Y}'\times\mathcal{Y}}\\ &\qquad + \langle F(p_h) - F_h(p_h), Q_h v\rangle_{\mathcal{Y}_h'\times\mathcal{Y}_h} + \langle F_h(p_h), Q_h v\rangle_{\mathcal{Y}_h'\times\mathcal{Y}_h}\\ &\quad\le \|(I - Q_h)'\tilde{F}_h(p_h)\|_{\mathcal{Y}'} + \left\|(I - Q_h)'\left[F(p_h) - \tilde{F}_h(p_h)\right]\right\|_{\mathcal{Y}'}\\ &\qquad +\|Q_h\|_{\mathcal{L}(\mathcal{Y},\mathcal{Y}_h)}\left(\|F(p_h) - F_h(p_h)\|_{\mathcal{Y}_h'} + \|F_h(p_h)\|_{\mathcal{Y}_h'}\right),\end{aligned}$$

which, together with (6.73), implies the desired result. □

When this theorem is applied to the examples presented in Sect. 6.3, $\mathcal{X}_h$ and $\mathcal{Y}_h$ are appropriate finite element spaces, the choice of Q_h is natural, and $\tilde{F}_h(p_h)$ is obtained by projecting $F(p_h)$ elementwise onto suitable finite element spaces. The construction of $\tilde{\mathcal{Y}}_h$ satisfying (6.73) is not so easy. The second term in the right-hand side of (6.74) measures the quality of the approximation of $\tilde{F}_h(p_h)$ to $F(p_h)$. The quantity $\|F(p_h) - F_h(p_h)\|_{\mathcal{Y}_h'}$ is a consistency error of the discretization. Finally, $\|F_h(p_h)\|_{\mathcal{Y}_h'}$ measures the residual of (6.72) and must be computed separately.

6.6.2 Applications

We now present an application of the theory to a-posteriori error estimators for the second-order problem (6.1). As an example, we focus on the residual estimator in Sect. 6.3.1.

Set

$$\mathcal{X} = \mathcal{Y} = V = \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_D\} ,$$

with norms $\|\cdot\|_{\mathcal{X}} = \|\cdot\|_{\mathcal{Y}} = \|\cdot\|_{H^1(\Omega)}$, and

$$\langle F(p), v\rangle_{\mathcal{Y}'\times\mathcal{Y}} = (\nabla p, \nabla v) - (f, v) - (g, v)_{\Gamma_N}, \qquad v \in \mathcal{Y} .$$

The solution $p \in \mathcal{X}$ satisfies (6.68) if and only if it is the solution of (6.2). Because the bilinear form $a(\cdot,\cdot) : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ is continuous and coercive, we see that $DF(v) \in \mathrm{Hom}(\mathcal{X}, \mathcal{Y}')$ for all $v \in \mathcal{X}$.

Next, set

$$\begin{aligned} &\mathcal{X}_h = \mathcal{Y}_h = \{v \in V : v|_K \in P_1(K),\ K \in K_h\} , \\ &\langle F_h(p_h), v\rangle_{\mathcal{Y}_h'\times\mathcal{Y}_h} = \langle F(p_h), v\rangle_{\mathcal{Y}'\times\mathcal{Y}} , \qquad p_h,\ v \in \mathcal{X}_h . \end{aligned}$$

It is clear that $p_h \in \mathcal{X}_h$ satisfies (6.72) if and only if it is the solution of (6.5).

Define $Q_h \in \mathcal{L}(\mathcal{Y}, \mathcal{Y}_h)$ as the interpolation operator of Clément (1975). Given $v \in \mathcal{Y}$ and $\mathbf{m} \in \mathcal{N}_h$, let $\pi_{\mathbf{m}} v$ be the L^2-projection of v into $P_1(\Omega_{\mathbf{m}})$; i.e.,

$$(\pi_{\mathbf{m}} v, w)_{\Omega_{\mathbf{m}}} = (v, w)_{\Omega_{\mathbf{m}}} \qquad \forall w \in P_1(\Omega_{\mathbf{m}}) .$$

Then define $Q_h v$ by

$$\begin{aligned} &(Q_h v)(\mathbf{m}) = (\pi_h v)(\mathbf{m}), && \mathbf{m} \in \mathcal{N}_h^o \cup \mathcal{N}_h^N , \\ &(Q_h v)(\mathbf{m}) = 0, && \mathbf{m} \in \mathcal{N}_h^D . \end{aligned}$$

This operator satisfies the following local error estimates:

$$\begin{aligned} &\|v - Q_h v\|_{L^2(K)} \le C h_K \|v\|_{H^1(\tilde{\Omega}_K)}, && K \in K_h , \\ &\|v - Q_h v\|_{L^2(e)} \le C h_e^{1/2} \|v\|_{H^1(\tilde{\Omega}_e)}, && e \in \mathcal{E}_h , \end{aligned} \tag{6.75}$$

where (cf. Fig. 6.13)

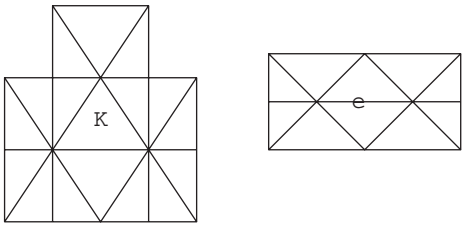

Fig. 6.13. An illustration of $\tilde{\Omega}_K$ (*left*) and $\tilde{\Omega}_e$ (*right*)

$$\tilde{\Omega}_K = \bigcup \{K' \in K_h : \bar{K} \cap \bar{K}' \neq \emptyset\} \ ,$$
$$\tilde{\Omega}_e = \bigcup \{K' \in K_h : \bar{e} \cap \bar{K}' \neq \emptyset\} \ ,$$

and the constants C depend only on C_2 in (6.4).

Also, define $\tilde{F}_h$ and $\tilde{\mathcal{Y}}_h$ by

$$\begin{aligned} \left\langle \tilde{F}_h(v), w \right\rangle_{\mathcal{Y}' \times \mathcal{Y}} &= (\nabla v, \nabla w) - \sum_{K \in K_h} (f_K, v)_K \\ &\quad - \sum_{e \in \mathcal{E}_h^N} (g_e, v)_e \, d\ell, \qquad v, \ w \in \mathcal{Y} \ , \\ \tilde{\mathcal{Y}}_h &= \operatorname{span} \{ b_K, \ K \in K_h; \ b_e, \ e \in \mathcal{E}_h^o \cup \mathcal{E}_h^N \} \ , \end{aligned}$$

where f_K and g_e are defined as in (6.13).

We are now ready to use Theorems 6.1 and 6.2 to prove residual a-posteriori error estimates stated in Sect. 6.3.1.

Theorem 6.3. *Let p and p_h be the respective solutions to* (6.2) *and* (6.5)*, and $\mathcal{R}_K$ be defined as in* (6.14)*, $K \in K_h$. Then there are constants C, depending only on C_2, such that*

$$\begin{aligned} \|p - p_h\|_{H^1(\Omega)} \leq C \Big\{ & \sum_{K \in K_h} \left(\mathcal{R}_K^2 + h_K^2 \|f - f_K\|_{L^2(K)}^2 \right) \\ & + \sum_{e \in \mathcal{E}_h^N} h_e \|g - g_e\|_{L^2(e)}^2 \Big\}^{1/2} , \end{aligned} \tag{6.76}$$

and

$$\begin{aligned} \mathcal{R}_K \leq C \Big\{ & \sum_{K' \in \Omega_K} \left(\|p - p_h\|_{H^1(K')}^2 + h_{K'}^2 \|f - f_{K'}\|_{L^2(K')}^2 \right) \\ & + \sum_{e \in \partial K \cap \mathcal{E}_h^N} h_e \|g - g_e\|_{L^2(e)}^2 \Big\}^{1/2} . \end{aligned} \tag{6.77}$$

Proof. It follows from (6.11) that, with $v \in \mathcal{Y}$,

$$\begin{aligned} \langle F(p_h), v \rangle_{\mathcal{Y}' \times \mathcal{Y}} = \ & -(f, v) - \sum_{e \in \mathcal{E}_h^N} (g - \nabla p_h \cdot \boldsymbol{\nu}, v)_e \\ & + \sum_{e \in \mathcal{E}_h^o} ([\![\nabla p_h \cdot \boldsymbol{\nu}]\!], v)_e \, , \end{aligned} \tag{6.78}$$

and that, with f and g replaced elementwise by f_K and g_e,

$$\begin{aligned} \left\langle \tilde{F}_h(p_h), v \right\rangle_{\mathcal{Y}' \times \mathcal{Y}} = & - \sum_{K \in K_h} (f_K, v)_K - \sum_{e \in \mathcal{E}_h^N} (g_e - \nabla p_h \cdot \boldsymbol{\nu}, v)_e \\ & + \sum_{e \in \mathcal{E}_h^o} ([\![\nabla p_h \cdot \boldsymbol{\nu}]\!], v)_e \ . \end{aligned} \tag{6.79}$$

Using (6.2) and (6.75), we see that

$$\begin{aligned}
&\Big\|(I-Q_h)'\Big[F(p_h)-\tilde{F}_h(p_h)\Big]\Big\|_{\mathcal{Y}'}\\
&\quad=\sup_{v\in\mathcal{Y},\ \|v\|_{\mathcal{Y}}=1}\Big\{\sum_{K\in K_h}(f_K-f,v-Q_hv)_K+\sum_{e\in\mathcal{E}_h^N}(g_e-g,v-Q_hv)_e\Big\}\\
&\le C\sup_{v\in\mathcal{Y},\ \|v\|_{\mathcal{Y}}=1}\Big\{\sum_{K\in K_h}h_K\|f_K-f\|_{L^2(K)}\|v\|_{H^1(\tilde{\Omega}_K)}\\
&\qquad\qquad+\sum_{e\in\mathcal{E}_h^N}h_e^{1/2}\|g_e-g\|_{L^2(e)}\|v\|_{H^1(\tilde{\Omega}_e)}\Big\}\\
&\le C\Big\{\sum_{K\in K_h}h_K^2\|f_K-f\|^2_{L^2(K)}+\sum_{e\in\mathcal{E}_h^N}h_e\|g_e-g\|^2_{L^2(e)}\Big\}^{1/2},
\end{aligned}\tag{6.80}$$

and

$$\begin{aligned}
&\Big\|F(p_h)-\tilde{F}_h(p_h)\Big\|_{\tilde{\mathcal{Y}}_h'}\\
&\quad=\sup_{v\in\tilde{\mathcal{Y}}_h,\ \|v\|_{\mathcal{Y}}=1}\Big\{\sum_{K\in K_h}(f_K-f,v)_K+\sum_{e\in\mathcal{E}_h^N}(g_e-g,v)_e\Big\}\\
&\le C\sup_{v\in\tilde{\mathcal{Y}}_h,\ \|v\|_{\mathcal{Y}}=1}\Big\{\sum_{K\in K_h}h_K\|f_K-f\|_{L^2(K)}\|v\|_{H^1(K)}\\
&\qquad\qquad+\sum_{e\in\mathcal{E}_h^N}h_e^{1/2}\|g_e-g\|_{L^2(e)}\|v\|_{H^1(\Omega_e)}\Big\}\\
&\le C\Big\{\sum_{K\in K_h}h_K^2\|f_K-f\|^2_{L^2(K)}+\sum_{e\in\mathcal{E}_h^N}h_e\|g_e-g\|^2_{L^2(e)}\Big\}^{1/2}.
\end{aligned}\tag{6.81}$$

Similarly, we have

$$\begin{aligned}
&\Big\|(I-Q_h)'\tilde{F}_h(p_h)\Big\|_{\mathcal{Y}'}\\
&\quad=\sup_{v\in\mathcal{Y},\ \|v\|_{\mathcal{Y}}=1}\Big\{\big(\nabla p_h,\nabla(v-Q_hv)\big)\\
&\qquad\qquad-\sum_{K\in K_h}(f_K,v-Q_hv)_K-\sum_{e\in\mathcal{E}_h^N}(g_e,v-Q_hv)_e\Big\}
\end{aligned}\tag{6.82}$$

$$\leq C\Bigg\{ \sum_{K\in K_h} h_K^2 \|f_K\|^2_{L^2(K)} + \sum_{e\in\mathcal{E}_h^o} h_e \|[\nabla p_h\cdot\boldsymbol{\nu}]\|^2_{L^2(e)} + \sum_{e\in\mathcal{E}_h^N} h_e \|g_e - \nabla p_h\cdot\boldsymbol{\nu}\|^2_{L^2(e)} \Bigg\}^{1/2}$$

$$\leq C\Bigg\{ \sum_{K\in K_h} \mathcal{R}_K^2 \Bigg\}^{1/2} .$$

Next, it follows from (6.50), (6.51), and (6.79) that

$$\begin{aligned}
&\Big\|\tilde{F}_h(p_h)\Big\|_{\tilde{\mathcal{Y}}_h'} \\
&\quad = \sup_{v\in\tilde{\mathcal{Y}}_h,\ \|v\|_{\mathcal{Y}}=1} \Bigg\{ \sum_{K\in K_h} (f_K, v)_K + \sum_{e\in\mathcal{E}_h^N} (g_e - \nabla p_h\cdot\boldsymbol{\nu}, v)_e \\
&\qquad\qquad - \sum_{e\in\mathcal{E}_h^o} ([\nabla p_h\cdot\boldsymbol{\nu}], v)_e \Bigg\} \\
&\leq C \sup_{v\in\tilde{\mathcal{Y}}_h,\ \|v\|_{\mathcal{Y}}=1} \Bigg\{ \sum_{K\in K_h} h_K \|f_K\|_{L^2(K)} \|v\|_{H^1(K)} \\
&\qquad\qquad + \sum_{e\in\mathcal{E}_h^N} h_e^{1/2} \|g_e - \nabla p_h\cdot\boldsymbol{\nu}\|_{L^2(e)} \|v\|_{H^1(\Omega_e)} \\
&\qquad\qquad + \sum_{e\in\mathcal{E}_h^o} h_e^{1/2} \|[\nabla p_h\cdot\boldsymbol{\nu}]\|_{L^2(e)} \|v\|_{H^1(\Omega_e)} \Bigg\} \\
&\leq C\Bigg\{ \sum_{K\in K_h} \mathcal{R}_K^2 \Bigg\}^{1/2} .
\end{aligned} \tag{6.83}$$

Thus, to prove (6.73), it suffices to show that

$$\Bigg\{ \sum_{K\in K_h} \mathcal{R}_K^2 \Bigg\}^{1/2} \leq \Big\|\tilde{F}_h(p_h)\Big\|_{\tilde{\mathcal{Y}}_h'} . \tag{6.84}$$

Toward that end, for each $K \in K_h$, set $w_K = f_K b_K$. Then, by (6.50) and (6.79), we observe that

$$\begin{aligned}
\frac{9}{20}\|f_K\|^2_{L^2(K)} &= (f_K, w_K)_K \\
&= -\Big\langle \tilde{F}_h(p_h), w_K \Big\rangle_{\mathcal{Y}'\times\mathcal{Y}} \\
&\leq \|w_K\|_{H^1(K)} \sup_{v\in\tilde{\mathcal{Y}}_h|_K,\ \|v\|_{\mathcal{Y}}=1} \Big\langle \tilde{F}_h(p_h), v \Big\rangle_{\mathcal{Y}'\times\mathcal{Y}} \\
&\leq C h_K^{-1} \|f_K\|_{L^2(K)} \sup_{v\in\tilde{\mathcal{Y}}_h|_K,\ \|v\|_{\mathcal{Y}}=1} \Big\langle \tilde{F}_h(p_h), v \Big\rangle_{\mathcal{Y}'\times\mathcal{Y}} ,
\end{aligned}$$

so that

$$h_K \|f_K\|_{L^2(K)} \le C \sup_{v \in \tilde{\mathcal{Y}}_h|_K,\ \|v\|_{\mathcal{Y}}=1} \left\langle \tilde{F}_h(p_h), v \right\rangle_{\mathcal{Y}' \times \mathcal{Y}} , \tag{6.85}$$

where $\tilde{\mathcal{Y}}_h|_K$ represents the set of functions $v \in \tilde{\mathcal{Y}}_h$ such that $\mathrm{supp}(v) \subset K$.

Also, for each $e \in \mathcal{R}_h^o$, let $w_e = [\nabla p_h \cdot \boldsymbol{\nu}]_e b_e$. Applying (6.51) and (6.79), we see that

$$\begin{aligned}
\frac{2}{3}\|[\nabla p_h \cdot \boldsymbol{\nu}]\|^2_{L^2(e)} &= ([\nabla p_h \cdot \boldsymbol{\nu}[, w_e)_e \\
&= \left\langle \tilde{F}_h(p_h), w_e \right\rangle_{\mathcal{Y}' \times \mathcal{Y}} + \sum_{K \in \Omega_e} (f_K, w_e)_K \\
&\le \|w_e\|_{H^1(\Omega_e)} \sup_{v \in \tilde{\mathcal{Y}}_h|_{\Omega_e},\ \|v\|_{\mathcal{Y}}=1} \left\langle \tilde{F}_h(p_h), v \right\rangle_{\mathcal{Y}' \times \mathcal{Y}} \\
&\qquad + \sum_{K \in \Omega_e} \|f_K\|_{L^2(K)} \|w_e\|_{L^2(K)} \\
&\le C\|[\nabla p_h \cdot \boldsymbol{\nu}]\|_{L^2(e)} \left\{ h_e^{-1/2} \sup_{v \in \tilde{\mathcal{Y}}_h|_{\Omega_e},\ \|v\|_{\mathcal{Y}}=1} \left\langle \tilde{F}_h(p_h), v \right\rangle_{\mathcal{Y}' \times \mathcal{Y}} \right. \\
&\qquad \left. + \sum_{K \in \Omega_e} h_e^{1/2} \|f_K\|_{L^2(K)} \right\} ,
\end{aligned}$$

so that, using (6.85),

$$h_e^{1/2} \|[\nabla p_h \cdot \nu]\|_{L^2(e)} \le C \sup_{v \in \tilde{\mathcal{Y}}_h|_{\Omega_e},\ \|v\|_{\mathcal{Y}}=1} \left\langle \tilde{F}_h(p_h), v \right\rangle_{\mathcal{Y}' \times \mathcal{Y}} . \tag{6.86}$$

Applying the same argument with $w_e = (g_e - \nabla p_h \cdot \boldsymbol{\nu}) b_e$, $e \in \mathcal{E}_h^N$, we can also prove that

$$h_e^{1/2} \|g_e - \nabla p_h \cdot \boldsymbol{\nu}\|_{L^2(e)} \le C \sup_{v \in \tilde{\mathcal{Y}}_h|_{\Omega_e},\ \|v\|_{\mathcal{Y}}=1} \left\langle \tilde{F}_h(p_h), v \right\rangle_{\mathcal{Y}' \times \mathcal{Y}} . \tag{6.87}$$

Combining inequalities (6.85)–(6.87), we obtain

$$\mathcal{R}_K \le C \sup_{v \in \tilde{\mathcal{Y}}_h|_{\Omega_K},\ \|v\|_{\mathcal{Y}}=1} \left\langle \tilde{F}_h(p_h), v \right\rangle_{\mathcal{Y}' \times \mathcal{Y}} , \qquad K \in K_h . \tag{6.88}$$

Summing over $K \in K_h$ yields (6.84). Note that, for all v, $w \in \mathcal{Y}$ with $\mathrm{supp}(w) \subset \Omega_K$,

$$(\nabla v, \nabla w) \le \|v\|_{H^1(\Omega_K)} \|w\|_{H^1(\Omega_K)} .$$

Using this inequality, Theorem 6.3 finally follows from Theorems 6.1 and 6.2 and inequalities (6.80)–(6.84) and (6.88). □

While we have only analyzed linear problems, the theory presented in Sect. 6.6.1 is quite general. It can be applied to quasilinear equations of second-order elliptic type and Navier-Stokes equations (Verfürth, 1996), for example.

6.7 Bibliographical Remarks

The theory and application of the adaptive finite element method has undergone significant advances in the last two decades. Significant advances in error estimation, data structure development, and adaptive strategies have been made for stationary (elliptic) and transient (parabolic) problems and for a certain type of hyperbolic problems. In this chapter, we have briefly reviewed the basic ideas behind the adaptive finite element method for the first two types of problems. In fact, over the last two decades, much of the interest has been concentrated on these two types of problems, and relatively little progress has been made on hyperbolic and advection-dominated problems. As discussed in the preceding two chapters, for the latter type of problems special precautions are needed. In this book, we have considered the discontinuous and characteristic finite element methods for solving them numerically; see Chaps. 4 and 5. Research into adaptive discontinuous and characteristic methods is needed. For adaptive techniques for other numerical methods (e.g., finite difference) for hyperbolic or advection-dominated problems, the reader can refer to Berger-Oliger (1984), for example. We mention that adaptive refinement techniques for the standard finite element method can be extended to the nonconforming and mixed finite element methods in Chaps. 2 and 3 (see, e.g., Ewing-Wang, 1992; Hoppe-Wohlmuth, 1997; Chen-Ewing, 2003). The discontinuous finite element method discussed in Chap. 4 is local; namely, functions used in the finite element spaces of this method are discontinuous across interelement boundaries. The adaptive method should take advantage of the locality of discontinuous finite elements. Again, much research is needed.

The content of Sects. 6.1 and 6.2, of Sects. 6.3 and 6.6, and of Sects. 6.4 and 6.5 is based on Oden-Demkowicz (1988), Verfürth (1996), and Eriksson-Johnson (1991), respectively. The residual estimator in Sect. 6.3.1 was first proposed and analyzed by Babuška-Rheinboldt (1978A,B). The estimators $\mathcal{R}_{D,\mathbf{m}}$, $\mathcal{R}_{D,K}$, and $\mathcal{R}_{N,K}$ in Sect. 6.3.2 were developed by Babuška-Rheinboldt (1978A), Bernardi et al. (1993), and Bank-Weiser (1985), respectively. The averaging-based estimator in Sect. 6.3.3 was introduced by Zienkiewicz-Zhu (1987). For hierarchical basis estimators in Sect. 6.3.4, see Deuflhard et al. (1989), Bank-Smith (1993), and Verfürth (1996). Finally, there are several books on the adaptive finite element method, e.g., Verfürth (1996), Ainsworth-Oden (2000), and Bangerth-Rannacher (2003).

6.8 Exercises

6.1. For the example in Fig. 6.2, use the Refinement Rule defined in Sect. 6.1.1 to convert irregular vertices to regular vertices.

6.2. For the problem

$$
\begin{aligned}
-\nabla \cdot (\mathbf{a}\nabla p) &= f && \text{in } \Omega \ , \\
p &= 0 && \text{on } \Gamma_D \ , \\
\mathbf{a}\nabla p \cdot \boldsymbol{\nu} &= g_N && \text{on } \Gamma_N \ ,
\end{aligned}
$$

derive an inequality similar to (6.12).

6.3. Apply (6.13) and (6.14) to derive (6.15) from (6.12).

6.4. For problem (6.1), extend the result (6.15) to the case $r \geq 2$, where r is the polynomial degree of the finite element space V_h.

6.5. For the problem

$$
\begin{aligned}
-\nabla \cdot (\mathbf{a}\nabla p) &= f && \text{in } \Omega \ , \\
p &= 0 && \text{on } \Gamma_D \ , \\
\mathbf{a}\nabla p \cdot \boldsymbol{\nu} &= g_N && \text{on } \Gamma_N \ ,
\end{aligned}
$$

define an error estimator and a discrete problem similar to (6.21) and (6.22), respectively.

6.6. Use Figs. 6.7 and 6.9 to show that the respective dimensions of the discrete problems (6.22), (6.27), and (6.31) are 12, 7, and 4.

6.7. Prove inequality (6.38) and equality (6.39).

6.8. Show that the bubble functions b_K and b_e satisfy (6.50) and (6.51), respectively.

6.9. Let B_h be given in Example 6.4 and $\mathcal{N}_h'$ be the set of nodes corresponding to B_h, i.e., $\mathcal{N}_h' = \mathcal{N}_{h/2}$. Define

$$
S_h = \{v \in B_h : v(\mathbf{x}) = 0 \quad \forall \mathbf{x} \in \mathcal{N}_h\} \ ,
$$

and

$$
b(v,w) = \sum_{\mathbf{x} \in \mathcal{N}_h' \setminus \mathcal{N}_h} v(\mathbf{x}) w(\mathbf{x}), \qquad v,\ w \in S_h \ .
$$

Show that (6.53) holds for the bilinear form $b(\cdot,\cdot)$.

6.10. Show that the error estimator in Sect. 6.3.1 is efficient according to the definition given in Sect. 6.3.5.

6.11. Consider the estimator $\mathcal{R}_{N,K}$ defined in (6.30). Show that if

$$
\begin{aligned}
&\text{the triangulation } K_h \text{ is parallel in } \Omega_1 \subset \Omega, \\
&p|_{\Omega_1} \in H^3(\Omega_1), \\
&\|\nabla(p - p_h)\|_{\mathbf{L}^2(\Omega_1)} \geq Ch, \text{ and} \\
&\|p - p_h\|_{L^2(\Omega_1)} \leq Ch^{1+\epsilon_1} \text{ for some } \epsilon_1 > 0 \ ,
\end{aligned}
$$

then the error estimator $\mathcal{R}_{N,K}$ is asymptotically exact in the subdomain Ω_1:

$$
\left\{ \sum_{K \subset \Omega_1, K \in K_h} \mathcal{R}_{N,K}^2 \right\}^{1/2} \Big/ \|\nabla(p - p_h)\|_{\mathbf{L}^2(\Omega_1)} \to 1 \text{ as } h \to 0 \ .
$$

6.12. Consider problem (6.1) with $\Omega = (0,1) \times (0,1)$, $\Gamma_N = (0,1) \times \{0\} \cup (0,1) \times \{1\}$, $g = 0$, and $f = 1$. The analytical solution is $p(x_1, x_2) = x_1(x_1 - 1)/2$. The partition K_h is constructed as follows: First, Ω is cut into m^2 squares of length $h = 1/m$, and each square is then divided into four triangles by drawing the diagonals (cf. Fig. 6.12). Show that the approximate solution p_h to (6.5) is determined by

$$p_h(\mathbf{m}) = \begin{cases} p(\mathbf{m}) & \text{if } \mathbf{m} \text{ is a vertex of a square ,} \\ p(\mathbf{m}) - \dfrac{h^2}{24} & \text{if } \mathbf{m} \text{ is the centroid of a square.} \end{cases}$$

Also, prove that for any square Q disjoint with Γ_N,

$$\left\{ \sum_{K \subset Q, K \in K_h} \mathcal{R}_{N,K}^2 \right\}^{1/2} \Big/ \|\nabla(p - p_h)\|_{\mathbf{L}^2(Q)} = \sqrt{\frac{17}{6}} \,,$$

where $\mathcal{R}_{N,K}$ is defined in (6.30).

6.13. Referring to Exercise 6.2, extend the analysis in Sect. 6.6.2 to the problem

$$\begin{aligned} -\nabla \cdot (\mathbf{a} \nabla p) &= f && \text{in } \Omega \,, \\ p &= 0 && \text{on } \Gamma_D \,, \\ \mathbf{a} \nabla p \cdot \boldsymbol{\nu} &= g_N && \text{on } \Gamma_N \,. \end{aligned}$$

7 Solid Mechanics

7.1 Introduction

Finite element methods have been widely employed to solve problems in *solid mechanics*. An important problem in this area involves calculating *deformations* and *stresses* of *elastic bodies* subject to loads. Elasticity theory has three ingredients: the *kinematics*, *equilibrium*, and *material* laws.

So far, we have solely considered the application of finite element methods to single partial differential equations. This and next three chapters present an application of these methods to a system of partial differential equations.

7.1.1 Kinematics

Suppose that the domain Ω is the reference configuration of a body under consideration. Initially, the body is in a natural state, which can be described by a mapping

$$\mathbf{R} : \Omega \to \mathbb{R}^3 \, .$$

We write this mapping in the form

$$\mathbf{R}(\mathbf{x}) = \mathbf{ID}(\mathbf{x}) + \mathbf{u}(\mathbf{x}), \qquad \mathbf{x} \in \Omega \, , \tag{7.1}$$

where $\mathbf{ID}$ is the identity function and $\mathbf{u}$ is the *displacement*. We often deal with the case where the displacement is small.

Define the gradient tensor

$$\nabla \mathbf{R} = \begin{pmatrix} \dfrac{\partial R_1}{\partial x_1} & \dfrac{\partial R_1}{\partial x_2} & \dfrac{\partial R_1}{\partial x_3} \\ \dfrac{\partial R_2}{\partial x_1} & \dfrac{\partial R_2}{\partial x_2} & \dfrac{\partial R_2}{\partial x_3} \\ \dfrac{\partial R_3}{\partial x_1} & \dfrac{\partial R_3}{\partial x_2} & \dfrac{\partial R_3}{\partial x_3} \end{pmatrix} \, ,$$

with $\mathbf{R} = (R_1, R_2, R_3)$. If the determinant of $\nabla \mathbf{R}$ is positive, i.e.,

$$\det(\nabla \mathbf{R}) > 0 \,,$$

the mapping $\mathbf{R}$ represents a *deformation*. The matrix

$$\mathbf{C} = \nabla \mathbf{R}^T \nabla \mathbf{R} \tag{7.2}$$

represents a transformation of the body and is termed the *Cauchy-Green strain tensor*. Its deviation from the identity is referred to as the *strain*:

$$\mathbf{E} = \frac{1}{2}(\mathbf{C} - \mathbf{I}) \ . \tag{7.3}$$

It follows from (7.1)–(7.3) that

$$E_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) + \frac{1}{2}\sum_{k=1}^{3} \frac{\partial u_i}{\partial x_k}\frac{\partial u_j}{\partial x_k}, \qquad i,\ j = 1,2,3 \,.$$

In *linear elasticity* the quadratic terms are assumed small and neglected, and the components of the resulting strain are denoted by

$$\epsilon_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right), \qquad i,\ j = 1,2,3 \,. \tag{7.4}$$

This quantity (*strain*) is one of the most important quantities in elasticity theory.

7.1.2 Equilibrium

Two types of forces are applied to a body: the surface force and the body force. A typical body force is gravitation, and a force applied by a load on the surface is a surface force.

We denote the body force by $\mathbf{f} : \Omega \to \mathbb{R}^3$, which acts on a volume. A major axiom in solid mechanics is that, in the equilibrium state of a body, all forces and moments add to zero. One of the implications of this axiom is that there is a symmetric tensor $\boldsymbol{\sigma}$, called the *stress tensor*, such that (Ciarlet, 1988)

$$\nabla \cdot \boldsymbol{\sigma} + \mathbf{f} = \mathbf{0}, \qquad \mathbf{x} \in \Omega \,. \tag{7.5}$$

7.1.3 Material Laws

The equilibrium relation (7.5) comprises three equations, and they are not sufficient to determine the six components of the symmetric stress tensor. The other three necessary equations arise from constitutive relationships, i.e., *material laws*. An important task in solid mechanics is to find these laws, which show how the deformation of a body depends on material properties and applied forces.

A state of deformation in which all the strain components are constant throughout the body is called a *homogeneous deformation*. On the other hand, if the properties of the body are identical in all directions, the material is termed *isotropic*. In the case where the displacement is small and the material is isotropic, the relationship between the stress and deformation (i.e., the *linear Hooke's law*) is

$$\boldsymbol{\sigma} = 2\mu\boldsymbol{\epsilon} + \lambda\nabla \cdot \mathbf{u}\,\mathbf{I}\,, \tag{7.6}$$

where μ and λ are the Lamé constants. λ describes the stress due to the change in density, and μ is the *shear modulus of the material*.

Equation (7.6) can be also written in terms of the *modulus of elasticity* (*Young modulus*) E and the *Poisson ratio* (*contraction ratio*) ν. These constants are related to μ and λ via

$$E = \frac{\mu(3\lambda + 2\mu)}{\lambda + \mu}, \qquad \nu = \frac{\lambda}{2(\lambda + \mu)}\,,$$
$$\lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}, \quad \mu = \frac{E}{2(1+\nu)}\,.$$

Equations (7.4)–(7.6) constitute the basic laws for a homogeneous, isotropic, elastic body. They, together with appropriate boundary conditions, are used to determine the displacement $\mathbf{u}$, the stress $\boldsymbol{\sigma}$, and the strain $\boldsymbol{\epsilon}$.

Let the boundary Γ be decomposed into two parts Γ_D and Γ_N, where Γ_D is fixed and a surface force $\mathbf{g}$ is applied on Γ_N (cf. Fig. 7.1). That is,

$$\begin{aligned} \mathbf{u} &= \mathbf{0} \qquad && \text{on } \Gamma_D\,, \\ \boldsymbol{\sigma} \cdot \boldsymbol{\nu} &= \mathbf{g} \qquad && \text{on } \Gamma_N, \end{aligned} \tag{7.7}$$

where $\boldsymbol{\nu}$ is the outward unit normal to Γ_N. If $\Gamma_D = \emptyset$ (respectively, $\Gamma_N = \emptyset$), the boundary value problem is called a *pure traction* (respectively, *pure displacement*) problem.

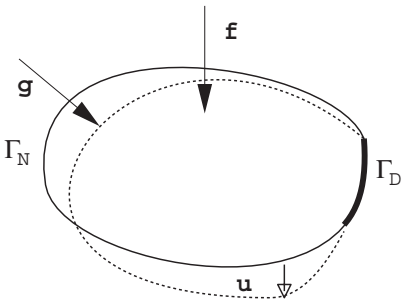

Fig. 7.1. An illustration of an elastic body Ω

In this chapter, we discuss various finite element methods for solving equations (7.4)–(7.7). In particular, the methods developed in Chaps. 1–3 are considered. In Sect. 7.2, we state variational formulations of these equations. Then, in Sect. 7.3, we describe the conforming, mixed, and nonconforming finite element methods. Sect. 7.4 is devoted to theoretical considerations. Finally, in Sect. 7.5, we give bibliographical information.

7.2 Variational Formulations

7.2.1 The Displacement Form

We substitute (7.6) into (7.5) to obtain the *Lamé differential equation*

$$-2\mu\nabla\cdot\boldsymbol{\epsilon}(\mathbf{u})-\lambda\nabla\nabla\cdot\mathbf{u}=\mathbf{f}\qquad\text{in }\Omega\,. \tag{7.8}$$

Define the space (cf. Sect. 1.2)

$$\mathbf{V}=\{\mathbf{v}\in(H^1(\Omega))^3:\mathbf{v}=\mathbf{0}\text{ on }\Gamma_D\}\,.$$

Applying Green's formula (1.19), (7.7) and (7.8) can be rewritten in the variational formulation (cf. Exercise 7.1): Find $\mathbf{u}\in\mathbf{V}$ such that

$$\begin{aligned}&2\mu\,(\boldsymbol{\epsilon}(\mathbf{u}),\boldsymbol{\epsilon}(\mathbf{v}))+\lambda\,(\nabla\cdot\mathbf{u},\nabla\cdot\mathbf{v})\\&\qquad=(\mathbf{f},\mathbf{v})+(\mathbf{g},\mathbf{v})_{\Gamma_N}\,,\qquad\mathbf{v}\in\mathbf{V}\,,\end{aligned} \tag{7.9}$$

where

$$(\boldsymbol{\epsilon}(\mathbf{u}),\boldsymbol{\epsilon}(\mathbf{v}))=\int_\Omega\boldsymbol{\epsilon}(\mathbf{u}):\boldsymbol{\epsilon}(\mathbf{v})\;d\mathbf{x}\,,$$

and the tensor product is defined by

$$\boldsymbol{\epsilon}(\mathbf{u}):\boldsymbol{\epsilon}(\mathbf{v})=\sum_{i,j=1}^{3}\epsilon_{ij}(\mathbf{u})\epsilon_{ij}(\mathbf{v})\,.$$

If the surface Γ_D has a positive area, system (7.9) can be shown to have a unique solution (cf. Exercise 7.2). In the case of the pure traction problem, $\mathbf{V}$ is simply $\left(H^1(\Omega)\right)^3$, and (7.9) is solvable under a compatibility condition between $\mathbf{f}$ and $\mathbf{g}$ (Brenner-Scott, 1994). In two dimensions, for example, for (7.9) to have a solution in the pure traction case, the necessary and sufficient condition is

$$(\mathbf{f},\mathbf{v})+(\mathbf{g},\mathbf{v})_\Gamma=0,\qquad\mathbf{v}\in\hat{\mathbf{V}}\,,$$

where

$$\hat{\mathbf{V}}=\{\mathbf{v}=\mathbf{b}+c(x_2,-x_1):\mathbf{b}\in\mathbb{R}^2,\;c\in\mathbb{R}\}\,.$$

7.2.2 The Mixed Form

Equations (7.4)–(7.7) can be also written in a mixed formulation. For this, define

$$\mathbf{V} = \left\{ \boldsymbol{\tau} \in (\mathbf{H}(\mathrm{div}, \Omega))^3 : \boldsymbol{\tau} \cdot \boldsymbol{\nu} = 0 \text{ on } \Gamma_N \right\}, \quad \mathbf{W} = \left(L^2(\Omega)\right)^3 ,$$

where we recall that

$$\mathbf{H}(\mathrm{div}, \Omega) = \left\{ \mathbf{v} \in (L^2(\Omega))^3 : \nabla \cdot \mathbf{v} \in L^2(\Omega) \right\} .$$

Next, using (7.6), we see that (cf. Exercise 7.3)

$$\begin{pmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{33} \\ \sigma_{12} \\ \sigma_{13} \\ \sigma_{23} \end{pmatrix} = \mathbf{B} \begin{pmatrix} \epsilon_{11} \\ \epsilon_{22} \\ \epsilon_{33} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{23} \end{pmatrix} , \tag{7.10}$$

where

$$\mathbf{B} = \begin{pmatrix} \lambda + 2\mu & \lambda & \lambda & 0 & 0 & 0 \\ \lambda & \lambda + 2\mu & \lambda & 0 & 0 & 0 \\ \lambda & \lambda & \lambda + 2\mu & 0 & 0 & 0 \\ 0 & 0 & 0 & 2\mu & 0 & 0 \\ 0 & 0 & 0 & 0 & 2\mu & 0 \\ 0 & 0 & 0 & 0 & 0 & 2\mu \end{pmatrix} .$$

Then (7.4)–(7.7) are written in an equivalent mixed formulation (cf. Exercise 7.4): Find $\boldsymbol{\sigma} \in (\mathbf{H}(\mathrm{div}, \Omega))^3$ and $\mathbf{u} \in \mathbf{W}$, with $\boldsymbol{\sigma} \cdot \boldsymbol{\nu} = \mathbf{g}$ on Γ_N, such that

$$\begin{aligned} \left(\mathbf{B}^{-1}\boldsymbol{\sigma}, \boldsymbol{\tau}\right) + (\nabla \cdot \boldsymbol{\tau}, \mathbf{u}) &= 0, \qquad && \boldsymbol{\tau} \in \mathbf{V} , \\ (\nabla \cdot \boldsymbol{\sigma}, \mathbf{v}) &= -(\mathbf{f}, \mathbf{v}), && \mathbf{v} \in \mathbf{W} . \end{aligned} \tag{7.11}$$

The analysis of this mixed formulation is similar to that for second-order differential problems given in Chap. 3.

The displacement formulation is simpler than the mixed one. However, in real applications, one is mostly interested in calculating directly the stress with more accuracy. The mixed method is more desirable in this aspect. There is also a mixed method that involves all three variables $\mathbf{u}$, $\boldsymbol{\sigma}$, and $\boldsymbol{\epsilon}$, i.e., the *Hu-Washizu* mixed method. However, from a computational perspective, this three variable mixed method is more complex than the two variable mixed method, so we do not consider it.

For the pure displacement problem, we must modify the definition of the space $\mathbf{V}$ in (7.11) by

$$\mathbf{V} = \left\{ \boldsymbol{\tau} \in (\mathbf{H}(\mathrm{div}, \Omega))^3 : \int_\Omega \mathrm{tr}(\boldsymbol{\tau}) \; d\mathbf{x} = 0 \right\} , \tag{7.12}$$

where the trace of a tensor $\boldsymbol{\tau}$ is defined by

$$\mathrm{tr}(\boldsymbol{\tau}) = \tau_{11} + \tau_{22} + \tau_{33} \; .$$

This can be seen as follows: From (7.4) it follows that

$$\mathrm{tr}(\boldsymbol{\epsilon}) = \nabla \cdot \mathbf{u} \; ;$$

consequently, by (7.6) and (1.17), we have

$$\begin{aligned} \int_\Omega \mathrm{tr}(\boldsymbol{\sigma}) \; d\mathbf{x} &= (2\mu + 3\lambda) \int_\Omega \mathrm{tr}(\boldsymbol{\epsilon}) \; d\mathbf{x} = (2\mu + 3\lambda) \int_\Omega \nabla \cdot \mathbf{u} \; d\mathbf{x} \\ &= (2\mu + 3\lambda) \int_\Gamma \mathbf{u} \cdot \boldsymbol{\nu} \; d\ell = 0 \; . \end{aligned}$$

7.3 Finite Element Methods

7.3.1 Finite Elements and Locking Effects

For simplicity, let Ω be a convex polygonal domain, and let K_h be a regular triangulation of Ω into tetrahedra as in Chap. 1. With $\mathbf{V}$ being defined as in Sect. 7.2.1, we define

$$\mathbf{V}_h = \{ \mathbf{v} \in \mathbf{V} : \mathbf{v}|_K \in (P_1(K))^3, \; K \in K_h \} \; .$$

Now, the finite element method in the displacement formulation can be stated as follows: Find $\mathbf{u}_h \in \mathbf{V}_h$ such that

$$\begin{aligned} &2\mu \, (\boldsymbol{\epsilon}(\mathbf{u}_h), \boldsymbol{\epsilon}(\mathbf{v})) + \lambda \, (\nabla \cdot \mathbf{u}_h, \nabla \cdot \mathbf{v}) \\ &\qquad = (\mathbf{f}, \mathbf{v}) + (\mathbf{g}, \mathbf{v})_{\Gamma_N} \; , \qquad \mathbf{v} \in \mathbf{V}_h \; . \end{aligned} \tag{7.13}$$

It can be shown that this discrete problem has a unique solution if Γ_D has a positive area (cf. Exercise 7.5). For the pure traction problem, $\mathbf{V}_h$ is a subspace of $\left(H^1(\Omega)\right)^3$, and $\mathbf{f}$ and $\mathbf{g}$ must satisfy a compatibility condition, as noted in Sect. 7.2.1. It can be proven that the following error estimate holds (cf. Sect. 1.9):

$$\|\mathbf{u} - \mathbf{u}_h\|_{\mathbf{H}^1(\Omega)} \leq C(\mu, \lambda) h \|\mathbf{u}\|_{\mathbf{H}^2(\Omega)} \; . \tag{7.14}$$

For a fixed pair of (μ, λ), estimate (7.14) gives a satisfactory convergence result. But the convergence of the finite element solution to the exact solution is not uniform in λ as $h \to 0$. In particular, the performance of the finite element method deteriorates as $\lambda \to \infty$. This phenomenon is known as *locking* (*Poisson locking* or *volume locking*). There are several approaches to reducing the effects of locking such as the mixed and nonconforming finite element methods, which will be discussed in the next two subsections.

7.3.2 Mixed Finite Elements

In general, it is difficult to find a stable pair of mixed finite element spaces for elasticity problems. For this reason, in this section, we concentrate on two dimensions. Even in solving three-dimensional problems, it is often possible to work in two (or even one) dimensions because the length of a body in one of the directions is much shorter than that in other directions. Some typical examples are *bars*, *membranes*, *beams*, *plates*, and *shells*. Here we consider a membrane problem.

Let $\Omega \subset \mathbb{R}^2$ be a planar domain, and the elasticity body have the form $\Omega \times (-l, l)$, where l is a real number. We assume that only external forces are exerted on this body and that their x_3-components vanish. The reduction in dimension cannot be achieved simply by eliminating the x_3-components in the stress or strain. Here we consider a case where boundary conditions are enforced at the ends $x_3 = \pm l$. These boundary conditions imply that the x_3-component of the displacement is zero. Thus we have the *plane strain state*, i.e.,

$$\begin{aligned} &\epsilon_{ij}(x_1, x_2, x_3) = \epsilon_{ij}(x_1, x_2), \quad && i,\ j = 1, 2\ , \\ &\epsilon_{3j} = \epsilon_{j3}, && j = 1, 2, 3\ . \end{aligned}$$

The corresponding displacements become

$$u_i(x_1, x_2, x_3) = u_i(x_1, x_2), \quad i = 1, 2, \quad u_3 = 0\ .$$

Then, by (7.6), we see that

$$\sigma_{33} = \frac{\lambda}{2(\lambda + \mu)}(\sigma_{11} + \sigma_{22})\ ,$$

so that σ_{33} can be eliminated. Also, (7.10) reduces to

$$\begin{pmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{12} \end{pmatrix} = \mathbf{B} \begin{pmatrix} \epsilon_{11} \\ \epsilon_{22} \\ \epsilon_{12} \end{pmatrix}\ , \tag{7.15}$$

with

$$\mathbf{B} = \begin{pmatrix} \lambda + 2\mu & \lambda & 0 \\ \lambda & \lambda + 2\mu & 0 \\ 0 & 0 & 2\mu \end{pmatrix}\ .$$

There are not many satisfactory mixed finite elements available for solving elasticity problems. In this section, we consider the PEERS (plane elasticity element with reduced symmetry) mixed finite element introduced by Arnold et al. (1984A). In this element, tensors are not required to be symmetric. Toward that end, we define the *antisymmetric operator*

$$\mathrm{as}(\boldsymbol{\tau}) = \tau_{12} - \tau_{21}, \qquad \boldsymbol{\tau} \in \left(L^2(\Omega)\right)^{2\times 2}\ .$$

For simplicity, we introduce the PEERS element for the pure displacement boundary problem.

The mixed formulation is defined as follows: Find $(\boldsymbol{\sigma}, \mathbf{u}, \gamma) \in (\mathbf{H}(\mathrm{div}, \Omega))^2 \times (L^2(\Omega))^2 \times L^2(\Omega)$ such that

$$\begin{aligned} &\left(\mathbf{B}^{-1}\boldsymbol{\sigma}, \boldsymbol{\tau}\right) + (\nabla \cdot \boldsymbol{\tau}, \mathbf{u}) + (\mathrm{as}(\boldsymbol{\tau}), \gamma) = 0, && \boldsymbol{\tau} \in (\mathbf{H}(\mathrm{div}, \Omega))^2 \ , \\ &(\nabla \cdot \boldsymbol{\sigma}, \mathbf{v}) = -(\mathbf{f}, \mathbf{v}), && \mathbf{v} \in (L^2(\Omega))^2 \ , \\ &(\mathrm{as}(\boldsymbol{\sigma}), \eta) = 0, && \eta \in L^2(\Omega) \ . \end{aligned} \tag{7.16}$$

Problem (7.16) has a unique solution (Arnold et al., 1984A). In fact, it can be seen that (7.16) is equivalent to the two-dimensional counterpart of (7.11) with

$$\gamma = \frac{1}{2}\left(\frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2}\right) \ .$$

If $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$ are symmetric, system (7.16) becomes (7.11) immediately. The last equation in system (7.16) is used to enforce symmetry of $\boldsymbol{\sigma}$.

Let K_h be a triangulation of Ω into triangles as in Chap. 1. Define the *cubic bubble functions* (cf. Sect. 6.3.2.1)

$$B_h = \left\{v \in H^1(\Omega) : v|_K \in P_3(K),\ K \in K_h \right. \\ \left. \text{and } v|_e = 0 \text{ on all edges in } K_h\right\} \ .$$

Let $\mathbf{V}_h \times W_h$ be the lowest-order Raviart-Thomas mixed element on triangles (cf. Sect. 3.4); i.e.,

$$\begin{aligned} \mathbf{V}_h &= \left\{\mathbf{v} \in \mathbf{H}(\mathrm{div}, \Omega) : \mathbf{v}\big|_K = (b_K x_1 + a_K, b_K x_2 + c_K),\ K \in K_h\right\} \ , \\ W_h &= \left\{w \in L^2(\Omega) : w|_K \in P_0(K),\ K \in K_h\right\} \ . \end{aligned}$$

Also, define the continuous finite element space

$$\Lambda_h = \left\{v \in H^1(\Omega) : v|_K \in P_1(K),\ K \in K_h\right\} \ ,$$

and the augmented space

$$\mathbf{X}_h = \mathbf{V}_h \oplus \mathbf{rot}(B_h) \ ,$$

where

$$\mathbf{rot}(v) = \left(\frac{\partial v}{\partial x_2}, -\frac{\partial v}{\partial x_1}\right) \ .$$

Now, the mixed finite element method is: Find $(\boldsymbol{\sigma}_h, \mathbf{u}_h, \gamma_h) \in (\mathbf{X}_h)^2 \times (W_h)^2 \times \Lambda_h$ such that

$$\begin{aligned} &\left(\mathbf{B}^{-1}\boldsymbol{\sigma}_h, \boldsymbol{\tau}\right) + (\nabla \cdot \boldsymbol{\tau}, \mathbf{u}_h) + (\mathrm{as}(\boldsymbol{\tau}), \gamma_h) = 0, && \boldsymbol{\tau} \in (\mathbf{X}_h)^2 \ , \\ &(\nabla \cdot \boldsymbol{\sigma}_h, \mathbf{v}) = -(\mathbf{f}, \mathbf{v}), && \mathbf{v} \in (W_h)^2 \ , \\ &(\mathrm{as}(\boldsymbol{\sigma}_h), \eta) = 0, && \eta \in \Lambda_h \ . \end{aligned} \tag{7.17}$$

Again, this system has a unique solution, and the following error estimate holds (Arnold et al., 1984A):

$$\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\mathbf{L}^2(\Omega)} + \|\mathbf{u} - \mathbf{u}_h\|_{\mathbf{L}^2(\Omega)} + \|\gamma - \gamma_h\|_{L^2(\Omega)} \leq Ch\|\mathbf{f}\|_{\mathbf{L}^2(\Omega)} ,$$

where the constant C depends on the positive upper and lower bounds of $\mu \in [\mu_1, \mu_2]$ $(0 < \mu_1 < \mu_2)$, but is independent of $\lambda \in [0, \infty)$. This implies that method (7.17) is locking-free.

7.3.3 Nonconforming Finite Elements

We consider the nonconforming finite element method for the pure displacement boundary problem. For this problem, we define

$$\mathbf{V} = \{\mathbf{v} \in (H^1(\Omega))^2 : \mathbf{v} = \mathbf{0} \text{ on } \Gamma\} .$$

With this space, formulation (7.13) becomes: Find $\mathbf{u} \in \mathbf{V}$ such that

$$\mu\,(\nabla\mathbf{u}, \nabla\mathbf{v}) + (\mu + \lambda)\,(\nabla \cdot \mathbf{u}, \nabla \cdot \mathbf{v}) = (\mathbf{f}, \mathbf{v}) , \qquad \mathbf{v} \in \mathbf{V} . \tag{7.18}$$

Equation (7.18) has a unique solution (cf. Sect. 7.4).

7.3.3.1 Nonconforming Finite Elements on Triangles

For a convex polygonal domain Ω, let K_h be a regular triangulation of Ω into triangles as in Chap. 1. Define the finite element space on triangles (cf. Sect. 2.1.1)

$$\begin{aligned}\mathbf{V}_h = \{\mathbf{v} \in \left(L^2(\Omega)\right)^2 : \mathbf{v}|_K \text{ is linear, } K \in K_h; \mathbf{v} \text{ is continuous} \\ \text{at the midpoints of interior edges and} \\ \text{is zero at the midpoints of edges on } \Gamma\} .\end{aligned}$$

Then the nonconforming finite element method in the displacement formulation is: Find $\mathbf{u}_h \in \mathbf{V}_h$ such that

$$\begin{aligned}\sum_{K \in K_h} \{\mu\,(\nabla\mathbf{u}_h, \nabla\mathbf{v})_K + (\mu + \lambda)\,(\nabla \cdot \mathbf{u}_h, \nabla \cdot \mathbf{v})_K\} \\ = (\mathbf{f}, \mathbf{v}) , \qquad \mathbf{v} \in \mathbf{V}_h .\end{aligned} \tag{7.19}$$

The analysis of this discrete problem will be given in the next section. Particularly, problem (7.19) possesses a unique solution, and the following error estimate holds:

$$\|\mathbf{u} - \mathbf{u}_h\|_{\mathbf{L}^2(\Omega)} + h\|\mathbf{u} - \mathbf{u}_h\|_{\mathbf{H}^1(\Omega)} \leq Ch^2\|\mathbf{f}\|_{\mathbf{L}^2(\Omega)} , \tag{7.20}$$

where again the constant C depends on μ_1 and μ_2, but is independent of $\lambda \in [0, \infty)$. Thus this method is also locking-free.

7.3.3.2 Nonconforming Finite Elements on Rectangles

Consider the case where Ω is a rectangular domain and K_h is a regular partition of Ω into rectangles such that the horizontal and vertical edges of rectangles are parallel to the x_1- and x_2-coordinate axes, respectively. Define the nonconforming finite element space on rectangles (cf. Sect. 2.1.2)

$$\begin{aligned}\mathbf{V}_h = \{\mathbf{v} \in \left(L^2(\Omega)\right)^2 : v_i|_K = a_K^{i,1} + a_K^{i,2}x_1 + a_K^{i,3}x_2 + a_K^{i,4}(x_1^2 - x_2^2)\ ,\\ i = 1, 2,\ a_K^{i,j} \in \mathbb{R},\ K \in K_h;\ \mathbf{v} \text{ is continuous}\\ \text{at the midpoints of interior edges and is}\\ \text{zero at the midpoints of edges on } \Gamma\}\ .\end{aligned}$$

The degrees of freedom in $\mathbf{V}_h$ are defined in terms of nodal values. They can also use mean values over edges, as in Sect. 2.1.2. With this space, the nonconforming method can be defined as in (7.19), and estimate (7.20) remains true (Chen et al., 2004A).

7.4 Theoretical Considerations

As an example, we give an analysis for the nonconforming finite element method for the pure displacement problem.

Recall the space

$$\mathbf{V} = \{\mathbf{v} \in (H^1(\Omega))^2 : \mathbf{v} = \mathbf{0} \text{ on } \Gamma\}\ ,$$

and the bilinear form $a(\cdot, \cdot) : \mathbf{V} \times \mathbf{V} \to \mathbb{R}$

$$a(\mathbf{v}, \mathbf{w}) = \mu\,(\nabla\mathbf{v}, \nabla\mathbf{w}) + (\mu + \lambda)\,(\nabla \cdot \mathbf{v}, \nabla \cdot \mathbf{w})\,, \qquad \mathbf{v},\ \mathbf{w} \in \mathbf{V}\ .$$

Then (7.18) becomes: Find $\mathbf{u} \in \mathbf{V}$ such that

$$a(\mathbf{u}, \mathbf{v}) = (\mathbf{f}, \mathbf{v})\,, \qquad \mathbf{v} \in \mathbf{V}\ . \tag{7.21}$$

Note that $a(\cdot, \cdot)$ is symmetric. Also, it is bounded:

$$|a(\mathbf{v}, \mathbf{w})| \le C(\mu, \lambda)\|\mathbf{v}\|_{\mathbf{H}^1(\Omega)}\|\mathbf{w}\|_{\mathbf{H}^1(\Omega)}, \qquad \mathbf{v},\ \mathbf{w} \in \mathbf{V}\ .$$

Using Poincaré's inequality (1.36), we easily see that

$$a(\mathbf{v}, \mathbf{v}) \ge C(\mu, \Omega)\|\mathbf{v}\|_{\mathbf{H}^1(\Omega)}, \qquad \mathbf{v} \in \mathbf{V}\ ;$$

i.e., $a(\cdot, \cdot)$ is $\mathbf{V}$-coercive. Applying these three properties, it follows from the Lax-Milgram Lemma (Theorem 1.1) that problem (7.21) (i.e., (7.18)) has a unique solution. Moreover, the following elliptic regularity result holds (Brenner-Scott, 1994):

$$\|\mathbf{u}\|_{\mathbf{H}^2(\Omega)} + \lambda\|\nabla \cdot \mathbf{u}\|_{H^1(\Omega)} \leq C\|\mathbf{f}\|_{\mathbf{L}^2(\Omega)} \,, \tag{7.22}$$

where the positive constant C is independent of $(\mu, \lambda) \in [\mu_1, \mu_2] \times (0, \infty)$.

To analyze problem (7.19), we define the mesh-dependent bilinear form $a_h(\cdot,\cdot) : \mathbf{V}_h \times \mathbf{V}_h \to \mathbb{R}$

$$a_h(\mathbf{v}, \mathbf{w}) = \sum_{K \in K_h} \left\{ \mu \left(\nabla \mathbf{v}, \nabla \mathbf{w}\right)_K + (\mu + \lambda) \left(\nabla \cdot \mathbf{v}, \nabla \cdot \mathbf{w}\right)_K \right\} \,, \qquad \mathbf{v},\ \mathbf{w} \in \mathbf{V}_h \,.$$

Then (7.19) is replaced with: Find $\mathbf{u}_h \in \mathbf{V}_h$ such that

$$a_h(\mathbf{u}_h, \mathbf{v}) = (\mathbf{f}, \mathbf{v}) \,, \qquad \mathbf{v} \in \mathbf{V}_h \,. \tag{7.23}$$

Introduce the nonconforming energy norm $\|\cdot\|_h$ on $\mathbf{V}_h \cup \left(H_0^1(\Omega)\right)^2$

$$\|\mathbf{v}\|_h = \sqrt{a_h(\mathbf{v}, \mathbf{v})}, \qquad \mathbf{v} \in \mathbf{V}_h \cup \left(H_0^1(\Omega)\right)^2 \,. \tag{7.24}$$

It is obvious that

$$\|\nabla \mathbf{v}\|_{\mathbf{L}^2(\Omega)} \leq \mu_1^{-1/2} \|\mathbf{v}\|_h, \qquad \mathbf{v} \in \mathbf{V}_h \cup \left(H_0^1(\Omega)\right)^2 \,, \tag{7.25}$$

where (here and below) any differential operator on $\mathbf{V}_h$ is defined element-by-element. Also, we define an interpolation operator $\mathbf{\Pi}_h : \left(H^2(\Omega)\right)^2 \cap \mathbf{V} \to \mathbf{V}_h$ by

$$\int_e \mathbf{\Pi}_h \mathbf{v} \; d\ell = \int_e \mathbf{v} \; d\ell, \qquad \mathbf{v} \in \left(H^2(\Omega)\right)^2 \cap \mathbf{V} \,, \tag{7.26}$$

where e is any edge in the partition K_h. By (7.26) and Green's formula (1.19), we see that

$$\int_K \nabla \cdot (\mathbf{\Pi}_h \mathbf{v}) \; d\mathbf{x} = \int_K \nabla \cdot \mathbf{v} \; d\mathbf{x} \qquad \forall K \in K_h \,.$$

Consequently, by the definition of $\mathbf{V}_h$ in Sect. 7.3.3.1, we have

$$\nabla \cdot (\mathbf{\Pi}_h \mathbf{v}) = \frac{1}{|K|} \int_K \nabla \cdot \mathbf{v} \; d\mathbf{x} \qquad \forall K \in K_h \,, \tag{7.27}$$

where $|K|$ denotes the area of the triangle K. It can be shown that the operator $\mathbf{\Pi}_h$ has the approximation property (cf. Sect. 1.9)

$$\|\mathbf{v} - \mathbf{\Pi}_h \mathbf{v}\|_{\mathbf{L}^2(\Omega)} + h\|\nabla(\mathbf{v} - \mathbf{\Pi}_h \mathbf{v})\|_{\mathbf{L}^2(\Omega)} \leq Ch^2 |\mathbf{v}|_{\mathbf{H}^2(\Omega)} \,, \tag{7.28}$$

where $|\cdot|_{\mathbf{H}^2(\Omega)}$ is the semi-norm of $\mathbf{H}^2(\Omega)$.

The proof of the next lemma is exactly the same as that of Lemma 2.1.

Lemma 7.1. *Let* $\mathbf{u}$ *and* $\mathbf{u}_h$ *be the respective solutions to* (7.21) *and* (7.23). *Then it holds that*

$$\|\mathbf{u}-\mathbf{u}_h\|_h \le \inf_{\mathbf{v}\in\mathbf{V}_h}\|\mathbf{u}-\mathbf{v}\|_h + \sup_{\mathbf{v}\in\mathbf{V}_h\setminus\{\mathbf{0}\}}\frac{|a_h(\mathbf{u},\mathbf{v})-(\mathbf{f},\mathbf{v})|}{\|\mathbf{v}\|_h}\ .$$

We also need the next result, which can be found in Arnold et al. (1988).

Lemma 7.2. *There exists a constant $C(\Omega)>0$ such that for any $\mathbf{g}\in\left(H^2(\Omega)\cap H_0^1(\Omega)\right)^2$, there is $\mathbf{u}_1\in\left(H^2(\Omega)\cap H_0^1(\Omega)\right)^2$ satisfying*

$$\nabla\cdot\mathbf{u}_1=\nabla\cdot\mathbf{g}\ ,$$

and

$$\|\mathbf{u}_1\|_{\mathbf{H}^2(\Omega)}\le C(\Omega)\|\nabla\cdot\mathbf{g}\|_{H^1(\Omega)}\ .$$

Theorem 7.3. *Let $\mathbf{u}$ and $\mathbf{u}_h$ be the respective solutions to* (7.21) *and* (7.23). *Then there is a constant $C>0$, independent of h and of $(\mu,\lambda)\in[\mu_1,\mu_2]\times(0,\infty)$, such that*

$$\|\mathbf{u}-\mathbf{u}_h\|_h\le Ch\|\mathbf{f}\|_{\mathbf{L}^2(\Omega)}\ .$$

Proof. By the definition of $a_h(\cdot,\cdot)$ and (7.21), we have

$$\begin{aligned}a_h(\mathbf{u},\mathbf{v})-(\mathbf{f},\mathbf{v}) &= \mu\left(\sum_{K\in K_h}\int_K\nabla\mathbf{u}:\nabla\mathbf{v}\ d\mathbf{x}+\int_\Omega\Delta\mathbf{u}\cdot\mathbf{v}\ d\mathbf{x}\right)+(\mu+\lambda)\\ &\quad\times\left(\sum_{K\in K_h}\int_K\nabla\cdot\mathbf{u}\nabla\cdot\mathbf{v}\ d\mathbf{x}+\int_\Omega\nabla(\nabla\cdot\mathbf{u})\cdot\mathbf{v}\ d\mathbf{x}\right).\end{aligned}\tag{7.29}$$

Applying a homogeneity argument (cf. Sect. 1.9) and the Bramble-Hilbert Lemma (Lemma 1.4), we have

$$\begin{aligned}&\left|\sum_{K\in K_h}\int_K\nabla\mathbf{u}:\nabla\mathbf{v}\ d\mathbf{x}+\int_\Omega\Delta\mathbf{u}\cdot\mathbf{v}\ d\mathbf{x}\right|\\ &\qquad\le Ch|\mathbf{u}|_{\mathbf{H}^2(\Omega)}\|\nabla\mathbf{v}\|_{\mathbf{L}^2(\Omega)}\ ,\end{aligned}\tag{7.30}$$

and

$$\begin{aligned}&\left|\sum_{K\in K_h}\int_K\nabla\cdot\mathbf{u}\nabla\cdot\mathbf{v}\ d\mathbf{x}+\int_\Omega\nabla(\nabla\cdot\mathbf{u})\cdot\mathbf{v}\ d\mathbf{x}\right|\\ &\qquad\le Ch|\nabla\cdot\mathbf{u}|_{H^1(\Omega)}\|\nabla\mathbf{v}\|_{\mathbf{L}^2(\Omega)}\ .\end{aligned}\tag{7.31}$$

Consequently, by (7.22), (7.25), and (7.29)–(7.31), we get

$$\begin{aligned}&|a_h(\mathbf{u},\mathbf{v})-(\mathbf{f},\mathbf{v})|\\ &\quad\le Ch\left(\mu\,|\mathbf{u}|_{\mathbf{H}^2(\Omega)}+(\mu+\lambda)\,|\nabla\cdot\mathbf{u}|_{H^1(\Omega)}\right)\|\nabla\mathbf{v}\|_{\mathbf{L}^2(\Omega)}\\ &\quad\le Ch\,\|\mathbf{f}\|_{\mathbf{L}^2(\Omega)}\,\|\mathbf{v}\|_h\ .\end{aligned}\tag{7.32}$$

Next, by (7.24) and the definition of $a_h(\cdot,\cdot)$, we have

$$\begin{aligned}\inf_{\mathbf{v}\in\mathbf{V}_h}\|\mathbf{u}-\mathbf{v}\|_h &\le \|\mathbf{u}-\boldsymbol{\Pi}_h\mathbf{u}\|_h \\ &= \left(\mu\,\|\nabla(\mathbf{u}-\boldsymbol{\Pi}_h\mathbf{u})\|^2_{\mathbf{L}^2(\Omega)} + (\mu+\lambda)\,\|\nabla\cdot(\mathbf{u}-\boldsymbol{\Pi}_h\mathbf{u})\|^2_{L^2(\Omega)}\right)^{1/2} . \end{aligned} \tag{7.33}$$

From Lemma 7.2, there is a $\mathbf{u}_1 \in \left(H^2(\Omega)\cap H_0^1(\Omega)\right)^2$ such that

$$\nabla\cdot\mathbf{u}_1 = \nabla\cdot\mathbf{u}\,, \tag{7.34}$$

and

$$\|\mathbf{u}_1\|_{\mathbf{H}^2(\Omega)} \le C\,\|\nabla\cdot\mathbf{u}\|_{H^1(\Omega)}\,. \tag{7.35}$$

Hence it follows from (7.22) and (7.35) that

$$\|\mathbf{u}_1\|_{\mathbf{H}^2(\Omega)} \le \frac{C}{1+\lambda}\,\|\mathbf{f}\|_{\mathbf{L}^2(\Omega)}\,. \tag{7.36}$$

Also, by (7.27) and (7.34), we observe that

$$\nabla\cdot(\boldsymbol{\Pi}_h\mathbf{u}_1) = \nabla\cdot(\boldsymbol{\Pi}_h\mathbf{u})\,. \tag{7.37}$$

Thus, using (7.34) and (7.37), we have

$$\|\nabla\cdot\mathbf{u}-\nabla\cdot(\boldsymbol{\Pi}_h\mathbf{u})\|_{L^2(\Omega)} = \|\nabla\cdot\mathbf{u}_1-\nabla\cdot(\boldsymbol{\Pi}_h\mathbf{u}_1)\|_{L^2(\Omega)}\,. \tag{7.38}$$

Now, by (7.22), (7.28), (7.33), (7.36), and (7.38), we obtain

$$\inf_{\mathbf{v}\in\mathbf{V}_h}\|\mathbf{u}-\mathbf{v}\|_h \le Ch\|\mathbf{f}\|_{\mathbf{L}^2(\Omega)}\,. \tag{7.39}$$

Finally, we apply (7.32) and (7.39) in Lemma 7.1 to yield the desired result. □

We now apply a duality argument to the derivation of an error estimate in the L^2-norm (cf. Sect. 2.4).

Theorem 7.4. *Let Ω be convex, and $\mathbf{u}$ and $\mathbf{u}_h$ be the respective solutions to (7.21) and (7.23). Then there exists a constant $C>0$, independent of h and of $(\mu,\lambda)\in[\mu_1,\mu_2]\times(0,\infty)$, such that*

$$\|\mathbf{u}-\mathbf{u}_h\|_{\mathbf{L}^2(\Omega)} \le Ch^2\|\mathbf{f}\|_{\mathbf{L}^2(\Omega)}\,.$$

Proof. It follows from duality that

$$\|\mathbf{u}-\mathbf{u}_h\|_{\mathbf{L}^2(\Omega)} = \sup_{\mathbf{v}\in(L^2(\Omega))^2\setminus\{\mathbf{0}\}}\frac{(\mathbf{u}-\mathbf{u}_h,\mathbf{v})}{\|\mathbf{v}\|_{\mathbf{L}^2(\Omega)}}\,. \tag{7.40}$$

For a fixed $\mathbf{v}\in\left(L^2(\Omega)\right)^2\setminus\{\mathbf{0}\}$, let $\mathbf{p}\in\left(H^2(\Omega)\cap H_0^1(\Omega)\right)^2$ satisfy

$$-\mu\Delta\mathbf{p} - (\mu+\lambda)\nabla(\nabla\cdot\mathbf{p}) = \mathbf{v} \qquad \text{in } \Omega\ , \tag{7.41}$$

and let $\mathbf{p}_h \in \mathbf{V}_h$ be the corresponding nonconforming finite element solution of

$$a_h(\mathbf{p}_h, \mathbf{w}) = (\mathbf{v}, \mathbf{w}) \qquad \forall \mathbf{w} \in \mathbf{V}_h\ . \tag{7.42}$$

Using (7.22), we see that

$$\|\mathbf{p}\|_{\mathbf{H}^2(\Omega)} + \lambda\|\nabla\cdot\mathbf{p}\|_{H^1(\Omega)} \le C\|\mathbf{v}\|_{\mathbf{L}^2(\Omega)}\ . \tag{7.43}$$

Also, it follows from Theorem 7.3 that

$$\|\mathbf{p} - \mathbf{p}_h\|_h \le Ch\|\mathbf{v}\|_{\mathbf{L}^2(\Omega)}\ . \tag{7.44}$$

By (7.23), (7.42), and Cauchy's inequality (1.10), we get

$$\begin{aligned}
&|(\mathbf{u}-\mathbf{u}_h, \mathbf{v})| \\
&\quad = |a_h(\mathbf{p}, \mathbf{u}) - a_h(\mathbf{p}_h, \mathbf{u}_h)| \\
&\quad = |a_h(\mathbf{p}-\mathbf{p}_h, \mathbf{u} - \boldsymbol{\Pi}_h\mathbf{u}) + a_h(\mathbf{p}-\mathbf{p}_h, \boldsymbol{\Pi}_h\mathbf{u}) \\
&\qquad + a_h(\mathbf{p}_h - \boldsymbol{\Pi}_h\mathbf{p}, \mathbf{u}-\mathbf{u}_h) + a_h(\boldsymbol{\Pi}_h\mathbf{p}, \mathbf{u}-\mathbf{u}_h)| \\
&\quad \le \|\mathbf{p}-\mathbf{p}_h\|_h \, \|\mathbf{u} - \boldsymbol{\Pi}_h\mathbf{u}\|_h + \|\mathbf{p}_h - \boldsymbol{\Pi}_h\mathbf{p}\|_h \, \|\mathbf{u}-\mathbf{u}_h\|_h \\
&\qquad + |a_h(\mathbf{p}-\mathbf{p}_h, \boldsymbol{\Pi}_h\mathbf{u})| + |a_h(\boldsymbol{\Pi}_h\mathbf{p}, \mathbf{u}-\mathbf{u}_h)|\ .
\end{aligned} \tag{7.45}$$

Using Theorem 7.3, (7.43), (7.44), and the same argument as for (7.39), we have

$$\begin{aligned}
\|\mathbf{p}-\mathbf{p}_h\|_h \, \|\mathbf{u} - \boldsymbol{\Pi}_h\mathbf{u}\|_h + \|\mathbf{p}_h - \boldsymbol{\Pi}_h\mathbf{p}\|_h \, \|\mathbf{u}-\mathbf{u}_h\|_h \\
\le Ch^2\|\mathbf{v}\|_{\mathbf{L}^2(\Omega)}\|\mathbf{f}\|_{\mathbf{L}^2(\Omega)}\ .
\end{aligned} \tag{7.46}$$

Next, by a homogeneity argument and the Bramble-Hilbert Lemma (Lemma 1.4), we see that

$$\begin{aligned}
\Big|\sum_{K\in K_h}\int_K \nabla\boldsymbol{\Pi}_h\mathbf{p} : \nabla\mathbf{u}\ d\mathbf{x} + \sum_{K\in K_h}\int_K \boldsymbol{\Pi}_h\mathbf{p}\cdot\Delta\mathbf{u}\ d\mathbf{x}\Big| \\
\le Ch^2|\mathbf{p}|_{\mathbf{H}^2(\Omega)}|\mathbf{u}|_{\mathbf{H}^2(\Omega)}\ ,
\end{aligned} \tag{7.47}$$

and

$$\begin{aligned}
\Big|\sum_{K\in K_h}\int_K \nabla\cdot(\boldsymbol{\Pi}_h\mathbf{p})\nabla\cdot\mathbf{u}\ d\mathbf{x} + \sum_{K\in K_h}\int_K \boldsymbol{\Pi}_h\mathbf{p}\cdot\nabla(\nabla\cdot\mathbf{v})\ d\mathbf{x}\Big| \\
\le Ch^2|\mathbf{p}|_{\mathbf{H}^2(\Omega)}|\nabla\cdot\mathbf{u}|_{H^1(\Omega)}\ .
\end{aligned} \tag{7.48}$$

Combining (7.47) and (7.48) gives

$$\begin{aligned}
|a_h(\boldsymbol{\Pi}_h\mathbf{p}, \mathbf{u}-\mathbf{u}_h)| &= |a_h(\boldsymbol{\Pi}_h\mathbf{p}, \mathbf{u}) - (\mathbf{f}, \boldsymbol{\Pi}_h\mathbf{p})| \\
&\le Ch^2|\mathbf{p}|_{\mathbf{H}^2(\Omega)}\left(\mu|\mathbf{u}|_{\mathbf{H}^2(\Omega)} + (\mu+\lambda)\|\nabla\cdot\mathbf{u}\|_{H^1(\Omega)}\right)\ ,
\end{aligned}$$

so that, by (7.22) and (7.43),

$$|a_h(\mathbf{\Pi}_h\mathbf{p}, \mathbf{u} - \mathbf{u}_h)| \leq Ch^2 \|\mathbf{v}\|_{\mathbf{L}^2(\Omega)} \|\mathbf{f}\|_{\mathbf{L}^2(\Omega)} . \tag{7.49}$$

Similarly, we can show that

$$\begin{aligned} |a_h(\mathbf{p} - \mathbf{p}_h, \mathbf{\Pi}_h\mathbf{u})| &= |a_h(\mathbf{p}, \mathbf{\Pi}_h\mathbf{u}) - (\mathbf{v}, \mathbf{\Pi}_h\mathbf{u})| \\ &\leq Ch^2 \|\mathbf{v}\|_{\mathbf{L}^2(\Omega)} \|\mathbf{f}\|_{\mathbf{L}^2(\Omega)} . \end{aligned} \tag{7.50}$$

Substituting (7.46), (7.49), and (7.50) into (7.45) yields

$$|(\mathbf{u} - \mathbf{u}_h, \mathbf{v})| \leq Ch^2 \|\mathbf{v}\|_{\mathbf{L}^2(\Omega)} \|\mathbf{f}\|_{\mathbf{L}^2(\Omega)} ,$$

which, together with (7.40), implies the desired result. □

Estimate (7.20) follows from Theorems 7.3 and 7.4. Also, the analysis in this subsection applies to the nonconforming finite element on rectangles discussed in Sect. 7.3.3.2 (Chen et al., 2004A).

7.5 Bibliographical Remarks

Sections 7.1 and 7.2 contain a very compact introduction to linear elasticity. For more details, see Ciarlet (1988) and Braess (1997). For the analysis of the PEERS mixed finite element discussed in Sect. 7.3.2, see Arnold et al. (1984A). Finally, the theoretical considerations outlined in Sect. 7.4 follow Brenner-Scott (1994).

7.6 Exercises

7.1. Derive (7.9) from (7.7) and (7.8) in detail.
7.2. Show that if the surface Γ_D has a positive area, system (7.9) has a unique solution. (If necessary, see Sect. 1.3.1.)
7.3. Prove (7.10).
7.4. Derive (7.11) from (7.4)–(7.7) (cf. Sect. 7.2.2).
7.5. Show that if the surface Γ_D has a positive area, the discrete problem (7.13) has a unique solution. (If necessary, see Sect. 1.3.2.)
7.6. We consider another two-dimensional version of the elastic model given in (7.4)–(7.7). Let $\Omega \subset \mathbb{R}^2$ be a planar domain, the elasticity body have the form $\Omega \times (-l, l)$, where l is a small real number, and $f_3 = g_3 = 0$. This problem corresponds to a thin elastic plate with a middle surface Ω subject to planar loads only (no transversal loads). Assuming a *planar stress state* (i.e., $\sigma_{i3} = 0$, $i = 1, 2, 3$), show that (7.4)–(7.7) in this case become

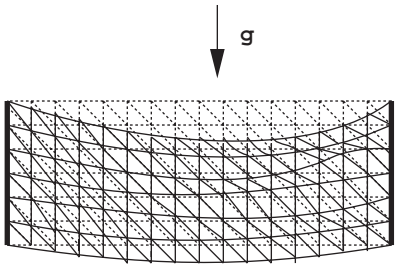

Fig. 7.2. An example of the computed displacements

$$
\begin{aligned}
&\epsilon_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right), && i,\ j = 1,2,\ \mathbf{x} \in \Omega\ ,\\
&\sum_{j=1}^{2} \frac{\partial \sigma_{ij}}{\partial x_j} + f_i = 0, && i = 1,2,\ \mathbf{x} \in \Omega\ ,\\
&\sigma_{ij} = 2\mu\epsilon_{ij}(\mathbf{u}) + \bar{\lambda}\big(\epsilon_{11}(\mathbf{u}) + \epsilon_{22}(\mathbf{u})\big)\delta_{ij}, && i,\ j = 1,2,\ \mathbf{x} \in \Omega\ ,\\
&u_1 = u_2 = 0, && \mathbf{x} \in \Gamma_D\ ,\\
&\sum_{j=1}^{2} \sigma_{ij}\nu_j = g_i, && i = 1,2,\ \mathbf{x} \in \Gamma_N\ ,
\end{aligned}
$$

where f_i and g_i are given forces and

$$\bar{\lambda} = \frac{E\nu}{1-\nu^2}\ .$$

Write down a variational formulation of this problem in the displacement form, and formulate the corresponding conforming and nonconforming finite element methods using linear displacements on triangles (cf. Sects. 7.3.1 and 7.3.3). A numerical example of the computed displacements using the conforming method is displayed in Fig. 7.2 for a thin plate fixed at both ends and subject to a distributed load as shown. The Young modulus E differs in the upper and lower halfs of the plate, with E larger in the lower half.

8 Fluid Mechanics

8.1 Introduction

The motion of a continuous medium is governed by the fundamental principles of classical mechanics and thermodynamics for the *conservation of mass*, *momentum*, and *energy*. In particular, the application of the first two principles in a frame of reference leads to the following differential equations:

$$\begin{aligned} &\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0 \ , \\ &\frac{\partial}{\partial t}(\rho \mathbf{u}) + \nabla \cdot (\rho \mathbf{u}\mathbf{u} - \boldsymbol{\sigma}) = \mathbf{f} \ , \end{aligned} \tag{8.1}$$

where ρ, $\mathbf{u}$, and $\boldsymbol{\sigma}$ are, respectively, the density, velocity, and stress tensor of the continuous medium, and $\mathbf{f}$ is the force (per unit volume). These equations are in *divergence form*. Their *nondivergence form* is (cf. Exercise 8.1)

$$\begin{aligned} &\frac{D\rho}{Dt} + \rho \nabla \cdot \mathbf{u} = 0 \ , \\ &\rho \frac{D\mathbf{u}}{Dt} - \nabla \cdot \boldsymbol{\sigma} = \mathbf{f} \ , \end{aligned} \tag{8.2}$$

where the *material derivative* is defined by

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \mathbf{u} \cdot \nabla \ .$$

These equations are based on the *Eulerian approach* for the description of continuum motion; i.e., the characteristic properties of the medium $(\rho, \mathbf{u}, \boldsymbol{\sigma})$ are treated as functions of time and space in the frame of reference. An alternative description is through the *Lagrangian approach* where the dependent variables are the characteristic properties of material particles that are followed in their motion; i.e., these properties are the functions of time and parameters used to identify the particles such as the particle coordinates at a fixed initial time. This approach, or more precisely the mixed Lagrangian-Eulerian approach, is mostly interesting for problems involving different media with interfaces. It is not as widely used in fluid mechanics as the Eulerian approach and thus is not presented.

The basic unknowns in (8.1) or (8.2) are $(\rho, \mathbf{u})$. A constitutive relationship is needed for the stress tensor $\boldsymbol{\sigma}$ as in the preceding chapter. A fluid is *Newtonian* if its stress tensor is a linear function of the velocity gradient. For this type of fluid, the *Newton law* (or *Navier-Stokes law*) applies:

$$\boldsymbol{\sigma} = -p\mathbf{I} + \boldsymbol{\tau}, \quad \boldsymbol{\tau} = \mu(\nabla\mathbf{u} + (\nabla\mathbf{u})^T) + \lambda\nabla\cdot\mathbf{u}\mathbf{I}\,, \tag{8.3}$$

where p and $\boldsymbol{\tau}$ are the pressure and viscous stress tensor, and μ and λ are the viscosity coefficients. Note that the second equation has the same form as (7.6).

We substitute (8.3) into the second (momentum) equation in (8.2) to obtain

$$\begin{aligned}\rho\frac{D\mathbf{u}}{Dt} + \nabla p = \mu\Delta\mathbf{u} + (\mu+\lambda)\nabla(\nabla\cdot\mathbf{u}) + \mathbf{f} \\ + \nabla\cdot\mathbf{u}\nabla\lambda + \nabla\mu\cdot(\nabla\mathbf{u} + (\nabla\mathbf{u})^T)\,.\end{aligned} \tag{8.4}$$

In general, the viscosity coefficients depend on temperature; in the present case where the temperature is fixed, they are constant. Consequently, (8.4) becomes

$$\rho\frac{D\mathbf{u}}{Dt} + \nabla p = \mu\Delta\mathbf{u} + (\mu+\lambda)\nabla(\nabla\cdot\mathbf{u}) + \mathbf{f}\,. \tag{8.5}$$

An *incompressible flow* is characterized by the condition

$$\nabla\cdot\mathbf{u} = 0\,. \tag{8.6}$$

Using (8.6), the first (mass conservation) equation in (8.2) becomes

$$\frac{D\rho}{Dt} = 0\,. \tag{8.7}$$

This equation implies that the density is constant along a fluid particle trajectory. In most cases, we can assume that ρ is constant so that (8.7) is satisfied everywhere.

Under condition (8.6), the momentum (8.5) becomes

$$\rho\left(\frac{\partial\mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u}\right) + \nabla p - \mu\Delta\mathbf{u} = \mathbf{f}\,. \tag{8.8}$$

This equation is known as the Navier-Stokes equation. In the incompressible case, the unknown variables are the pressure and velocity field. They can be determined from (8.6) and (8.8). The Navier-Stokes equation can be also presented in the *stream-function vorticity formulation*, which is not discussed in this chapter.

We observe that the Navier-Stokes equation is nonlinear. If we neglect the nonlinear term, we derive the *Stokes equation*

$$\rho\frac{\partial\mathbf{u}}{\partial t} + \nabla p - \mu\Delta\mathbf{u} = \mathbf{f}\,. \tag{8.9}$$

Strictly speaking, the Stokes equation is valid only for a viscous Newtonian fluid over a limited range of flow rates where *turbulence*, *inertial*, and other high velocity effects are negligible. As the flow velocity is increased, deviations from the Stokes flow are observed. The generally accepted explanation is that, as the velocity is increased, deviations are due to inertial effects first, followed later by turbulent effects. Such a phenomenon can be characterized by the well known *Reynolds number* that expresses the ratio between the inertial force and the viscous (frictional) force and can be defined, for example, by

$$Re = \frac{Lu^*}{\mu} ,$$

where L and u^* are some reference length and velocity of a medium, respectively. This number can be used as a criterion to distinguish between *laminar flow* occurring at low velocities and *turbulent flow*. The critical number Re between these two types of flows in pipes is about 2,100, for example. In this chapter, we consider a low velocity flow of an incompressible Newtonian fluid. Especially, we concentrate on the Stokes equation and make remarks on the extension of the presentation and analysis to the Navier-Stokes equation. Turbulent flow is beyond the scope of this chapter.

In Sect. 8.2, we introduce variational formulations of the Stokes equation. Then, in Sect. 8.3, we develop the conforming, mixed, and nonconforming finite element methods. In Sect. 8.4, we remark on an extension to the Navier-Stokes equation. Section 8.5 is devoted to theoretical considerations. Finally, in Sect. 8.6, we give bibliographical information.

8.2 Variational Formulations

8.2.1 The Galerkin Approach

We recall the stationary Stokes equation, together with a boundary condition, in a domain $\Omega \subset \mathbb{R}^3$:

$$\begin{aligned} -\mu\Delta\mathbf{u} + \nabla p &= \mathbf{f} && \text{in } \Omega , \\ \nabla \cdot \mathbf{u} &= 0 && \text{in } \Omega , \\ \mathbf{u} &= \mathbf{0} && \text{on } \Gamma , \end{aligned} \tag{8.10}$$

where Γ is the boundary of Ω. The boundary condition in (8.10) is often called the *no-slip condition*. We write (8.10) in a variational formulation. For this, define

$$\mathbf{V} = \left\{ \mathbf{v} \in \left(H_0^1(\Omega)\right)^3 : \nabla \cdot \mathbf{v} = 0 \text{ in } \Omega \right\} .$$

Then, using Green's formula (1.19), we are led to the variational formulation of (8.10): Find $\mathbf{u} \in \mathbf{V}$ such that

$$\mu\,(\nabla \mathbf{u}, \nabla \mathbf{v}) = (\mathbf{f}, \mathbf{v})\,, \qquad \mathbf{v} \in \mathbf{V}\,. \tag{8.11}$$

It can be checked that (8.11) has a unique solution $\mathbf{u} \in \mathbf{V}$ (cf. Exercise 8.3). Note that the pressure p disappears in (8.11), which stems from the fact that we are working with the space $\mathbf{V}$ of velocities that satisfy the incompressibility condition.

8.2.2 The Mixed Formulation

As we will see in the next section, problem (8.10) can be solved more appropriately by the mixed finite element method studied in Chap. 3. Note that (8.10) determines the pressure p only up to an additive constant, which is usually fixed by enforcing the integral condition

$$\int_\Omega p\; d\mathbf{x} = 0\,.$$

We introduce the spaces

$$\mathbf{V} = \left(H_0^1(\Omega)\right)^3, \qquad W = \left\{ w \in L^2(\Omega) : \int_\Omega w\; d\mathbf{x} = 0 \right\}\,.$$

As in Chap. 3 (cf. Exercise 8.4), problem (8.10) can be now written in a mixed formulation: Find $\mathbf{u} \in \mathbf{V}$ and $p \in W$ such that

$$\begin{aligned} &\mu\,(\nabla \mathbf{u}, \nabla \mathbf{v}) - (\nabla \cdot \mathbf{v}, p) = (\mathbf{f}, \mathbf{v})\,, && \mathbf{v} \in \mathbf{V}\,, \\ &(\nabla \cdot \mathbf{u}, w) = 0, && w \in W\,. \end{aligned} \tag{8.12}$$

System (8.12) can be shown to have a unique solution $(\mathbf{u}, p) \in \mathbf{V} \times W$ (cf. Sect. 8.5). Moreover, this problem satisfies an *inf-sup* condition similar to (3.30) (cf. (8.25)):

$$\inf_{w \in W} \sup_{\mathbf{v} \in \mathbf{V}} \frac{b(\mathbf{v}, w)}{\|\mathbf{v}\|_{\mathbf{H}^1(\Omega)} \|w\|_{L^2(\Omega)}} \geq b_* > 0\,.$$

8.3 Finite Element Methods

8.3.1 Galerkin Finite Elements

To introduce a finite element method based on (8.11), we need to construct a finite element space $\mathbf{V}_h$ that is a subspace of the space $\mathbf{V}$ defined in Sect. 8.2.1. This is not an easy task because the elements $\mathbf{v}$ in $\mathbf{V}_h$ have to satisfy the condition $\nabla \cdot \mathbf{v} = 0$, i.e., the *divergence free* condition. For simplicity, let Ω be a convex polygonal domain in the plane. It follows from a theorem in the advanced calculus that if Ω does not contain any "holes", i.e.,

if Ω is simply connected, then $\nabla \cdot \mathbf{v} = 0$ if and only if there exists a function $w \in H^2(\Omega)$ such that (Kaplan, 1991)

$$\mathbf{v} = \mathbf{rot}\ w \equiv \left(\frac{\partial w}{\partial x_2}, -\frac{\partial w}{\partial x_1} \right) .$$

More precisely, it holds that

$$\mathbf{v} \in \mathbf{V} \text{ if and only if } \mathbf{v} = \mathbf{rot}\ w, \quad w \in H_0^2(\Omega) . \tag{8.13}$$

The function w is called the *stream function* associated with the velocity $\mathbf{v}$.

Let K_h be a regular triangulation of Ω into triangles as in Chap. 1. Define

$$W_h = \left\{ w \in H_0^2(\Omega) : w|_K \in P_5(K),\ K \in K_h \right\} .$$

As discussed in Example 1.5 in Chap. 1, because the first partial derivatives of functions in W_h are required to be continuous on Ω, there are at least six degrees of freedom on each interior edge in K_h. Thus the polynomial degree of the finite element space W_h must be at least five. Each function in W_h is in $C^1(\bar{\Omega})$ (cf. Exercise 1.17). Now, set

$$\mathbf{V}_h = \{ \mathbf{v} \in \mathbf{V} : \mathbf{v} = \mathbf{rot}\ w,\ \ w \in W_h \} .$$

The Galerkin finite element method for (8.10) reads: Find $\mathbf{u}_h \in \mathbf{V}_h$ such that

$$\mu\,(\nabla \mathbf{u}_h, \nabla \mathbf{v}) = (\mathbf{f}, \mathbf{v}) , \qquad \mathbf{v} \in \mathbf{V}_h . \tag{8.14}$$

It possesses a unique solution (cf. Exercise 8.5). Furthermore, we have the error estimate (Ciarlet, 1978)

$$\|\mathbf{u} - \mathbf{u}_h\|_{\mathbf{H}^1(\Omega)} \le C h^4 |\mathbf{u}|_{\mathbf{H}^5(\Omega)} .$$

Note that the elements in $\mathbf{V}_h$ must satisfy the incompressibility condition exactly. To be able to satisfy this condition, we use the space W_h that consists of piecewise polynomials of degree five. To utilize a finite element space of polynomials of lower degree, we will employ the mixed and nonconforming finite element methods introduced in the next two subsections.

8.3.2 Mixed Finite Elements

We now construct the mixed finite element method based on (8.12). For the Stokes problem, the velocity has a derivative of higher order than the pressure. This suggests the rule of thumb that the degree of the piecewise polynomials used to approximate the velocity should be higher than that of the polynomials for the pressure. However, it is known that this rule does not suffice to guarantee stability, and the spaces $\mathbf{V}_h$ and W_h have to be constructed very carefully. In this section, we state a couple of mixed finite elements that satisfy the discrete *inf-sup* condition (cf. (8.35)):

$$\inf_{w \in W_h} \sup_{\mathbf{v} \in \mathbf{V}_h} \frac{b(\mathbf{v}, w)}{\|\mathbf{v}\|_{\mathbf{H}^1(\Omega)} \|w\|_{L^2(\Omega)}} \ge b_{**} > 0 ,$$

where b_{**} is independent of h.

8.3.2.1 Example One

Let K_h be a regular triangulation of Ω into triangles. Define

$$\mathbf{V}_h = \left\{ \mathbf{v} \in \left(H_0^1(\Omega)\right)^2 : \mathbf{v}|_K \in (P_2(K))^2,\ K \in K_h \right\},$$
$$W_h = \{ w \in W : w|_K \in P_0(K),\ K \in K_h \}.$$

Then the mixed finite element method for (8.10) is: Find $\mathbf{u}_h \in \mathbf{V}_h$ and $p_h \in W_h$ such that

$$\begin{aligned} &\mu\,(\nabla \mathbf{u}_h, \nabla \mathbf{v}) - (\nabla \cdot \mathbf{v}, p_h) = (\mathbf{f}, \mathbf{v}), && \mathbf{v} \in \mathbf{V}_h, \\ &(\nabla \cdot \mathbf{u}_h, w) = 0, && w \in W_h. \end{aligned} \tag{8.15}$$

System (8.15) has a unique solution (cf. Sect. 8.5). Moreover, if Ω is convex, the solution satisfies (Girault-Raviart, 1981)

$$\|\mathbf{u} - \mathbf{u}_h\|_{\mathbf{L}^2(\Omega)} \le Ch^2 \left(\|\mathbf{u}\|_{\mathbf{H}^2(\Omega)} + \|p\|_{H^1(\Omega)} \right),$$
$$\|p - p_h\|_{L^2(\Omega)} \le Ch \left(\|\mathbf{u}\|_{\mathbf{H}^2(\Omega)} + \|p\|_{H^1(\Omega)} \right).$$

8.3.2.2 Example Two

The second example is the so-called *MINI element* (Arnold et al., 1984B). To introduce this element, let λ_1, λ_2, and λ_3 be the barycentric coordinates of a triangle (they are x_1, x_2, and $1 - x_1 - x_2$ in the unit triangle with vertices $(0,0)$, $(1,0)$, and $(0,1)$; see Sect. 1.4). Define

$$B_h = \left\{ v \in H^1(\Omega) : v|_K \in \text{span}\{\lambda_1\lambda_2\lambda_3\},\ K \in K_h \right\};$$

that is, B_h is the space of *cubic bubble functions* (cf. Sect. 6.3.2.1). Now, define the spaces

$$\mathbf{V}_h = \left\{ \mathbf{v} \in \left(H_0^1(\Omega)\right)^2 : \mathbf{v}|_K \in (P_1(K))^2,\ K \in K_h \right\} \oplus (B_h)^2,$$
$$W_h = \left\{ w \in W \cap H^1(\Omega) : w|_K \in P_1(K),\ K \in K_h \right\}.$$

With these choices, the mixed method can be defined as in (8.15). Moreover, if Ω is convex, the mixed finite element solution satisfies (cf. Sect. 8.5)

$$\begin{aligned} &\|\mathbf{u} - \mathbf{u}_h\|_{\mathbf{L}^2(\Omega)} + \|p - p_h\|_{L^2(\Omega)} \\ &\qquad \le Ch^2 \left(\|\mathbf{u}\|_{\mathbf{H}^2(\Omega)} + \|p\|_{H^2(\Omega)} \right). \end{aligned} \tag{8.16}$$

8.3.3 Nonconforming Finite Elements

We now develop the nonconforming finite element method discussed in Chap. 2 for the solution of (8.10).

8.3.3.1 Nonconforming Finite Elements on Triangles

The variational formulation is defined as in (8.11). For a convex polygonal domain Ω, let K_h be a regular triangulation of Ω into triangles as in Chap. 1. Define the finite element space on triangles (cf. Sect. 2.1.1)

$$\begin{aligned}\mathbf{V}_h = \{\mathbf{v} \in \left(L^2(\Omega)\right)^2 : \mathbf{v}|_K \text{ is linear, } K \in K_h; \mathbf{v} \text{ is continuous} \\ \text{at the midpoints of interior edges and} \\ \text{is zero at the midpoints of edges on } \Gamma; \\ \nabla \cdot \mathbf{v} = 0 \text{ on all } K \in K_h\} \,.\end{aligned}$$

Namely, all the functions in $\mathbf{V}_h$ are the nonconforming P_1-elements that are divergence-free on each triangle $K \in K_h$. The nonconforming method for (8.10) is: Find $\mathbf{u}_h \in \mathbf{V}_h$ such that

$$\mu \sum_{K \in K_h} (\nabla \mathbf{u}_h, \nabla \mathbf{v})_K = (\mathbf{f}, \mathbf{v}) \,, \qquad \mathbf{v} \in \mathbf{V}_h \,. \tag{8.17}$$

Existence and uniqueness of a solution to this problem can be shown exactly in the same way as for (2.3) in Sect. 2.1.1 (cf. Exercise 8.6). Furthermore, it can be proven in the same manner as in Sect. 2.4.2 that (Crouzeix-Raviart, 1973)

$$\|\mathbf{u} - \mathbf{u}_h\|_{\mathbf{L}^2(\Omega)} + h\|\mathbf{u} - \mathbf{u}_h\|_h \le Ch^2 \left(|\mathbf{u}|_{\mathbf{H}^2(\Omega)} + |p|_{H^1(\Omega)}\right) , \tag{8.18}$$

where

$$\begin{aligned}&\|\mathbf{v}\|_h = \sqrt{a_h(\mathbf{v}, \mathbf{v})}, \\ &a_h(\mathbf{v}, \mathbf{v}) = \mu \sum_{K \in K_h} (\nabla \mathbf{v}, \nabla \mathbf{v})_K \,, \qquad \mathbf{v} \in \mathbf{V}_h \cup \left(H^1(\Omega)\right)^2 \,.\end{aligned}$$

A basis in $\mathbf{V}_h$ can be constructed as follows: By the divergence formula (1.17), we see that

$$0 = \int_K \nabla \cdot \mathbf{v} \; d\mathbf{x} = \int_{\partial K} \mathbf{v} \cdot \boldsymbol{\nu} \; d\ell = \sum_{i=1}^{3} \mathbf{v} \cdot \boldsymbol{\nu}(\mathbf{m}_K^i)|e_i|, \quad K \in K_h \,, \tag{8.19}$$

where e_i is an edge of K, $\mathbf{m}_K^i$ is the midpoint of e_i, $|e_i|$ represents the length of e_i, and $\boldsymbol{\nu}$ is the outward unit normal to ∂K. The basis functions must satisfy property (8.19).

Let e be an edge in K_h. Denote by φ_e the piecewise linear function defined on K_h that is one at the midpoint of e and zero at the midpoints of all other edges in K_h (cf. Sect. 2.1.1). Define $\boldsymbol{\varphi}_e = \varphi_e \mathbf{t}_e$, where e is an internal edge of K_h and $\mathbf{t}_e$ is a unit vector tangential to e. This basis function satisfies (8.19).

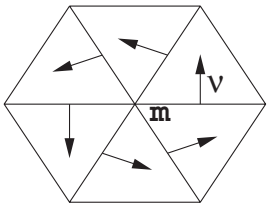

Fig. 8.1. An illustration of the normal direction on triangles

Let $\mathbf{m}$ be an internal vertex and let $e_1, e_2, \ldots, e_l$ be the edges in K_h that have $\mathbf{m}$ as a common vertex. Define

$$\boldsymbol{\varphi}_{\mathbf{m}} = \sum_{i=1}^{l} \frac{\varphi_{e_i}}{|e_i|} \boldsymbol{\nu}_{e_i} \ ,$$

where $\boldsymbol{\nu}_{e_i}$ is a unit vector normal to e_i pointing in the counterclockwise direction (cf. Fig. 8.1). Again, this function satisfies (8.19). It can be shown that a basis for $\mathbf{V}_h$ is given by the union of the two sets (cf. Exercise 8.7)

$$\{\boldsymbol{\varphi}_e : e \text{ is an internal edge in } K_h\}$$

and

$$\{\boldsymbol{\varphi}_{\mathbf{m}} : \mathbf{m} \text{ is an internal vertex in } K_h\} \ .$$

8.3.3.2 Nonconforming Finite Elements on Rectangles

We now consider the case where Ω is a rectangular domain and K_h is a regular partition of Ω into rectangles such that the horizontal and vertical edges of rectangles are parallel to the x_1- and x_2-coordinate axes, respectively. Define the nonconforming finite element space on rectangles (cf. Sect. 2.1.2)

$$\begin{aligned} \mathbf{V}_h = \Big\{ \mathbf{v} \in \left(L^2(\Omega)\right)^2 : \ & v_i|_K = a_K^{i,1} + a_K^{i,2} x_1 + a_K^{i,3} x_2 + a_K^{i,4}(x_1^2 - x_2^2), \\ & i = 1, 2, \ a_K^{i,j} \in \mathbb{R}; \text{ if } K_1 \text{ and } K_2 \text{ share an} \\ & \text{edge } e, \text{ then } \int_e \mathbf{v}|_{\partial K_1} \, d\ell = \int_e \mathbf{v}|_{\partial K_2} \, d\ell; \\ & \int_{e \cap \Gamma} \mathbf{v}|_e \, d\ell = 0; \ \nabla \cdot \mathbf{v} = 0 \text{ on all } K \in K_h \Big\} \ . \end{aligned}$$

That is, $\mathbf{V}_h$ is the nonconforming finite element space of rotated Q_1 functions that are divergence-free locally. With this space, the nonconforming method can be defined as in (8.17), and estimate (8.18) is satisfied.

A basis in $\mathbf{V}_h$ can be constructed similarly. By (1.17), we see that

$$0 = \int_K \nabla \cdot \mathbf{v} \, d\mathbf{x} = \int_{\partial K} \mathbf{v} \cdot \boldsymbol{\nu} \, d\ell = \sum_{i=1}^{4} \int_{e_i} \mathbf{v} \cdot \boldsymbol{\nu} \, d\ell, \quad K \in K_h \, . \tag{8.20}$$

The first kind of basis functions are associated with internal edges as in Sect. 8.3.3.1. Let e be an edge in K_h. Denote by φ_e the piecewise rotated Q_1 function defined on K_h such that the mean value of its integral on e equals one and it is zero on all other edges in K_h (cf. Sect. 2.1.2). Define $\boldsymbol{\varphi}_e = \varphi_e \mathbf{t}_e$, where e is an internal edge of K_h and $\mathbf{t}_e$ is a unit vector tangential to e. This basis function satisfies (8.20).

The second kind of basis functions are associated with internal vertices. Let $\mathbf{m}$ be an internal vertex and let e_1, e_2, e_3, and e_4 be the edges in K_h that have $\mathbf{m}$ as a common vertex. Define

$$\boldsymbol{\varphi}_{\mathbf{m}} = \sum_{i=1}^{4} \frac{\varphi_{e_i}}{|e_i|} \boldsymbol{\nu}_{e_i} \, ,$$

where $\boldsymbol{\nu}_{e_i}$ is a unit vector normal to e_i pointing in the counterclockwise direction (cf. Fig. 8.2). Again, this function satisfies (8.20). The basis functions for $\mathbf{V}_h$ consist of these two kinds of functions.

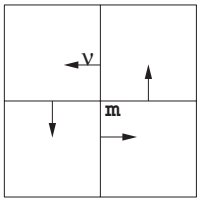

Fig. 8.2. An illustration of the normal direction on rectangles

8.4 The Navier-Stokes Equation

We make remarks on extensions of the development in the previous two sections to the Navier-Stokes equation

$$\begin{aligned} -\mu \Delta \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p &= \mathbf{f} && \text{in } \Omega \, , \\ \nabla \cdot \mathbf{u} &= 0 && \text{in } \Omega \, , \\ \mathbf{u} &= \mathbf{0} && \text{on } \Gamma \, . \end{aligned} \tag{8.21}$$

We introduce the *trilinear form*

$$a\left(\mathbf{w};\mathbf{u},\mathbf{v}\right)=\left((\mathbf{w}\cdot\nabla)\mathbf{u},\mathbf{v}\right)\,.$$

Then, in the same fashion as for (8.12), (8.21) can be recast in the mixed formulation: Find $\mathbf{u}\in\mathbf{V}$ and $p\in W$ such that

$$\begin{aligned} &a\left(\mathbf{u};\mathbf{u},\mathbf{v}\right)+\mu\left(\nabla\mathbf{u},\nabla\mathbf{v}\right)-\left(\nabla\cdot\mathbf{v},p\right)=\left(\mathbf{f},\mathbf{v}\right), && \mathbf{v}\in\mathbf{V}\,,\\ &\left(\nabla\cdot\mathbf{u},w\right)=0, && w\in W\,, \end{aligned} \tag{8.22}$$

where the spaces $\mathbf{V}$ and W are defined as in Sect. 8.2.2. System (8.22) can be shown to possess at least a solution $\mathbf{u}\in\mathbf{V}$ and $p\in W$. A proof of uniqueness of the solution requires some strong conditions on the trilinear form a, the function $\mathbf{f}$, and the viscosity μ (Girault-Raviart, 1981).

With an appropriate choice of the mixed finite element spaces $\mathbf{V}_h$ and W_h as in Sect. 8.3.2, the mixed finite element method for the Navier-Stokes problem is: Find $\mathbf{u}_h\in\mathbf{V}_h$ and $p_h\in W_h$ such that

$$\begin{aligned} &a\left(\mathbf{u}_h;\mathbf{u}_h,\mathbf{v}\right)+\mu\left(\nabla\mathbf{u}_h,\nabla\mathbf{v}\right)\\ &\qquad\qquad-\left(\nabla\cdot\mathbf{v},p_h\right)=\left(\mathbf{f},\mathbf{v}\right), && \mathbf{v}\in\mathbf{V}_h\,,\\ &\left(\nabla\cdot\mathbf{u}_h,w\right)=0, && w\in W_h\,. \end{aligned} \tag{8.23}$$

Again, under suitable conditions, problem (8.23) has a unique solution (Girault-Raviart, 1981). Note that (8.23) is a nonlinear system and can be solved using the solution techniques discussed in Sect. 1.8 such as the linearization and Newton methods.

8.5 Theoretical Considerations

As an example, we present an analysis for the mixed finite element method for the Stokes problem (8.10). We prove that the general theory for this method discussed in Sect. 3.8 applies.

We recall the spaces

$$\mathbf{V}=\left(H_0^1(\Omega)\right)^3,\qquad W=\left\{w\in L^2(\Omega):\int_\Omega w\;d\mathbf{x}=0\right\}\,.$$

We also introduce the bilinear forms $a(\cdot,\cdot):\mathbf{V}\times\mathbf{V}\to\mathbb{R}$ and $b(\cdot,\cdot):\mathbf{V}\times W\to\mathbb{R}$ as

$$\begin{aligned} &a(\mathbf{v},\mathbf{w})=\mu\left(\nabla\mathbf{v},\nabla\mathbf{w}\right), && \mathbf{v},\ \mathbf{w}\in\mathbf{V}\,,\\ &b(\mathbf{v},w)=-\left(\nabla\cdot\mathbf{v},w\right), && \mathbf{v}\in\mathbf{V},\ w\in W\,. \end{aligned}$$

Then (8.12) is rewritten: Find $\mathbf{u} \in \mathbf{V}$ and $p \in W$ such that

$$\begin{aligned} a(\mathbf{u}, \mathbf{v}) + b(\mathbf{v}, p) &= (\mathbf{f}, \mathbf{v}) , \qquad \mathbf{v} \in \mathbf{V} , \\ b(\mathbf{u}, w) &= 0, \qquad\qquad\quad w \in W . \end{aligned} \tag{8.24}$$

To apply the general theory in Sect. 3.8, set

$$\mathbf{Z} = \{\mathbf{v} \in \mathbf{V} : b(\mathbf{v}, w) = 0 \quad \forall w \in W\} .$$

By Poincaré's inequality (1.36), we see that $|\mathbf{v}|_{\mathbf{H}^1(\Omega)} = \sqrt{a(\mathbf{v}, \mathbf{v})}$ is a norm on $\mathbf{V}$. Hence the bilinear form $a(\cdot, \cdot)$ is $\mathbf{V}$-elliptic (thus $\mathbf{Z}$-elliptic). It remains to verify an *inf-sup* condition similar (3.67) for the bilinear form $b(\cdot, \cdot)$.

The next lemma is similar to Lemma 7.2. Its proof can be found in Arnold et al. (1988).

Lemma 8.1. *There exists a constant $C(\Omega) > 0$ such that for any $g \in L^2(\Omega)$, there is $\mathbf{v} \in \left(H^1(\Omega)\right)^2$ satisfying*

$$\nabla \cdot \mathbf{v} = g ,$$

and

$$\|\mathbf{v}\|_{\mathbf{H}^1(\Omega)} \le C(\Omega) \|g\|_{L^2(\Omega)} .$$

In addition, if $g \in W$, $\mathbf{v}$ can be chosen in $\left(H_0^1(\Omega)\right)^2$.

Theorem 8.2. *The bilinear form $b(\cdot, \cdot)$ satisfies*

$$\inf_{w \in W} \sup_{\mathbf{v} \in \mathbf{V}} \frac{b(\mathbf{v}, w)}{\|\mathbf{v}\|_{\mathbf{H}^1(\Omega)} \|w\|_{L^2(\Omega)}} \ge b_* , \tag{8.25}$$

where $b_* > 0$ is a constant; i.e., the inf-sup condition holds.

Proof. For $w \in W$, it follows from Lemma 8.1 that there exists $\mathbf{v} \in \left(H_0^1(\Omega)\right)^2$ such that $\nabla \cdot \mathbf{v} = -w$ and

$$\|\mathbf{v}\|_{\mathbf{H}^1(\Omega)} \le C(\Omega) \|w\|_{L^2(\Omega)} .$$

Then we see that

$$C(\Omega) \|w\|_{L^2(\Omega)} = C(\Omega) \frac{b(\mathbf{v}, w)}{\|w\|_{L^2(\Omega)}} \le \frac{b(\mathbf{v}, w)}{\|\mathbf{v}\|_{\mathbf{H}^1(\Omega)}} ,$$

which implies (8.25) with $b_* = C(\Omega)$. □

It follows from this theorem that Theorem 3.2 applies, and thus (8.24) (or (8.12)) has a unique solution.

Let $\mathbf{V}_h$ and W_h be defined as in Example 2 in Sect. 8.3.2.2 (i.e., the MINI element). The discrete counterpart of (8.24) is: Find $\mathbf{u}_h \in \mathbf{V}_h$ and $p_h \in W_h$ such that

$$\begin{aligned} a(\mathbf{u}_h, \mathbf{v}) + b(\mathbf{v}, p_h) &= (\mathbf{f}, \mathbf{v}), \qquad && \mathbf{v} \in \mathbf{V}_h , \\ b(\mathbf{u}_h, w) &= 0, && w \in W_h , \end{aligned} \tag{8.26}$$

which is (8.15). For the Stokes problem, we show that Theorem 3.6 applies.

Theorem 8.3. *Assume that K_h is a quasi-uniform triangulation of Ω (cf. (1.78)). For the MINI element, there exists a projection operator $\mathbf{\Pi}_h : \mathbf{V} \to \mathbf{V}_h$ such that*

$$b(\mathbf{v} - \mathbf{\Pi}_h \mathbf{v}, w) = 0 \qquad \forall w \in W_h . \tag{8.27}$$

Moreover, we have

$$\|\mathbf{\Pi}_h \mathbf{v}\|_{\mathbf{H}^1(\Omega)} \le C \|\mathbf{v}\|_{\mathbf{H}^1(\Omega)} , \tag{8.28}$$

where the constant C is independent of h.

Proof. Let

$$\mathbf{\Pi}_h^1 : \mathbf{V} \to \left\{ \mathbf{v} \in \left(H_0^1(\Omega)\right)^2 : \mathbf{v}|_K \in (P_1(K))^2 , \ K \in K_h \right\}$$

be the standard L^2-projection, with the following properties:

$$\|\mathbf{\Pi}_h^1 \mathbf{v}\|_{\mathbf{H}^1(\Omega)} \le C \|\mathbf{v}\|_{\mathbf{H}^1(\Omega)}, \tag{8.29}$$

and

$$\|\mathbf{v} - \mathbf{\Pi}_h^1 \mathbf{v}\|_{\mathbf{L}^2(\Omega)} \le Ch \|\mathbf{v}\|_{\mathbf{H}^1(\Omega)} . \tag{8.30}$$

Also, define $\mathbf{\Pi}_h^2 : \left(L^2(\Omega)\right)^2 \to (B_h)^2$ by

$$\int_K \left(\mathbf{v} - \mathbf{\Pi}_h^2 \mathbf{v}\right) \, d\mathbf{x} = \mathbf{0}, \qquad K \in K_h . \tag{8.31}$$

The projection operator $\mathbf{\Pi}_h^2$ is bounded in the L^2-norm:

$$\|\mathbf{\Pi}_h^2 \mathbf{v}\|_{\mathbf{L}^2(\Omega)} \le C \|\mathbf{v}\|_{\mathbf{L}^2(\Omega)} . \tag{8.32}$$

Now, for $\mathbf{v} \in \mathbf{V}$, we define

$$\mathbf{\Pi}_h \mathbf{v} = \mathbf{\Pi}_h^1 \mathbf{v} + \mathbf{\Pi}_h^2 (\mathbf{v} - \mathbf{\Pi}_h^1 \mathbf{v}) . \tag{8.33}$$

Then it follows from (8.31) that

$$\begin{aligned} \textstyle\int_K (\mathbf{v} - \mathbf{\Pi}_h \mathbf{v}) \, d\mathbf{x} \ &= \textstyle\int_K \left(\mathbf{I} - \mathbf{\Pi}_h^2\right) \left(\mathbf{v} - \mathbf{\Pi}_h^1 \mathbf{v}\right) \, d\mathbf{x} = \mathbf{0} , \\ & \qquad\qquad K \in K_h , \end{aligned} \tag{8.34}$$

where $\mathbf{I}$ is the identity operator. Because a function w in W_h is continuous and ∇w is piecewise constant, it follows from Green's formula (1.19) and (8.34) that

$$\begin{aligned} b(\mathbf{v} - \mathbf{\Pi}_h \mathbf{v}, w) &= -\left(\nabla \cdot [\mathbf{v} - \mathbf{\Pi}_h \mathbf{v}], w\right) \\ &= -\left([\mathbf{v} - \mathbf{\Pi}_h \mathbf{v}] \cdot \boldsymbol{\nu}, w\right)_\Gamma + (\mathbf{v} - \mathbf{\Pi}_h \mathbf{v}, \nabla w) \\ &= 0 \,, \end{aligned}$$

which implies (8.27). Next, it follows from (8.29), (8.30), (8.32), (8.33), and an inverse inequality (cf. (1.139)) that

$$\begin{aligned} \|\mathbf{\Pi}_h \mathbf{v}\|_{\mathbf{H}^1(\Omega)} &\le \|\mathbf{\Pi}_h^1 \mathbf{v}\|_{\mathbf{H}^1(\Omega)} + \|\mathbf{\Pi}_h^2 (\mathbf{v} - \mathbf{\Pi}_h^1 \mathbf{v})\|_{\mathbf{H}^1(\Omega)} \\ &\le C \left(\|\mathbf{v}\|_{\mathbf{H}^1(\Omega)} + h^{-1} \|\mathbf{\Pi}_h^2 (\mathbf{v} - \mathbf{\Pi}_h^1 \mathbf{v})\|_{\mathbf{L}^2(\Omega)} \right) \\ &\le C \left(\|\mathbf{v}\|_{\mathbf{H}^1(\Omega)} + h^{-1} \|\mathbf{v} - \mathbf{\Pi}_h^1 \mathbf{v}\|_{\mathbf{L}^2(\Omega)} \right) \\ &\le C \|\mathbf{v}\|_{\mathbf{H}^1(\Omega)} \,, \end{aligned}$$

which yields (8.28). □

Using Theorems 8.2 and 8.3, Theorems 3.2 and 3.6 apply. In particular, the discrete *inf-sup* condition holds:

$$\inf_{w \in W_h} \sup_{\mathbf{v} \in \mathbf{V}_h} \frac{b(\mathbf{v}, w)}{\|\mathbf{v}\|_{\mathbf{H}^1(\Omega)} \|w\|_{L^2(\Omega)}} \ge b_{**} > 0 \,, \tag{8.35}$$

where b_{**} is independent of h, and Theorem 3.4 implies that (8.26) (and thus (8.15)) has a unique solution. Also, applying Theorems 8.2 and 8.3, the error estimate (8.16) can be shown as in Sect. 3.8.5 (Arnold et al., 1984B).

8.6 Bibliographical Remarks

For more details on the mixed finite element method for the Stokes and Navier-Stokes equations considered in Sects. 8.3.2 and 8.4, the reader should refer to Girault-Raviart (1981). For the nonconforming finite element method on triangles for the Stokes problem described in Sect. 8.3.3.1, see Crouzeix-Raviart (1973); for the corresponding method on rectangles in Sect. 8.3.3.2, see Rannacher-Turek (1992). The proof of Theorem 8.3 follows Braess (1997). The book by Glowinski (2003) gives a thorough treatment of the Navier-Stokes equation.

8.7 Exercises

8.1. Defining the material derivative

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \mathbf{u} \cdot \nabla \,,$$

derive (8.2) from (8.1) in detail.

8.2. Prove that the Stokes equation (8.10) in a simply connected domain $\Omega \subset \mathbb{R}^2$ can be written as the biharmonic problem (1.57) by introducing a suitable stream function as an unknown.

8.3. Show that problem (8.11) has a unique solution $\mathbf{u} \in \mathbf{V}$ (cf. Sect. 1.9).

8.4. Derive (8.12) from (8.10) in detail.

8.5. Show that the discrete problem (8.14) has a unique solution $\mathbf{u}_h \in \mathbf{V}_h$ (cf. Sect. 1.9).

8.6. With the nonconforming finite element space $\mathbf{V}_h$ defined in Sect. 8.3.3.1 or Sect. 8.3.3.2, show that problem (8.17) possesses a unique solution $\mathbf{u}_h \in \mathbf{V}_h$.

8.7. Let $\mathbf{V}_h$ be the P_1-nonconforming finite element space defined in Sect. 8.3.3.1. Prove that a basis for $\mathbf{V}_h$ is given by the union of the two sets

$$\{\boldsymbol{\varphi}_e : e \text{ is an internal edge in } K_h\}$$

and

$$\{\boldsymbol{\varphi}_{\mathbf{m}} : \mathbf{m} \text{ is an internal vertex in } K_h\} ,$$

where the functions $\boldsymbol{\varphi}_e$ and $\boldsymbol{\varphi}_{\mathbf{m}}$ are defined as in Sect. 8.3.3.1.

8.8. Formulate a Stokes problem with a suitable right-hand side to prove that for every function $g \in L^2(\Omega)$, there is $\mathbf{u} \in \mathbf{H}_0^1(\Omega)$ such that

$$\nabla \cdot \mathbf{u} = g \quad \text{and} \quad \|\mathbf{u}\|_{\mathbf{H}^1(\Omega)} \leq C \|g\|_{L^2(\Omega)} ,$$

where C is independent of g.

8.9. In this chapter, we have developed the finite element methods only for the stationary Stokes and Navier-Stokes equations. The corresponding

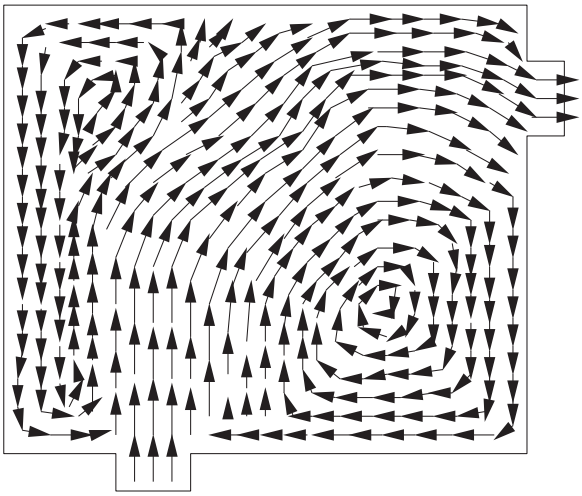

Fig. 8.3. A numerical cavity problem

transient equations can be treated using the techniques of this chapter with those in Sect. 1.7. As an example, develop the nonconforming finite element method discussed in Sect. 8.3.3.1, with the backward Euler scheme (cf. Sect. 1.7) for the time derivative, for the transient Navier-Stokes equation

$$\begin{aligned}
&\frac{\partial \mathbf{u}}{\partial t} - \mu \Delta \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p = \mathbf{0} && \text{in } \Omega\,, \\
&\nabla \cdot \mathbf{u} = 0 && \text{in } \Omega\,, \\
&\mathbf{u} \cdot \boldsymbol{\nu} = \mathbf{u} \cdot \mathbf{t} = 0 && \text{on } \Gamma\,, \\
&\mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0 && \text{in } \Omega\,,
\end{aligned}$$

where $\Omega \subset \mathbb{R}^2$. A numerical cavity problem is presented in Fig. 8.3. The initial velocity $\mathbf{u}_0$ is zero, the inlet velocity is one, and the viscosity μ equals 10^{-3}. An example of the computed velocities after 10 time steps is given in this figure on a 20×10 grid (triangles constructed as in Fig. 1.7) where $\Delta t = h$ is used.

9 Fluid Flow in Porous Media

The basic problems to be addressed in modeling and simulation of fluid flows in both petroleum and ground water reservoirs are analogous. In this chapter, as an example, we focus on petroleum reservoirs.

A petroleum reservoir is a porous medium which contains a number of hydrocarbons. The primary goal of reservoir simulation is to predict future performance of a reservoir and find the ways and means of optimizing the recovery of some of the hydrocarbons.

There are two important characteristics of a petroleum reservoir, the nature of the rock and that of the fluids filling it. A reservoir is usually heterogeneous; its properties heavily depend on the location. For example, a *fractured reservoir* is heterogeneous; it consists of a set of blocks of porous media (the *matrix*) and a net of fractures. The rock properties in such a reservoir vary dramatically; its permeability may vary from one in the matrix to thousands in the fractures, for example. While the governing equations for the fractured reservoir are similar to those for an ordinary reservoir, they pose additional difficulties in simulation.

The nature of the fluids filling a petroleum reservoir strongly depends on the stage of *oil recovery*. In the very early stage, the reservoir usually contains a single fluid such as gas or oil. Often the pressure at this stage is so high that the fluid is produced by simple natural decompression without any pumping effort at wells. This stage is referred to as *primary recovery*, and it ends when a pressure equilibrium between the oil field and the atmosphere is reached. Often primary recovery leaves 70–85% of hydrocarbons in the reservoir.

To recover part of the remaining oil, a fluid (usually water) is injected into some wells (*injection wells*) while oil is produced at other wells (*production wells*). This process serves to maintain the high reservoir pressure and flow rates. It also displaces some of the oil and pushes it toward the production wells. This stage of oil recovery is called *secondary recovery* or *water flooding*.

In the secondary recovery, if the pressure is above a bubble point pressure of the oil phase, the flow is two-phase immiscible, one phase being water and the other being oil, without mass transfer between the phases. If the pressure drops below the bubble point pressure, then the oil (more precisely, the hydrocarbon phase) is split into a liquid phase and a gaseous phase in thermodynamical equilibrium. In this case, the flow is of the *black-oil* type;

the water phase does not exchange mass with other two phases, and the liquid and gaseous phases exchange mass between them.

Water flooding is still not very effective and 50% or more of hydrocarbons often remain in a reservoir. Due to strong surface tension, a large amount of oil is trapped in small pores and cannot be washed out with this technique. Also, when oil is heavy and viscous, water is extremely mobile. If the flow rate is sufficiently high, instead of producing oil, the production wells primarily produce water.

To recover more of the hydrocarbons, several enhanced recovery techniques have been developed. These techniques involve complex chemical and thermal effects and are termed *tertiary recovery* or *enhanced recovery*. There are many different versions of enhanced recovery techniques, but one of the major objectives of these techniques is to achieve miscibility and thus to eliminate the residual oil saturation. Miscibility is achieved by increasing temperature (e.g., in-situ combustion) or by injecting other chemical species like carbon dioxide. A typical flow in enhanced recovery is *compositional flow*, where only the number of chemical species is given a-priori, and the number of phases and the composition of each phase in terms of the given species depend on the thermodynamical conditions and the overall concentration of each species.

In this chapter, as an example, we consider two-phase flow in a porous medium. Single phase flow is simpler, and other more complex flows (e.g., three-phase and compositional flow) can be also handled (Chen-Ewing, 1997A; Chen, 2002; Chen et al., 2000). In Sect. 9.1, we state the governing equations for two-phase flow and their variants defined in terms of pressure and saturation. In Sect. 9.2, we apply the mixed finite element method for the solution of the pressure equation. Then, in Sect. 9.3, we employ the characteristic finite element method to solve the saturation equation. In Sect. 9.4, we present a numerical example. Section 9.5 is devoted to theoretical considerations. Finally, in Sect. 9.6, we give bibliographical information.

9.1 Two-Phase Immiscible Flow

In this section, we consider two-phase flow where the fluids are incompressible and immiscible and there is no mass transfer between them. One phase wets a porous medium more than the other, is called the wetting phase, and is indicated by a subscript w. The other phase is termed the nonwetting phase, and is represented by o.

In a porous medium $\Omega \subset \mathbb{R}^d$ $(d \leq 3)$, the mass balance equation for each of the fluid phases is (Peaceman, 1977B; Aziz-Settari, 1979):

$$\phi \frac{\partial(\rho_\alpha s_\alpha)}{\partial t} + \nabla \cdot (\rho_\alpha \mathbf{u}_\alpha) = \rho_\alpha q_\alpha, \quad \alpha = w, o \,, \tag{9.1}$$

where $\alpha = w$ denotes the wetting phase (e.g., water), $\alpha = o$ indicates the nonwetting phase (e.g., oil), ϕ is the porosity of the reservoir, and ρ_α, s_α, $\mathbf{u}_\alpha$, and q_α are, respectively, the density, saturation, volumetric velocity, and external volumetric flow rate of the α-phase. The volumetric velocity $\mathbf{u}_\alpha$ is given by Darcy's law

$$\mathbf{u}_\alpha = -\frac{\boldsymbol{\kappa}\kappa_{r\alpha}}{\mu_\alpha}(\nabla p_\alpha + \rho_\alpha g \nabla Z), \quad \alpha = w, o \,, \tag{9.2}$$

where $\boldsymbol{\kappa}$ is the absolute permeability of the reservoir, p_α, μ_α, and $\kappa_{r\alpha}$ are the pressure, viscosity, and relative permeability of the α-phase, respectively, g denotes the gravitational constant, Z is the depth, and the x_3-coordinate (or the z-coordinate) is in the vertical upward direction. In addition to (9.1) and (9.2), the customary property for the saturations is

$$s_w + s_o = 1 \,, \tag{9.3}$$

and the two pressures are related by the capillary pressure function

$$p_c(s_w) = p_o - p_w \,. \tag{9.4}$$

Finally, we define the sink/source term q_α in (9.1) by

$$q_\alpha = \sum_l q_\alpha^{(l)} \delta(\mathbf{x} - \mathbf{x}^{(l)}), \qquad \alpha = w, o \,,$$

where $q_\alpha^{(l)}$ indicates the volume of the fluid produced or injected per unit time at the lth well, $\mathbf{x}^{(l)}$, for phase α and δ is the Dirac delta function. Following Peaceman (1977A), $q_\alpha^{(l)}$ can be defined by

$$q_\alpha^{(l)} = \frac{2\pi \bar{\kappa}^{(l)} \kappa_{r\alpha} \Delta L^{(l)}}{\mu_\alpha \ln \dfrac{r_e^{(l)}}{r_c^{(l)}}} \left(p^{(l)} - p_\alpha + \rho_\alpha g (Z^{(l)} - Z) \right) \,, \tag{9.5}$$

where $p^{(l)}$ is the flowing bottom hole pressure at depth $Z^{(l)}$, $\Delta L^{(l)}$, $r_e^{(l)}$, and $r_c^{(l)}$ are, respectively, the length, the equivalent radius, and the radius of the lth well, and the quantity $\bar{\kappa}^{(l)}$ is some average of $\boldsymbol{\kappa}$ at the lth well (Peaceman, 1991). To devise suitable numerical methods for solving (9.1)–(9.5), these equations will be rewritten in various formulations below.

A typical curve of the capillary pressure p_c is shown in Fig. 9.1. The capillary pressure depends on the wetting phase saturation and the direction of saturation change (drainage or imbibition). The phenomenon of dependence of the curve on the history of saturation is called *hysteresis*.

Typical curves of relative permeabilities κ_{rw} and κ_{ro} suitable for an oil-water system with water displacing oil are presented in Fig. 9.2. The value of s_w at which water starts to flow is termed the *critical saturation*, s_{wc}, and the value s_{nc} at which oil ceases to flow is called the *residual saturation*. Analogously, during a drainage cycle s_{nc} and s_{wc} are referred to as the critical and residual saturations, respectively. Hysteresis can also occur in relative permeabilities as in capillary pressures.

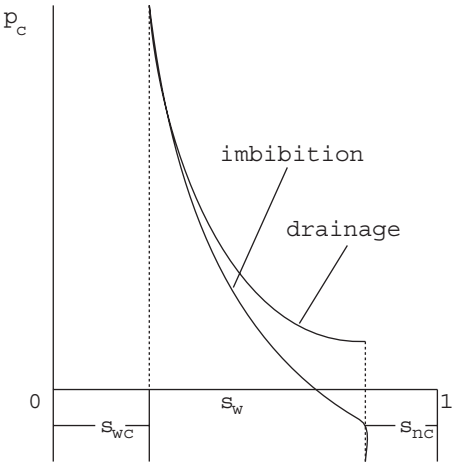

Fig. 9.1. Typical capillary pressure curve

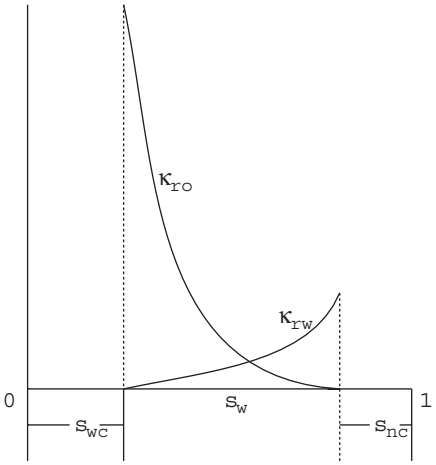

Fig. 9.2. Typical relative permeability curves

9.1.1 The Phase Formulation

We introduce the phase mobilities

$$\lambda_\alpha(s_\alpha) = \kappa_{r\alpha}/\mu_\alpha, \quad \alpha = w, o\ ,$$

and the total mobility

$$\lambda(s) = \lambda_w + \lambda_o\ ,$$

where $s = s_w$. The fractional flow functions are defined by

$$f_\alpha(s) = \lambda_\alpha/\lambda, \quad \alpha = w, o\ .$$

We use the oil phase pressure as the pressure variable in this subsection

$$p = p_o \ , \tag{9.6}$$

and define the total velocity by

$$\mathbf{u} = \mathbf{u}_w + \mathbf{u}_o \ . \tag{9.7}$$

Under the assumption that the fluids are incompressible (i.e., the phase densities are assumed constant; see Chap. 8), we apply (9.3) and (9.7) to (9.1) to see that (cf. Exercise 9.1)

$$\nabla \cdot \mathbf{u} = q(p,s) \equiv q_w + q_o \ , \tag{9.8}$$

and (9.4) and (9.7) to (9.2) to obtain

$$\mathbf{u} = -\boldsymbol{\kappa}\big\{\lambda(s)\nabla p - \lambda_w(s)\nabla p_c + \big(\lambda_w\rho_w + \lambda_o\rho_o\big)g\nabla Z\big\} \ . \tag{9.9}$$

Similarly, apply (9.4), (9.7), and (9.9) to (9.1) and (9.2) with $\alpha = w$ to have

$$\begin{aligned} \phi\frac{\partial s}{\partial t} + \nabla \cdot \Bigg\{ \boldsymbol{\kappa} f_w(s)\lambda_o(s)\Bigg[\frac{dp_c}{ds}\nabla s \\ + \ (\rho_o - \rho_w)g\nabla Z\Bigg] + f_w(s)\mathbf{u}\Bigg\} = q_w(p,s) \ . \end{aligned} \tag{9.10}$$

In (9.8) and (9.10), the well terms are now defined in terms of the phase pressure p and saturation s:

$$q_o^{(l)}(p,s) = \frac{2\pi\bar{\kappa}^{(l)}\kappa_{ro}\Delta L^{(l)}}{\mu_o \ln\dfrac{r_e^{(l)}}{r_c^{(l)}}}\Big(p^{(l)} - p + \rho_o g(Z^{(l)} - Z)\Big) \ , \tag{9.11a}$$

and

$$q_w^{(l)}(p,s) = \frac{2\pi\bar{\kappa}^{(l)}\kappa_{rw}\Delta L^{(l)}}{\mu_w \ln\dfrac{r_e^{(l)}}{r_c^{(l)}}}\Big(p^{(l)} - p + p_c + \rho_w g(Z^{(l)} - Z)\Big) \ . \tag{9.11b}$$

The pressure equation is given by (9.8) and (9.9), while the saturation equation is described by (9.10). They determine the main unknowns p, $\mathbf{u}$, and s. While the phase mobilities λ_α can be zero (cf. Fig. 9.2), the total mobility λ is always positive, so the pressure equation is elliptic. If one of the densities varies, this equation becomes parabolic. The saturation equation is parabolic for s, and it is degenerate in the sense that the capillary diffusion coefficient $f_w\lambda_o dp_c/ds$ can be zero. Furthermore, this equation becomes hyperbolic if the capillary pressure is ignored. The mathematical properties of this system such as existence, uniqueness, regularity, and asymptotic behavior of solutions have been studied (Chen, 2001B; 2002A).

9.1.2 The Weighted Formulation

We now introduce a smoother pressure than the phase pressure, i.e., a weighted pressure

$$p = s_w p_w + s_o p_o \ . \tag{9.12}$$

Note that even if a saturation is zero (i.e., a phase disappears), we still have a non-zero smooth variable p. The total velocity is defined as in (9.7), and (9.8) remains the same:

$$\nabla \cdot \mathbf{u} = q(p, s) \ . \tag{9.13}$$

Now, apply (9.3), (9.4), (9.7), and (9.12) to (9.2) to see that (cf. Exercise 9.2)

$$\begin{aligned} \mathbf{u} = -\boldsymbol{\kappa}\Big\{ & \lambda(s)\nabla p + \big(s\lambda(s) - \lambda_w(s)\big)\nabla p_c \\ & + \lambda(s) p_c \nabla s + \big(\lambda_w \rho_w + \lambda_o \rho_o\big) g \nabla Z \Big\}. \end{aligned} \tag{9.14}$$

The saturation (9.10) is the same:

$$\begin{aligned} \phi \frac{\partial s}{\partial t} + \nabla \cdot \Bigg\{ \boldsymbol{\kappa} f_w(s) \lambda_o(s) \Bigg[& \frac{dp_c}{ds} \nabla s \\ & + (\rho_o - \rho_w) g \nabla Z \Bigg] + f_w(s) \mathbf{u} \Bigg\} = q_w(p, s) \ . \end{aligned} \tag{9.15}$$

In (9.13) and (9.15), the well terms are evaluated using the weighted pressure. Now, the pressure equation consists of (9.13) and (9.14), and the saturation equation is (9.15).

9.1.3 The Global Formulation

Note that p_c appears in both (9.9) and (9.14). To eliminate it, following Antontsev (1972) and Chavent-Jaffré (1978), we define the global pressure

$$p = p_o - \int^s \left(f_w \frac{dp_c}{ds} \right) (\xi) \mathrm{d}\xi \ . \tag{9.16}$$

Again, (9.8) remains the same:

$$\nabla \cdot \mathbf{u} = q(p, s) \ . \tag{9.17}$$

Now, apply (9.4), (9.7), and (9.16) to (9.2) to obtain (cf. Exercise 9.3)

$$\mathbf{u} = -\boldsymbol{\kappa} \left\{ \lambda(s) \nabla p + \big(\lambda_w \rho_w + \lambda_o \rho_o\big) g \nabla Z \right\} \ . \tag{9.18}$$

The pressure equation is given by (9.17) and (9.18). The saturation equation is the same as previously:

$$\phi\frac{\partial s}{\partial t} + \nabla\cdot\left\{\boldsymbol{\kappa} f_w(s)\lambda_o(s)\left[\frac{dp_c}{ds}\nabla s + (\rho_o-\rho_w)g\nabla Z\right] + f_w(s)\mathbf{u}\right\} = q_w(p,s)\ . \tag{9.19}$$

In (9.17) and (9.19), the well terms are of the same form as in (9.11) with p now being the global pressure.

It follows from (9.4) and (9.16) that

$$\lambda\nabla p = \lambda_w\nabla p_w + \lambda_o\nabla p_o\ .$$

This implies that the global pressure is the pressure that would produce a flow of a fluid (with mobility λ) equal to the sum of the flows of fluids w and o.

The total velocity is used in all three formulations. This velocity is smoother than the phase velocities $\mathbf{u}_\alpha$, $\alpha = w, o$. As noted, the capillary pressure p_c appears in the phase and weighted formulations, but does not appear in the global formulation. Thus the coupling between the pressure and saturation equations in the latter formulation is less than that in the former two formulations. When p_c is ignored, all three formulations are the same. In this case, the saturation equation becomes the classical Buckley-Leverett equation where the flux function f_w is generally nonconvex over the range of saturation values where this function is nonzero (Aziz-Settari, 1979). A numerical comparison between these formulations is given in Sect. 9.4.

9.2 Mixed Finite Elements for Pressure

In this and next sections, we present numerical methods for solving the pressure and saturation equations developed in the previous section. As an example, we present them for the global formulation. The model in this formulation is completed by specifying boundary and initial conditions. For simplicity, in subsequent sections, *no-flow boundary conditions* are used:

$$\begin{aligned} &\mathbf{u}\cdot\boldsymbol{\nu} = 0, && \mathbf{x}\in\Gamma\ ,\\ &\boldsymbol{\kappa} f_w(s)\lambda_o(s)\left(\frac{dp_c}{ds}\nabla s + (\rho_o-\rho_w)g\nabla Z\right)\cdot\boldsymbol{\nu} = 0, && \mathbf{x}\in\Gamma\ , \end{aligned} \tag{9.20}$$

where $\boldsymbol{\nu}$ is the outward unit normal to the boundary Γ of Ω. These boundary conditions are derived from those for phase velocities (Chen et al., 1995). The initial condition is given by

$$s(\mathbf{x},0) = s_0(\mathbf{x}), \qquad \mathbf{x}\in\Omega\ .$$

By (9.17) and the first equation of (9.20), compatibility to incompressibility of the fluids requires

$$\int_\Omega q \, d\mathbf{x} = 0, \qquad t \geq 0 \, .$$

The saturation equation (9.19) depends on the pressure p explicitly through the velocity $\mathbf{u}$. Also, physical transport generally dominates diffusion in two-phase flow. These two facts suggest that obtaining an accurate approximate velocity be important. This motivates the use of the mixed finite element method in the computation of pressure and velocity (Chavent-Jaffré, 1978; Douglas et al., 1983).

Set (cf. Sect. 3.2)

$$\mathbf{V} = \{\mathbf{v} \in \mathbf{H}(\mathrm{div}, \Omega) : \mathbf{v} \cdot \boldsymbol{\nu} = 0 \text{ on } \Gamma\}, \quad W = L^2(\Omega) \, .$$

For simplicity, let Ω be a convex polygonal domain. For $0 < h < 1$, let K_h be a regular partition of Ω into elements, say, tetrahedra, rectangular parallelepipeds, or prisms, with maximum mesh size h (cf. Sect. 3.4). The partition for pressure is not necessarily the same as that for saturation. For notational convenience, we simply use the same partition.

Associated with the partition K_h, let $\mathbf{V}_h \times W_h \subset \mathbf{V} \times W$ be the Raviart-Thomas-Nedelec (1977, 1980), Brezzi et al. (if $d = 2$; 1985), Brezzi et al. (if $d = 3$; 1987A), Brezzi et al. (1987B), or Chen-Douglas (1989) mixed finite element space; see Sect. 3.4.

Let $J = (0, T]$ be the time interval of interest. In petroleum reservoir simulations using two-phase flow, pressure changes less rapidly in time than saturation. Thus it is appropriate to take a much longer time step for the former than for the latter. For each positive integer N, let $0 = t^0 < t^1 < \cdots < t^N = T$ be a partition of J for pressure into subintervals $J^n = (t^{n-1}, t^n]$, with length $\Delta t_p^n = t^n - t^{n-1}$. We may vary Δt_p, but except for Δt_p^1 we drop the superscript. The subinterval J^n is divided into sub-subintervals for saturation:

$$t^{n-1,m} = t^{n-1} + m\Delta t_p^n / M^n, \qquad m = 1, 2, \ldots, M^n \, .$$

The number of steps, M^n, can depend on n. Below we simply write $t^{n-1,0} = t^{n-1}$, and set $v^{n,m} = v(\cdot, t^{n,m})$.

Now, the mixed method for (9.17) and (9.18) is given as follows: For any $0 \leq n \leq N$, find $\mathbf{u}_h^n \in \mathbf{V}_h$ and $p_h^n \in W_h$ such that

$$\begin{aligned} &(\nabla \cdot \mathbf{u}_h^n, w) = \big(q\big(p_h^n, s_h^n\big), w\big) \, , && w \in W_h \, , \\ &\Big(\big(\boldsymbol{\kappa}\lambda(s_h^n)\big)^{-1}\mathbf{u}_h^n, \mathbf{v}\Big) - (p_h^n, \nabla \cdot \mathbf{v}) = -\left(\boldsymbol{\gamma}_{\mathbf{1}}(s_h^n), \mathbf{v}\right) , && \mathbf{v} \in \mathbf{V}_h \, , \end{aligned} \tag{9.21}$$

where

$$\boldsymbol{\gamma}_{\mathbf{1}}(s) = \big(f_w\rho_w + f_o\rho_o\big)g\nabla Z \, .$$

9.3 Characteristic Methods for Saturation

As noted earlier, physical transport dominates diffusive effects in incompressible flow in petroleum reservoirs, and the capillary diffusion coefficient in the saturation equation (9.19) can be zero. Thus it is appropriate to use the characteristic finite element method introduced in Chap. 5 to solve this equation. As an example, we present the MMOC procedure in this section; other characteristic procedures can be similarly described as in Chap. 5.

We introduce the notation

$$q_1(p,s) = q_w(p,s) - q(p,s)f_w(s) - \nabla\cdot(\boldsymbol{\kappa} f_w(s)\lambda_o(s)(\rho_o-\rho_w)g\nabla Z)\ .$$

Then, using (9.17) and (9.19), the saturation equation can be written as follows:

$$\phi\frac{\partial s}{\partial t} + \frac{df_w}{ds}\mathbf{u}\cdot\nabla s + \nabla\cdot\left\{\boldsymbol{\kappa} f_w(s)\lambda_o(s)\frac{dp_c}{ds}\nabla s\right\} = q_1(p,s)\ . \tag{9.22}$$

Let

$$\mathbf{b}(\mathbf{x},t) = \frac{df_w}{ds}\mathbf{u}, \quad \psi(\mathbf{x},t) = \left(\phi^2(\mathbf{x}) + \|\mathbf{b}(\mathbf{x},t)\|^2\right)^{1/2}\ ,$$

and let the characteristic direction associated with the operator $\phi\frac{\partial s}{\partial t} + \mathbf{b}\cdot\nabla s$ be denoted by $\boldsymbol{\tau}(\mathbf{x},t)$, so

$$\frac{\partial}{\partial\boldsymbol{\tau}} = \frac{\phi(\mathbf{x})}{\psi(\mathbf{x},t)}\frac{\partial}{\partial t} + \frac{\mathbf{b}(\mathbf{x},t)}{\psi(\mathbf{x},t)}\cdot\nabla\ .$$

Then (9.22) becomes

$$\psi\frac{\partial s}{\partial\boldsymbol{\tau}} + \nabla\cdot\left\{\boldsymbol{\kappa} f_w(s)\lambda_o(s)\frac{dp_c}{ds}\nabla s\right\} = q_1(p,s)\ . \tag{9.23}$$

Note that the characteristic direction $\boldsymbol{\tau}$ depends on the velocity $\mathbf{u}$. Since the saturation step $t^{n-1,m}$ relates to pressure steps by $t^{n-1} < t^{n-1,m} \le t^n$, we need a velocity approximation for (9.23) based on $\mathbf{u}_h^{n-1}$ and earlier values. For this, we utilize a *linear extrapolation approach* (cf. Sect. 5.6): If $n \ge 2$, take the linear extrapolation of $\mathbf{u}_h^{n-2}$ and $\mathbf{u}_h^{n-1}$ determined by

$$E\mathbf{u}_h^{n-1,m} = \left(1 + \frac{t^{n-1,m}-t^{n-1}}{t^{n-1}-t^{n-2}}\right)\mathbf{u}_h^{n-1} - \frac{t^{n-1,m}-t^{n-1}}{t^{n-1}-t^{n-2}}\mathbf{u}_h^{n-2}\ .$$

For $n = 1$, define

$$E\mathbf{u}_h^{0,m} = \mathbf{u}_h^0\ .$$

$E\mathbf{u}_h^{n-1,m}$ is first-order accurate in time in the first pressure step and second-order accurate in the later steps.

The MMOC procedure is generally defined with periodic boundary conditions (cf. Sect. 5.2). For this reason, we assume that Ω is a rectangular domain,

and all functions in (9.23) are spatially Ω-periodic. Let $M_h \subset H^1(\Omega)$ be any finite element space introduced in Chap. 1. Then the MMOC procedure for (9.23) is defined as follows: For each $0 \le n \le N$ and $1 \le m \le M^n$, find $s_h^{n,m} \in M_h$ such that

$$\begin{aligned}\left(\phi\frac{s_h^{n,m}-\check{s}_h^{n,m-1}}{t^{n,m}-t^{n,m-1}},w\right)+\left(\mathbf{a}\left(s_h^{n,m-1}\right)\nabla s_h^{n,m},\nabla w\right)\\ =\left(q_1\left(p_h^n,s_h^{n,m-1}\right),w\right),\qquad w\in M_h\ ,\end{aligned}\tag{9.24}$$

where

$$\mathbf{a}(s)=-\boldsymbol{\kappa}f_w(s)\lambda_o(s)\frac{dp_c}{ds}\ ,$$

$$\check{s}_h^{n,m-1}=s_h^{n,m-1}\left(\mathbf{x}-\frac{df_w}{ds}\left(s_h^{n,m-1}\right)\frac{E\mathbf{u}_h^{n,m}}{\phi(\mathbf{x})}\Delta t_s^{n,m},t^{n,m-1}\right)\ ,$$

with $\Delta t_s^{n,m}=t^{n,m}-t^{n,m-1}$. The initial approximate solution s_h^0 can be defined as any appropriate projection of s_0 in M_h (e.g., the L^2-projection of s_0 in M_h).

Equations (9.21) and (9.24) can be solved as follows: After startup for s_h^0, we obtain $(\mathbf{u}_h^0,p_h^0)$ from (9.21) and then $s_h^{0,m}$, $m=1,2,\ldots,M^0$, from (9.24); this process proceeds in a sequential fashion. Other solution approaches such as the IMPES (*implicit pressure-explicit saturation* (Sheldon et al., 1959)) and *simultaneous solution* approaches (Douglas et al., 1959) can be also presented.

9.4 A Numerical Example

In this section, we present a numerical comparison between the three formulations developed in Sect. 9.1. The porous medium is two-dimensional, with dimensions 1,000 ft × 1,000 ft. The relative permeability curves are

$$\kappa_{rw}=\kappa_{rwm}\left(\frac{s_w-s_{wc}}{1-s_{or}-s_{wc}}\right)^2,\quad \kappa_{ro}=\left(\frac{s_o-s_{or}}{1-s_{or}-s_{wc}}\right)^2\ ,\tag{9.25}$$

where $\kappa_{rwm}=0.65$, $s_{wc}=0.22$, and $s_{or}=0.2$. Other physical data are chosen as follows:

$$\phi=0.2,\quad \mu_w=0.096\text{ cp},\quad \mu_o=1.14\text{ cp}\ .\tag{9.26}$$

The example considered is in a five spot pattern: An injection well is located at a corner of the reservoir, and a production well is located at its opposite corner. Water is injected, and oil and/or water is produced. In addition to the above data, we also need

$$\boldsymbol{\kappa} = 0.1\mathbf{I} \text{ darcy}, \quad r_c = 0.2291667 \text{ ft}, \quad s_0 = s_{wc} \,, \tag{9.27}$$

where $\mathbf{I}$ is the identity matrix. The flowing bottom hole pressure is $3,700$ psi at the injection well and $3,500$ psi at the production well. Finally, the capillary curve is given by

$$p_c = p_{cmin} + (p_{cmax} - p_{cmin}) \frac{1 - s_w}{1 - s_{wc}} \,, \tag{9.28}$$

where $p_{cmin} = 0$ psi and $p_{cmax} = 70$ psi.

In the computations, we employ the lowest-order Raviart-Thomas mixed finite elements on triangles on a 10×10 grid (triangles constructed as in Fig. 1.7) to solve the pressure equation for all three formulations. On the same grid, the MMOC procedure is used to solve the saturation equation; the finite element space used in this procedure is composed of continuous piecewise linear functions. The oil and water production versus time (RB/day), the characterization curve of displacement (percent), and the oil recovery curve (percent) are shown in Figs. 9.3–9.5, where $-\cdot-$, #, and $-\circ-$ denote the phase, weighted, and global formulations, respectively. The characterization curve is defined as the logarithm of the cumulative water production versus the cumulative oil production. From these figures we see that the numerical results of the global and phase formulations match very well. This is probably due to the fact that the global form resembles the phase form more.

We also check the CPU (Central Processing Unit) times (in seconds) for the three formulations at the final time, $T = 8{,}000$ days; the results performed on a Dec Alpha workstation are displaced in Table 9.1. There is

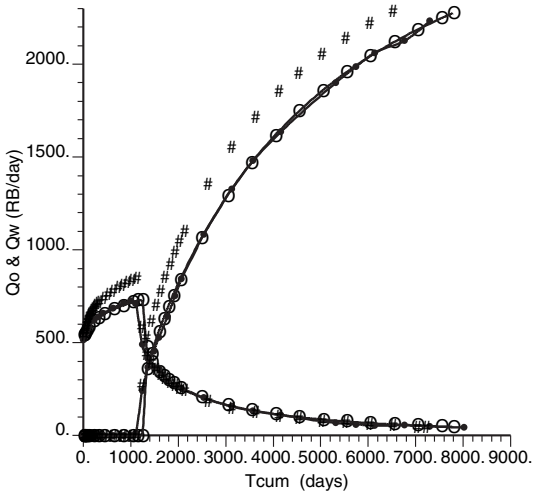

Fig. 9.3. Water (*upper*) and oil (*below*) productions

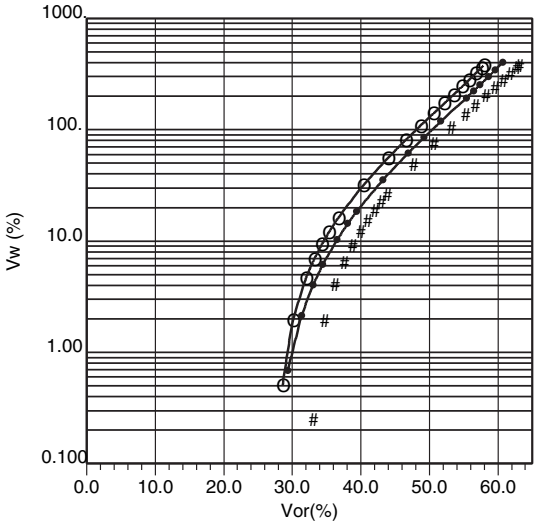

Fig. 9.4. Characterization curves of displacement

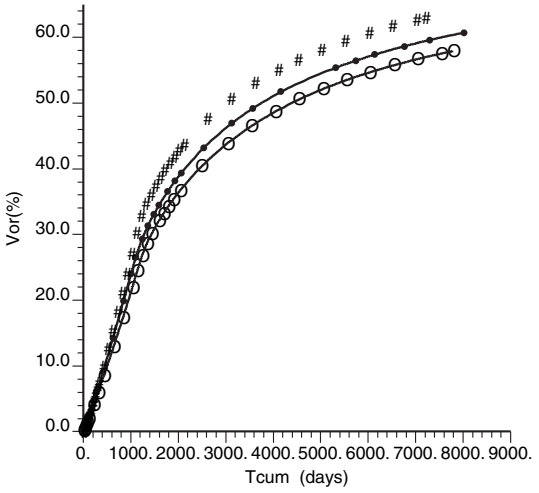

Fig. 9.5. Oil recovery curves

Table 9.1. CPU times for three formulations

	Global	Phase	Weighted
CPU times	38.6252	38.3705	38.7518

not much difference between the CPU times for this example. It appears that the three forms for two-phase flow do not differ much from the computational perspective. In terms of mathematical and numerical analysis, researchers have preferred to use the global form since this form has the least coupling between the pressure and saturation equations and is easiest to analyze.

9.5 Theoretical Considerations

We give a theoretical analysis for the system of (9.21) and (9.24).

9.5.1 Analysis for the Pressure Equation

We recall the approximation properties of the RTN, BDM, BDFM, BDDF, and CD mixed finite element spaces (cf. Sect. 3.5):

$$\begin{aligned}
&\inf_{\mathbf{v}_h\in\mathbf{V}_h} \|\mathbf{v}-\mathbf{v}_h\|_{\mathbf{L}^2(\Omega)} \le Ch^l\|\mathbf{v}\|_{\mathbf{H}^l(\Omega)}, \quad 1\le l\le r+1\ ,\\
&\inf_{\mathbf{v}_h\in\mathbf{V}_h} \|\nabla\cdot(\mathbf{v}-\mathbf{v}_h)\|_{L^2(\Omega)} \le Ch^l\|\nabla\cdot\mathbf{v}\|_{H^l(\Omega)}, \quad 0\le l\le r^*\ ,\\
&\inf_{w_h\in W_h} \|w-w_h\|_{L^2(\Omega)} \le Ch^l\|w\|_{H^l(\Omega)}, \quad 0\le l\le r^*\ ,
\end{aligned} \tag{9.29}$$

where $r^* = r+1$ for the RTN, BDFM, and first and third CD spaces and $r^* = r$ for the BDM, BDDF, and second CD spaces. Also, each of these spaces possesses the property that there are projection operators $\mathbf{\Pi}_h : (H^1(\Omega))^d \to \mathbf{V}_h$ and $P_h : W \to W_h$ such that

$$\begin{aligned}
&\big(\nabla\cdot(\mathbf{v}-\mathbf{\Pi}_h\mathbf{v}), w\big) = 0 \qquad \forall w\in W_h\ ,\\
&(\nabla\cdot\mathbf{y}, z-P_hz) = 0 \qquad \forall \mathbf{y}\in\mathbf{V}_h\ .
\end{aligned} \tag{9.30}$$

That is, on $(H^1(\Omega))^d\cap\mathbf{V}_h$ and with $\mathrm{div} = \nabla\cdot$,

$$\mathrm{div}\mathbf{\Pi}_h = P_h\mathrm{div}\ . \tag{9.31}$$

Relation (9.31) means that $\mathbf{\Pi}_h$ and P_h satisfies a commuting diagram (cf. Sect. 3.8.4). Moreover, these two operators have the approximation properties given in (9.29); i.e.,

$$\begin{aligned}
&\|\mathbf{v}-\mathbf{\Pi}_h\mathbf{v}\|_{\mathbf{L}^2(\Omega)} \le Ch^l\|\mathbf{v}\|_{\mathbf{H}^l(\Omega)}, && 1\le l\le r+1\ ,\\
&\|\nabla\cdot(\mathbf{v}-\mathbf{\Pi}_h\mathbf{v})\|_{L^2(\Omega)} \le Ch^l\|\nabla\cdot\mathbf{v}\|_{H^l(\Omega)}, && 0\le l\le r^*\ ,\\
&\|w-P_hw\|_{L^2(\Omega)} \le Ch^l\|w\|_{H^l(\Omega)}, && 0\le l\le r^*\ .
\end{aligned} \tag{9.32}$$

For the analysis of the pressure equation, we apply $\mathbf{\Pi}_h$ to the velocity $\mathbf{u}$, which cannot be done unless $\mathbf{u}$ is sufficiently smooth. Thus we explicitly assume that

$$\nabla p \in (L^\infty(\Omega \times J))^d \quad \text{and} \quad \mathbf{u} \in \left(L^2(J; H^1(\Omega))\right)^d . \tag{9.33}$$

The assumption that $\nabla p \in (L^\infty(\Omega \times J))^d$ was shown by Chen (2001B) under reasonable conditions on the data, and the assumption that $\mathbf{u} \in \left(L^2(J; H^1(\Omega))\right)^d$ was proven by Chen-Ewing (2001).

Set $\tilde{\boldsymbol{\kappa}} = (\boldsymbol{\kappa}\lambda)^{-1}$, and assume that it is a bounded, symmetric, and uniformly positive definite matrix; i.e.,

$$\begin{aligned} 0 < \tilde{\kappa}_* \le |\mathbf{y}|^{-2} \sum_{i,j=1}^{d} \tilde{\kappa}_{ij}(s, \mathbf{x}, t) y_i y_j \le \tilde{\kappa}^* < \infty , \\ \mathbf{x} \in \Omega,\ t \in J,\ \mathbf{y} \ne \mathbf{0} \in \mathbb{R}^d,\ s \in \mathbb{R} . \end{aligned} \tag{9.34}$$

We restate Lemma 3.7 below.

Lemma 9.1. *Given $w \in W_h$, there exists $\mathbf{v} \in \mathbf{V}_h$ such that $\nabla \cdot \mathbf{v} = w$ and*

$$\|\mathbf{v}\|_{\mathbf{V}} \le C \|w\|_{L^2(\Omega)} ,$$

where

$$\|\mathbf{v}\|_{\mathbf{V}} = \|\mathbf{v}\|_{\mathbf{H}(\mathrm{div},\Omega)} = \left\{ \|\mathbf{v}\|^2_{\mathbf{L}^2(\Omega)} + \|\nabla \cdot \mathbf{v}\|^2_{L^2(\Omega)} \right\}^{1/2} .$$

Below ϵ is a positive constant, as small as we please. For simplicity of proof, we assume that q and q_1 do not explicitly depend on p.

Theorem 9.2. *For the solution $\mathbf{u}_h^n \in \mathbf{V}_h$ and $p_h^n \in W_h$ of* (9.21)*, under assumptions* (9.33) *and* (9.34)*, we have*

$$\begin{aligned} \|\nabla \cdot (\mathbf{u} - \mathbf{u}_h^n)\|_{L^2(\Omega)} &\le C \Big\{ \|q(s) - q(s_h^n)\|_{L^2(\Omega)} \\ &\qquad + \|\nabla \cdot (\mathbf{u} - \boldsymbol{\Pi}_h \mathbf{u})\|_{L^2(\Omega)} \Big\}, \\ \|\mathbf{u} - \mathbf{u}_h^n\|_{\mathbf{L}^2(\Omega)} &+ \|p - p_h^n\|_{L^2(\Omega)} \\ &\le C \Big\{ \|\tilde{\boldsymbol{\kappa}}(s) - \tilde{\boldsymbol{\kappa}}(s_h^n)\|_{\mathbf{L}^2(\Omega)} + \|q(s) - q(s_h^n)\|_{L^2(\Omega)} \\ &\qquad + \|\boldsymbol{\gamma}_{\mathbf{1}}(s) - \boldsymbol{\gamma}_{\mathbf{1}}(s_h^n)\|_{\mathbf{L}^2(\Omega)} + \|\mathbf{u} - \boldsymbol{\Pi}_h \mathbf{u}\|_{\mathbf{L}^2(\Omega)} \\ &\qquad + \|p - P_h p\|_{L^2(\Omega)} \Big\} , \end{aligned}$$

for $t \in J^n$, $n = 1, 2, \ldots, N$.

Proof. It follows from (9.17) and (9.18) that $\mathbf{u} \in \mathbf{V}$ and $p \in W$ satisfy

$$\begin{aligned} (\nabla \cdot \mathbf{u}, w) &= \big(q(s), w\big) , && w \in W , \\ (\tilde{\boldsymbol{\kappa}}(s)\mathbf{u}, \mathbf{v}) - (p, \nabla \cdot \mathbf{v}) &= -\big(\boldsymbol{\gamma}_{\mathbf{1}}(s), \mathbf{v}\big) , && \mathbf{v} \in \mathbf{V} . \end{aligned} \tag{9.35}$$

Subtracting (9.21) from (9.35) for $t \in J^n$ gives the error equations

$$
\begin{aligned}
&(\nabla\cdot[\mathbf{u}-\mathbf{u}_h^n],w)=(q(s)-q(s_h^n),w)\quad\forall w\in W_h\ ,\\
&(\tilde{\boldsymbol{\kappa}}(s_h^n)[\mathbf{u}-\mathbf{u}_h^n],\mathbf{v})-(p-p_h^n,\nabla\cdot\mathbf{v})=(\boldsymbol{\gamma}_{\mathbf{1}}(s_h^n)-\boldsymbol{\gamma}_{\mathbf{1}}(s),\mathbf{v})\\
&\qquad\qquad+([\tilde{\boldsymbol{\kappa}}(s_h^n)-\tilde{\boldsymbol{\kappa}}(s)]\mathbf{u},\mathbf{v})\quad\forall\mathbf{v}\in\mathbf{V}_h\ .
\end{aligned}
\tag{9.36}
$$

First, take $w=\nabla\cdot(\boldsymbol{\Pi}_h\mathbf{u}-\mathbf{u}_h^n)$ in (9.36) to show the first inequality in this theorem. Second, choose $w=P_h(p-p_h^n)$ and $\mathbf{v}=\boldsymbol{\Pi}_h(\mathbf{u}-\mathbf{u}_h^n)$ and add the resulting two equations to see that

$$
\begin{aligned}
&(\tilde{\boldsymbol{\kappa}}(s_h^n)[\mathbf{u}-\mathbf{u}_h^n],\boldsymbol{\Pi}_h[\mathbf{u}-\mathbf{u}_h^n])+(\nabla\cdot[\mathbf{u}-\mathbf{u}_h^n],P_h[p-p_h^n])\\
&\quad-(p-p_h^n,\nabla\cdot\boldsymbol{\Pi}_h[\mathbf{u}-\mathbf{u}_h^n])=(q(s)-q(s_h^n),P_h[p-p_h^n])\\
&\quad+(\boldsymbol{\gamma}_{\mathbf{1}}(s_h^n)-\boldsymbol{\gamma}_{\mathbf{1}}(s),\boldsymbol{\Pi}_h[\mathbf{u}-\mathbf{u}_h^n])+([\tilde{\boldsymbol{\kappa}}(s_h^n)-\tilde{\boldsymbol{\kappa}}(s)]\mathbf{u},\boldsymbol{\Pi}_h[\mathbf{u}-\mathbf{u}_h^n])\ .
\end{aligned}
$$

It follows from (9.31) that the second two terms in the left-hand side of the above equation cancel, so that, by (9.34),

$$
\begin{aligned}
\|\mathbf{u}-\mathbf{u}_h^n\|_{\mathbf{L}^2(\Omega)}\le C\Big\{&\|\tilde{\boldsymbol{\kappa}}(s_h^n)-\tilde{\boldsymbol{\kappa}}(s)\|_{\mathbf{L}^2(\Omega)}+\|q(s)-q(s_h^n)\|_{L^2(\Omega)}\\
&+\|\boldsymbol{\gamma}_{\mathbf{1}}(s)-\boldsymbol{\gamma}_{\mathbf{1}}(s_h^n)\|_{\mathbf{L}^2(\Omega)}+\|\mathbf{u}-\boldsymbol{\Pi}_h\mathbf{u}\|_{\mathbf{L}^2(\Omega)}\Big\}\\
&+\epsilon\|P_h(p-p_h^n)\|_{L^2(\Omega)}\ .
\end{aligned}
$$

Third, take $\mathbf{v}$ in (9.36) associated with $P_h(p-p_h^n)$ according to Lemma 9.1:

$$
\begin{aligned}
(P_h[p-p_h^n],p-p_h^n)&=(\nabla\cdot\mathbf{v},p-p_h^n)\\
&=(\tilde{\boldsymbol{\kappa}}(s_h^n)[\mathbf{u}-\mathbf{u}_h^n],\mathbf{v})-(\boldsymbol{\gamma}_{\mathbf{1}}(s)-\boldsymbol{\gamma}_{\mathbf{1}}(s_h^n),\mathbf{v})\\
&\quad-([\tilde{\boldsymbol{\kappa}}(s_h^n)-\tilde{\boldsymbol{\kappa}}(s)]\mathbf{u},\mathbf{v})\\
&\le C\Big\{\|\mathbf{u}-\mathbf{u}_h^n\|^2_{\mathbf{L}^2(\Omega)}+\|\boldsymbol{\gamma}_{\mathbf{1}}(s)-\boldsymbol{\gamma}_{\mathbf{1}}(s_h^n)\|^2_{\mathbf{L}^2(\Omega)}\\
&\quad+\|\tilde{\boldsymbol{\kappa}}(s_h^n)-\tilde{\boldsymbol{\kappa}}(s)\|^2_{\mathbf{L}^2(\Omega)}\Big\}+\epsilon\|P_h(p-p_h^n)\|^2_{L^2(\Omega)}\ .
\end{aligned}
$$

Finally, combine these two inequalities to have the desired result. □

The proof of Theorem 9.2 is similar to that of Theorem 3.8.

9.5.2 Analysis for the Saturation Equation

As mentioned, the capillary diffusion coefficient $\mathbf{a}(s)$ can vanish at some values of s. However, in the subsequent analysis, it is assumed to be bounded, symmetric, and uniformly positive definite:

$$
\begin{aligned}
&0<a_*\le|\mathbf{y}|^{-2}\sum_{i,j=1}^{d}a_{ij}(s,\mathbf{x},t)y_iy_j\le a^*<\infty\ ,\\
&\qquad\mathbf{x}\in\Omega,\ t\in J,\ \mathbf{y}\ne\mathbf{0}\in\mathbb{R}^d,\ s\in\mathbb{R}\ .
\end{aligned}
\tag{9.37}
$$

For an analysis without the positive-definiteness assumption, see Chen et al. (2002; 2003C).

The error analysis uses a technique by Wheeler (1973) that relies on a projection of the exact saturation s in M_h: Find $\tilde{s}_h \in M_h$ such that

$$(\mathbf{a}(s)\nabla[\tilde{s}_h - s], \nabla w) + (\tilde{s}_h - s, w) = 0 \qquad \forall w \in M_h,\ t \in J\ .$$

By (9.23), this equation becomes

$$\begin{aligned}(\mathbf{a}(s)\nabla\tilde{s}_h, \nabla w) + (\tilde{s}_h - s, w) = (q_1(s), w) - \left(\psi\frac{\partial s}{\partial \boldsymbol{\tau}}, w\right) \\ \forall w \in M_h,\ t \in J\ .\end{aligned} \tag{9.38}$$

We assume that the solution $\tilde{s}_h$ satisfies

$$\begin{aligned}&\|\tilde{s}_h\|_{L^\infty(J;W^{1,\infty}(\Omega))} \le C\ ,\\ &\|s - \tilde{s}_h\|_{L^\infty(J;L^2(\Omega))} + h\|s - \tilde{s}_h\|_{L^\infty(J;H^1(\Omega))}\\ &\qquad\qquad \le Ch^{r_1+1}\|s\|_{L^\infty(J;H^{r+1}(\Omega))}\ ,\\ &\left\|\frac{\partial}{\partial t}(s - \tilde{s}_h)\right\|_{L^2(\Omega\times J)} \le Ch^{r_1+1}\|s\|_{H^1(J;H^{r+1}(\Omega))}\ ,\end{aligned} \tag{9.39}$$

where $r_1 \ge 1$ and C is independent of h. These estimates can be obtained under appropriate conditions on the coefficients $\mathbf{a}$ and q_1 and on the solution s (Wheeler, 1973). Also, we assume the explicit hypotheses on the coefficients

$$\begin{aligned}&0 < \phi_* \le \phi(\mathbf{x}) \le \phi^* < \infty,\ \mathbf{x} \in \Omega,\ \frac{d\phi}{dx_i} \in L^\infty(\Omega)\ ,\\ &\frac{dq}{ds},\ \frac{d\tilde{\kappa}_{ij}}{ds},\ \frac{d\gamma_{1,i}}{ds},\ \frac{dq_1}{ds},\ \frac{da_{ij}}{ds},\ \frac{d^2 f_w}{ds^2} \in L^\infty(\Omega\times J)\ ,\end{aligned} \tag{9.40}$$

for $i, j = 1, 2, \ldots, d$.

Define

$$\check{\mathbf{x}} = \check{\mathbf{x}}^{n,m} = \mathbf{x} - \frac{df_w}{ds}\left(s_h^{n,m-1}\right)\frac{E\mathbf{u}_h^{n,m}}{\phi(\mathbf{x})}\Delta t_s^{n,m}, \qquad \mathbf{x} \in \Omega\ .$$

The convergence analysis also uses an analogue of $\check{\mathbf{x}}^{n,m}$ defined in terms of the exact solutions s and $\mathbf{u}$. If v is a function on Ω, we define

$$\hat{\mathbf{x}} = \hat{\mathbf{x}}^{n,m} = \mathbf{x} - \frac{df_w}{ds}\left(s^{n,m}\right)\frac{E\mathbf{u}^{n,m}}{\phi(\mathbf{x})}\Delta t_s^{n,m}, \quad \hat{v}^{n,m} = v\left(\hat{\mathbf{x}}\right)\ .$$

Below we carry out the proof for two space dimensions in detail; the three-dimensional case will be mentioned later. For simplicity, we choose the initial approximation $s_h^0 = \tilde{s}_h^0$. The proof of Theorem 9.3 below follows

Ewing et al. (1984), where a differential system for the single-phase, miscible displacement of one incompressible fluid by another in a porous medium was considered, while a two-phase incompressible, immiscible flow is being treated. For the analysis of the MMOC for a linear differential problem, the reader may refer to Sect. 5.8. In the following proof, special care needs to be taken of on the nonlinearity of (9.19) and the coupling between (9.17)–(9.19).

Theorem 9.3. *Assume that K_h is a quasi-uniform triangulation of Ω (cf. (1.78)). For the solution $s_h^{n,m} \in M_h$ of (9.24), under assumptions (9.33), (9.34), (9.37), (9.39), (9.40), and $\Delta t_s = o(h)$, we have*

$$
\begin{aligned}
&\|s - s_h\|_{L^\infty(J;L^2(\Omega))} \\
&\le C\Bigg\{ h^{r_1+1}\|s\|_{L^\infty(J;H^{r_1+1}(\Omega))} + h^{r_1+1}\,\|s\|_{H^1(J;H^{r_1+1}(\Omega))} \\
&\quad + h^{r+1}\|\mathbf{u}\|_{\mathbf{L}^\infty(J;\mathbf{H}^{r+1}(\Omega))} + h^{r^*}\|p\|_{L^\infty(J;H^{r^*}(\Omega))} \\
&\quad + \Delta t_s \left\| \frac{\partial s}{\partial t} \right\|_{L^2(\Omega\times J)} + \Delta t_s \left\| \frac{\partial^2 s}{\partial \boldsymbol{\tau}^2} \right\|_{L^2(\Omega\times J)} \\
&\quad + (\Delta t_p)^2 \left\| \frac{\partial \mathbf{u}}{\partial t} \right\|_{\mathbf{L}^2(\Omega\times J)} + \left(\Delta t_p^1\right)^{3/2} \left\| \frac{\partial \mathbf{u}}{\partial t} \right\|_{\mathbf{L}^\infty(J;\mathbf{L}^2(\Omega))} \Bigg\} .
\end{aligned}
\tag{9.41}
$$

Proof. Set $\xi = s - \tilde{s}_h$ and $\eta = s_h - \tilde{s}_h$. By (9.39), it suffices to estimate η. Subtracting (9.38) from (9.24) and performing simple manipulations give the error equation

$$
\begin{aligned}
&\left(\phi\frac{\eta^{n,m}-\eta^{n,m-1}}{\Delta t_s^{n,m}}, w\right) + \left(\mathbf{a}\left(s_h^{n,m-1}\right)\nabla\eta^{n,m}, \nabla w\right) \\
&= \left(q_1\left(s_h^{n,m-1}\right) - q_1\left(s^{n,m}\right), w\right) - \left(\xi^{n,m}, w\right) \\
&- \left(\left[\mathbf{a}\left(s_h^{n,m-1}\right) - \mathbf{a}\left(s^{n,m}\right)\right]\nabla\tilde{s}_h^{n,m}, \nabla w\right) \\
&+ \left(\left[\phi\frac{\partial s^{n,m}}{\partial t} + \frac{df_w(s^{n,m})}{ds}E\mathbf{u}^{n,m}\cdot\nabla s^{n,m}\right] - \phi\frac{s^{n,m}-\hat{s}^{n,m-1}}{\Delta t_s^{n,m}}, w\right) \\
&+ \left(\frac{df_w(s^{n,m})}{ds}\left[\mathbf{u}^{n,m} - E\mathbf{u}^{n,m}\right]\cdot\nabla s^{n,m}, w\right) + \left(\phi\frac{\xi^{n,m}-\xi^{n,m-1}}{\Delta t_s^{n,m}}, w\right) \\
&+ \left(\phi\frac{\check{s}^{n,m-1}-\hat{s}^{n,m-1}}{\Delta t_s^{n,m}}, w\right) - \left(\phi\frac{\check{\xi}^{n,m-1}-\hat{\xi}^{n,m-1}}{\Delta t_s^{n,m}}, w\right) \\
&+ \left(\phi\frac{\check{\eta}^{n,m-1}-\hat{\eta}^{n,m-1}}{\Delta t_s^{n,m}}, w\right) + \left(\phi\frac{\xi^{n,m-1}-\hat{\xi}^{n,m-1}}{\Delta t_s^{n,m}}, w\right) \\
&- \left(\phi\frac{\eta^{n,m-1}-\hat{\eta}^{n,m-1}}{\Delta t_s^{n,m}}, w\right), \qquad w \in M_h\ .
\end{aligned}
$$

For notational convenience, set $\Delta t_s = \Delta t_s^{n,m}$. Take $w = \eta^{n,m}$ in this equation and write the resulting terms as

$$\left(\phi\frac{\eta^{n,m}-\eta^{n,m-1}}{\Delta t_s},\eta^{n,m}\right)+\left(\mathbf{a}\left(s_h^{n,m-1}\right)\nabla\eta^{n,m},\nabla\eta^{n,m}\right)=\sum_{i=1}^{11}T_i\ , \quad (9.42)$$

with the obvious definition of T_i, $i = 1, 2, \ldots, 11$.

Using the inequality

$$b(b-c)\geq\frac{1}{2}(b^2-c^2),\quad b,\ c\in\mathbb{R}\ ,$$

we see that

$$\left(\phi\frac{\eta^{n,m}-\eta^{n,m-1}}{\Delta t_s},\eta^{n,m}\right)\geq\frac{1}{2\Delta t_s}\left\{(\phi\eta^{n,m},\eta^{n,m})-\left(\phi\eta^{n,m-1},\eta^{n,m-1}\right)\right\}\ . \quad (9.43)$$

Clearly, by (9.39) and (9.40), we have

$$\begin{aligned}
|T_1| &\leq C\Bigg(\|\eta^{n,m-1}\|^2_{L^2(\Omega)}+\|\xi^{n,m-1}\|^2_{L^2(\Omega)}\\
&\qquad +\|\eta^{n,m}\|^2_{L^2(\Omega)}+\|s^{n,m-1}-s^{n,m}\|^2_{L^2(\Omega)}\Bigg)\ ,\\
|T_2| &\leq C\left(\|\xi^{n,m}\|^2_{L^2(\Omega)}+\|\eta^{n,m}\|^2_{L^2(\Omega)}\right)\ , \qquad (9.44)\\
|T_3| &\leq C\Bigg(\|\eta^{n,m-1}\|^2_{L^2(\Omega)}+\|\xi^{n,m-1}\|^2_{L^2(\Omega)}\\
&\qquad +\|s^{n,m-1}-s^{n,m}\|^2_{L^2(\Omega)}\Bigg)+\epsilon\|\nabla\eta^{n,m}\|^2_{\mathbf{L}^2(\Omega)}\ .
\end{aligned}$$

The term T_4 can be bounded as in Theorem 5.4 (cf. (5.99)):

$$|T_4|\leq C\left(\Delta t_s\left\|\frac{\partial^2 s}{\partial\boldsymbol{\tau}^2}\right\|^2_{L^2(\Omega\times(t^{n,m-1},t^{n,m}))}+\|\eta^{n,m}\|^2_{L^2(\Omega)}\right)\ . \quad (9.45)$$

By the definition of $E\mathbf{u}^{n,m}$, we have

$$|T_5|\leq C\left(\|\eta^{n,m}\|^2_{L^2(\Omega)}+(\Delta t_p)^3\left\|\frac{\partial\mathbf{u}}{\partial t}\right\|^2_{\mathbf{L}^2(\Omega\times(t^{n-1},t^{n+1}))}\right)\ . \quad (9.46)$$

If $t^{n,m}\leq t^1$, $E\mathbf{u}^{n,m}=\mathbf{u}^0$, so the temporal error term in (9.46) is replaced by

$$C\left(\Delta t_p^1\right)^2\left\|\frac{\partial\mathbf{u}}{\partial t}\right\|^2_{\mathbf{L}^\infty((t^0,t^1);\mathbf{L}^2(\Omega))}\ .$$

Next, it is obvious that

$$|T_6| \le C \left(\|\eta^{n,m}\|^2_{L^2(\Omega)} + (\Delta t_s)^{-1} \left\| \frac{\partial \xi}{\partial t} \right\|^2_{L^2(\Omega \times (t^{n,m-1}, t^{n,m}))} \right) . \tag{9.47}$$

The estimates of T_7, T_8, and T_9 fit into the following framework. Let v be a function defined on Ω; in T_7, T_8, and T_9, v is s, ξ, and η, respectively. Let $\mathbf{z}$ be the unit vector in the direction of

$$\frac{df_w}{ds}\left(s_h^{n,m-1}\right) E\mathbf{u}_h^{n,m} - \frac{df_w}{ds}\left(s^{n,m}\right) E\mathbf{u}^{n,m} .$$

Then we have

$$\begin{aligned} &\int_\Omega \phi \frac{\check{v}^{n,m-1} - \hat{v}^{n,m-1}}{\Delta t_s} \eta^{n,m} \, d\mathbf{x} \\ &\quad = \frac{1}{\Delta t_s} \int_\Omega \phi \left[\int_{\hat{\mathbf{x}}}^{\check{\mathbf{x}}} \frac{\partial v^{n,m-1}}{\partial \mathbf{z}} \, d\mathbf{z} \right] \eta^{n,m} \, d\mathbf{x} \\ &\quad = \frac{1}{\Delta t_s} \int_\Omega \phi \left[\int_0^1 \frac{\partial v^{n,m-1}}{\partial \mathbf{z}} \big((1-\ell)\hat{\mathbf{x}} + \ell \check{\mathbf{x}} \big) \, d\ell \right] \\ &\qquad\qquad \cdot \|\check{\mathbf{x}} - \hat{\mathbf{x}}\| \eta^{n,m} \, d\mathbf{x} , \end{aligned} \tag{9.48}$$

where the parameter $\ell \in [0,1]$ describes the segment from $\hat{\mathbf{x}}$ to $\check{\mathbf{x}}$. Set

$$g_v(\mathbf{x}) = \int_0^1 \frac{\partial v^{n,m-1}}{\partial \mathbf{z}} \big((1-\ell)\hat{\mathbf{x}} + \ell \check{\mathbf{x}} \big) \, d\ell .$$

Because

$$\check{\mathbf{x}} - \hat{\mathbf{x}} = \left(\frac{df_w}{ds}\left(s^{n,m}\right) E\mathbf{u}^{n,m} - \frac{df_w}{ds}\left(s_h^{n,m-1}\right) E\mathbf{u}_h^{n,m} \right) \frac{\Delta t_s}{\phi} ,$$

the terms T_7, T_8, and T_9, with an application of (9.48) with $v = s$, ξ, and η, respectively, can be bounded as follows:

$$\begin{aligned} |T_7| &\le \left\| \frac{df_w}{ds}\left(s^{n,m}\right) E\mathbf{u}^{n,m} - \frac{df_w}{ds}\left(s_h^{n,m-1}\right) E\mathbf{u}_h^{n,m} \right\|_{\mathbf{L}^2(\Omega)} \\ &\qquad \cdot \|g_s\|_{L^\infty(\Omega)} \|\eta^{n,m}\|_{L^2(\Omega)} , \\ |T_8| &\le \left\| \frac{df_w}{ds}\left(s^{n,m}\right) E\mathbf{u}^{n,m} - \frac{df_w}{ds}\left(s_h^{n,m-1}\right) E\mathbf{u}_h^{n,m} \right\|_{\mathbf{L}^2(\Omega)} \\ &\qquad \cdot \|g_\xi\|_{L^2(\Omega)} \|\eta^{n,m}\|_{L^\infty(\Omega)} , \\ |T_9| &\le \left\| \frac{df_w}{ds}\left(s^{n,m}\right) E\mathbf{u}^{n,m} - \frac{df_w}{ds}\left(s_h^{n,m-1}\right) E\mathbf{u}_h^{n,m} \right\|_{\mathbf{L}^2(\Omega)} \\ &\qquad \cdot \|g_\eta\|_{L^2(\Omega)} \|\eta^{n,m}\|_{L^\infty(\Omega)} . \end{aligned} \tag{9.49}$$

By (9.33) and (9.40), we see that

$$\begin{aligned}
&\left\| \frac{df_w}{ds}\left(s^{n,m}\right) E\mathbf{u}^{n,m} - \frac{df_w}{ds}\left(s_h^{n,m-1}\right) E\mathbf{u}_h^{n,m} \right\|_{\mathbf{L}^2(\Omega)} \\
&\quad \le C\Bigg(\|\eta^{n,m-1}\|^2_{L^2(\Omega)} + \|\xi^{n,m-1}\|^2_{L^2(\Omega)} \\
&\qquad + \|s^{n,m-1} - s^{n,m}\|^2_{L^2(\Omega)} + \|\mathbf{u}^n - \mathbf{u}_h^n\|^2_{\mathbf{L}^2(\Omega)} \\
&\qquad + \|\mathbf{u}^{n-1} - \mathbf{u}_h^{n-1}\|^2_{\mathbf{L}^2(\Omega)} \Bigg) .
\end{aligned} \tag{9.50}$$

Since g_s is an average of certain first partial derivatives of $s^{n,m-1}$, which are bounded by $\|s^{n,m-1}\|_{W^{1,\infty}(\Omega)}$, it follows from (9.49) and (9.50) that

$$\begin{aligned}
|T_7| \le C\Bigg(&\|\eta^{n,m-1}\|^2_{L^2(\Omega)} + \|\xi^{n,m-1}\|^2_{L^2(\Omega)} \\
&+ \|s^{n,m-1} - s^{n,m}\|^2_{L^2(\Omega)} + \|\mathbf{u}^n - \mathbf{u}_h^n\|^2_{\mathbf{L}^2(\Omega)} \\
&+ \|\mathbf{u}^{n-1} - \mathbf{u}_h^{n-1}\|^2_{\mathbf{L}^2(\Omega)} + \|\eta^{n,m}\|^2_{L^2(\Omega)} \Bigg) .
\end{aligned} \tag{9.51}$$

To estimate $\|g_\xi\|_{L^2(\Omega)}$ and $\|g_\eta\|_{L^2(\Omega)}$, we make the induction hypothesis

$$\|\mathbf{u}_h^n\|_{\mathbf{L}^\infty(\Omega)},\ \|\mathbf{u}_h^{n-1}\|_{\mathbf{L}^\infty(\Omega)},\ \|s_h^{n,m-1}\|_{L^\infty(\Omega)} \le \left(\frac{h}{\Delta t_s}\right)^{1/2} , \tag{9.52}$$

which will be shown at the end of the proof. Observe that

$$\|g_v\|^2_{L^2(\Omega)} \le \int_0^1 \int_\Omega \left| \frac{\partial v^{n,m-1}}{\partial \mathbf{z}}\big((1-\ell)\hat{\mathbf{x}} + \ell\check{\mathbf{x}}\big) \right|^2 d\mathbf{x}\, d\ell . \tag{9.53}$$

Defining the transformation

$$\begin{aligned}
\mathbf{G}_\ell(\mathbf{x}) &= (1-\ell)\hat{\mathbf{x}} + \ell\check{\mathbf{x}} \\
&= \mathbf{x} - \Bigg(\frac{df_w}{ds}\left(s^{n,m}\right) E\mathbf{u}^{n,m} \\
&\quad + \ell\left[\frac{df_w}{ds}\left(s_h^{n,m-1}\right) E\mathbf{u}_h^{n,m} - \frac{df_w}{ds}\left(s^{n,m}\right) E\mathbf{u}^{n,m} \right] \Bigg) \frac{\Delta t_s}{\phi} ,
\end{aligned} \tag{9.54}$$

inequality (9.53) becomes

$$\|g_v\|^2_{L^2(\Omega)} \le \int_0^1 \sum_{K\in K_h} \int_K \left| \frac{\partial v^{n,m-1}}{\partial \mathbf{z}}\big(\mathbf{G}_\ell(\mathbf{x})\big) \right|^2 d\mathbf{x}\, d\ell . \tag{9.55}$$

It follows from (9.54) that the Jacobian of $\mathbf{G}_\ell$ is the identity matrix, plus Δt_s times terms involving first partial derivatives of ϕ, $E\mathbf{u}$, and s (that are bounded) and of $E\mathbf{u}_h$ and s_h. Note that, by (9.52) and an inverse inequality (cf. (1.139)),

$$\begin{aligned} &\|\nabla \left(E\mathbf{u}_h^{n,m}\right)\|_{\mathbf{L}^2(\Omega)}\Delta t_s \\ &\quad \le Ch^{-1}\max\left\{\|\mathbf{u}_h^n\|_{\mathbf{L}^\infty(\Omega)}, \|\mathbf{u}_h^{n-1}\|_{\mathbf{L}^\infty(\Omega)}\right\}\Delta t_s \\ &\quad \le C\left(\frac{\Delta t_s}{h}\right)^{1/2} = o(1)\ , \end{aligned} \tag{9.56}$$

since $\Delta t_s = o(h)$. Similarly, we can show that

$$\|\nabla s_h^{n,m-1}\|_{\mathbf{L}^2(\Omega)} \le C\left(\frac{\Delta t_s}{h}\right)^{1/2} = o(1)\ . \tag{9.57}$$

Consequently, by (9.40), the determinant of the Jacobian of $\mathbf{G}_\ell$ equals

$$|\mathbf{J}(\mathbf{G}_\ell)| = 1 + o(1)\ . \tag{9.58}$$

Thus a change of variable in (9.55) yields

$$\|g_v\|_{L^2(\Omega)}^2 \le 2\int_0^1 \sum_{K\in K_h}\int_{\mathbf{G}_\ell(K)}\left|\frac{\partial v^{n,m-1}}{\partial \mathbf{z}}(\mathbf{x})\right|^2 d\mathbf{x}\ d\ell\ . \tag{9.59}$$

By (9.40), (9.54), (9.56), and (9.57), we see that

$$\|\mathbf{G}_\ell(\mathbf{x}_1) - \mathbf{G}_\ell(\mathbf{x}_2)\| \ge \|\mathbf{x}_1 - \mathbf{x}_2\|\big(1 - o(1)\big)\ , \tag{9.60}$$

so $\mathbf{G}_\ell$ is a one-to-one mapping on each element $K \in K_h$. Also, because

$$\begin{aligned} \|\mathbf{G}_\ell(\mathbf{x}) - \mathbf{x}\| &= \left(\mathcal{O}(1) + \mathcal{O}\left(\left[\frac{h}{\Delta t_s}\right]^{1/2}\right)\right)\Delta t_s \\ &= \mathcal{O}\left(\Delta t_s\right) + \mathcal{O}\left(h^{1/2}\Delta t_s^{1/2}\right) = o(h)\ , \end{aligned} \tag{9.61}$$

the mapping $\mathbf{G}_\ell$ maps K into itself and its immediate-neighbor elements. Therefore, $\mathbf{G}_\ell$ is globally at most finitely-many-to-one (with a repetition factor bounded by the number of neighbors of an element) and maps Ω into itself and its immediate-neighbor periodic copies. As a result, the sum in (9.59) is bounded by finitely many multiples of an Ω-integral, so that

$$\|g_v\|_{L^2(\Omega)} \le C\|\nabla v^{n,m-1}\|_{\mathbf{L}^2(\Omega)}\ . \tag{9.62}$$

Using the inequality in two dimensions (Bramble, 1966)

$$\|\eta^{n,m}\|_{L^\infty(\Omega)} \le C\left|\ln h\right|^{1/2}\|\eta^{n,m}\|_{H^1(\Omega)}\ , \tag{9.63}$$

and inequality (9.62) (with $v = \xi$), the second inequality of (9.49) implies

$$|T_8| \le \left\| \frac{df_w}{ds}\left(s^{n,m}\right) E\mathbf{u}^{n,m} - \frac{df_w}{ds}\left(s_h^{n,m-1}\right) E\mathbf{u}_h^{n,m} \right\|_{\mathbf{L}^2(\Omega)} \cdot \|\nabla \xi^{n,m-1}\|_{\mathbf{L}^2(\Omega)} \left|\ln h\right|^{1/2} \|\eta^{n,m}\|_{H^1(\Omega)} ,$$

so, by (9.39) and (9.50),

$$\begin{aligned} |T_8| \le C \Big(& \|\eta^{n,m-1}\|^2_{L^2(\Omega)} + \|\xi^{n,m-1}\|^2_{L^2(\Omega)} \\ & + \|s^{n,m-1} - s^{n,m}\|^2_{L^2(\Omega)} + \|\mathbf{u}^n - \mathbf{u}_h^n\|^2_{\mathbf{L}^2(\Omega)} \\ & + \|\mathbf{u}^{n-1} - \mathbf{u}_h^{n-1}\|^2_{\mathbf{L}^2(\Omega)} \Big) + \epsilon \|\eta^{n,m}\|^2_{H^1(\Omega)} . \end{aligned} \tag{9.64}$$

Using (9.50), inequality (9.41) inductively shows that

$$\begin{aligned} & \left\| \frac{df_w}{ds}\left(s^{n,m}\right) E\mathbf{u}^{n,m} - \frac{df_w}{ds}\left(s_h^{n,m-1}\right) E\mathbf{u}_h^{n,m} \right\|_{\mathbf{L}^2(\Omega)} \\ & \quad \le C \left(h^{r_1+1} + h^{r+1} + h^{r^*} + \Delta t_s + (\Delta t_p)^2 + \left(\Delta t_p^1\right)^{3/2} \right) , \end{aligned}$$

which, together with an application of (9.62) (with $v = \eta$) and (9.63) to the third inequality of (9.49), yields

$$\begin{aligned} |T_9| \le & \left\| \frac{df_w}{ds}\left(s^{n,m}\right) E\mathbf{u}^{n,m} - \frac{df_w}{ds}\left(s_h^{n,m-1}\right) E\mathbf{u}_h^{n,m} \right\|_{\mathbf{L}^2(\Omega)} \\ & \cdot \|\nabla \eta^{n,m-1}\|_{\mathbf{L}^2(\Omega)} \left|\ln h\right|^{1/2} \|\eta^{n,m}\|_{H^1(\Omega)} \\ \le & \epsilon \left(\|\nabla \eta^{n,m-1}\|^2_{\mathbf{L}^2(\Omega)} + \|\eta^{n,m}\|^2_{H^1(\Omega)} \right) . \end{aligned} \tag{9.65}$$

We now estimate T_{10} and T_{11}. These two terms are of the form

$$\int_\Omega \phi \frac{v^{n,m-1} - \hat{v}^{n,m-1}}{\Delta t_s} \eta^{n,m} \, d\mathbf{x} ,$$

with $v = \xi$ or η, which can be bounded by

$$\begin{aligned} & \left| \int_\Omega \phi \frac{v^{n,m-1} - \hat{v}^{n,m-1}}{\Delta t_s} \eta^{n,m} \, d\mathbf{x} \right| \\ & \quad \le C \left\| \frac{v^{n,m-1} - \hat{v}^{n,m-1}}{\Delta t_s} \right\|^2_{H^{-1}(\Omega)} + \epsilon \|\eta^{n,m}\|^2_{H^1(\Omega)} . \end{aligned} \tag{9.66}$$

To estimate

$$\begin{aligned}
&\left\|v^{n,m-1}-\hat{v}^{n,m-1}\right\|_{H^{-1}(\Omega)}\\
&\quad=\sup_{w\in H^1(\Omega)}\left\{\frac{1}{\|w\|_{H^1(\Omega)}}\int_\Omega\left[v^{n,m-1}(\mathbf{x})-v^{n,m-1}(\hat{\mathbf{x}})\right]w(\mathbf{x})\;d\mathbf{x}\right\}\;,
\end{aligned}$$

set

$$\mathbf{G}(\mathbf{x})=\mathbf{x}-\frac{df_w}{ds}\left(s^{n,m}\right)\frac{E\mathbf{u}^{n,m}}{\phi(\mathbf{x})}\Delta t_s\;.$$

By the periodicity assumption on s, $\mathbf{u}$, and ϕ, $\mathbf{G}$ is a differentiable mapping of Ω onto itself. Then changing variables leads to

$$\begin{aligned}
&\left\|v^{n,m-1}-\hat{v}^{n,m-1}\right\|_{H^{-1}(\Omega)}\\
&=\sup_{w\in H^1(\Omega)}\left\{\frac{1}{\|w\|_{H^1(\Omega)}}\left[\int_\Omega v^{n,m-1}(\mathbf{x})w(\mathbf{x})\;d\mathbf{x}\right.\right.\\
&\qquad\left.\left.-\int_\Omega v^{n,m-1}(\mathbf{x})w\big(\mathbf{G}^{-1}(\mathbf{x})\big)\left|\mathbf{J}(\mathbf{G})\right|^{-1}\;d\mathbf{x}\right]\right\}\\
&\le\sup_{w\in H^1(\Omega)}\left\{\frac{1}{\|w\|_{H^1(\Omega)}}\int_\Omega v^{n,m-1}(\mathbf{x})w(\mathbf{x})\left(1-\left|\mathbf{J}(\mathbf{G})\right|^{-1}\right)\;d\mathbf{x}\right\}\\
&\quad+\sup_{w\in H^1(\Omega)}\left\{\frac{1}{\|w\|_{H^1(\Omega)}}\int_\Omega v^{n,m-1}(\mathbf{x})\left(w(\mathbf{x})-w\left(\mathbf{G}^{-1}(\mathbf{x})\right)\right)\right.\\
&\qquad\left.\cdot\left|\mathbf{J}(\mathbf{G})\right|^{-1}\;d\mathbf{x}\right\}\\
&\equiv R_1+R_2\;.
\end{aligned}$$

As in (9.58) (with $\mathcal{O}(\Delta t_s)$ in place of $o(1)$), we see that

$$\begin{aligned}
|R_1|&\le C\sup_{w\in H^1(\Omega)}\left\{\frac{\|v^{n,m-1}\|_{L^2(\Omega)}\|w\|_{L^2(\Omega)}}{\|w\|_{H^1(\Omega)}}\right\}\Delta t_s\\
&\le C\|v^{n,m-1}\|_{L^2(\Omega)}\Delta t_s\;.
\end{aligned}\tag{9.67}$$

Also, as for (9.48), we have

$$\begin{aligned}
|R_2|\le 2\sup_{w\in H^1(\Omega)}\left\{\frac{1}{\|w\|_{H^1(\Omega)}}\int_\Omega v^{n,m-1}(\mathbf{x})\bar{g}_w(\mathbf{x})\right.\\
\left.\cdot\left\|\mathbf{x}-\mathbf{G}^{-1}(\mathbf{x})\right\|\;d\mathbf{x}\right\},
\end{aligned}\tag{9.68}$$

where

$$\bar{g}_w(\mathbf{x})=\int_0^1\frac{\partial w}{\partial\ell}\left((1-\ell)\mathbf{G}^{-1}(\mathbf{x})+\ell\mathbf{x}\right)\;d\ell\;.$$

Similarly to (9.61), the following inequality holds:

$$\left\|\mathbf{x} - \mathbf{G}^{-1}(\mathbf{x})\right\| \leq C\Delta t_s \,. \tag{9.69}$$

Also, as for (9.62), we have

$$\|\bar{g}_w\|_{L^2(\Omega)} \leq C\|w\|_{H^1(\Omega)} \,. \tag{9.70}$$

Applying (9.69) and (9.70) to (9.68), we see that

$$|R_2| \leq C\|v^{n,m-1}\|_{L^2(\Omega)}\Delta t_s \,. \tag{9.71}$$

Combining (9.66) (with $v = \xi$ for T_{10} and $v = \eta$ for T_{11}, respectively), (9.67), and (9.71), we get

$$\begin{aligned} |T_{10}| &\leq C\|\xi^{n,m-1}\|^2_{L^2(\Omega)} + \epsilon\|\eta^{n,m}\|^2_{H^1(\Omega)} \,, \\ |T_{11}| &\leq C\|\eta^{n,m-1}\|^2_{L^2(\Omega)} + \epsilon\|\eta^{n,m}\|^2_{H^1(\Omega)} \,. \end{aligned} \tag{9.72}$$

Finally, we combine (9.37), (9.42)–(9.47), (9.51), (9.64), (9.65), and (9.72) and use (9.32), (9.39), (9.40), and Theorem 9.2 to obtain

$$\begin{aligned} &\frac{1}{2\Delta t_s}\left\{(\phi\eta^{n,m},\eta^{n,m}) - \left(\phi\eta^{n,m-1},\eta^{n,m-1}\right)\right\} + \|\nabla\eta^{n,m}\|^2_{\mathbf{L}^2(\Omega)} \\ &\leq C\Bigg\{h^{2r_1+2}\|s\|^2_{L^\infty(J;H^{r_1+1}(\Omega))} + \Delta t_s\left\|\frac{\partial s}{\partial t}\right\|^2_{L^2(\Omega\times(t^{n,m-1},t^{n,m}))} \\ &\quad + \Delta t_s\left\|\frac{\partial^2 s}{\partial\boldsymbol{\tau}^2}\right\|^2_{L^2(\Omega\times(t^{n,m-1},t^{n,m}))} + (\Delta t_p)^3\left\|\frac{\partial\mathbf{u}}{\partial t}\right\|^2_{\mathbf{L}^2(\Omega\times(t^{n-2},t^n))} \\ &\quad + (\Delta t_s)^{-1}h^{2r_1+2}\|s\|^2_{H^1((t^{n,m-1},t^{n,m});H^{r_1+1}(\Omega))} + \|\eta^{n,m-1}\|^2_{L^2(\Omega)} \\ &\quad + h^{2r+2}\|\mathbf{u}\|^2_{\mathbf{L}^\infty(J;\mathbf{H}^{r+1}(\Omega))} + h^{2r^*}\|p\|^2_{L^\infty(J;H^{r^*}(\Omega))} \\ &\quad + \|\eta^{n,m}\|^2_{L^2(\Omega)} + \|\eta^{n-1}\|^2_{L^2(\Omega)} + \|\eta^n\|^2_{L^2(\Omega)}\Bigg\} \\ &\quad + \epsilon\|\nabla\eta^{n,m-1}\|^2_{\mathbf{L}^2(\Omega)} \,. \end{aligned}$$

If $t^{n,m} \leq t^1$, the remark after (9.46) applies. Multiply this inequality by Δt_s, sum on n and m, and use the discrete Gronwall lemma (cf. Lemma 5.5) and the fact that $\eta^0 = 0$ to obtain the desired result (9.41).

It remains to verify the induction hypothesis (9.52) for $t = t^{n+1}$. Applying an inverse inequality, Theorem 9.2, (9.41), and the fact that $\Delta t_s = o(h)$, we see that

$$
\begin{aligned}
\|\mathbf{u}_h^{n+1}\|_{\mathbf{L}^\infty(\Omega)} &\le \|\mathbf{\Pi}_h\mathbf{u}^{n+1}\|_{\mathbf{L}^\infty(\Omega)} + \|\mathbf{u}_h^{n+1} - \mathbf{\Pi}_h\mathbf{u}^{n+1}\|_{\mathbf{L}^\infty(\Omega)} \\
&\le C\left(1 + h^{-1}\|\mathbf{u}_h^{n+1} - \mathbf{\Pi}_h\mathbf{u}^{n+1}\|_{\mathbf{L}^2(\Omega)}\right) \\
&\le C\Bigg(1 + h^{-1}\Bigg[\|\mathbf{u}_h^{n+1} - \mathbf{u}^{n+1}\|_{\mathbf{L}^2(\Omega)} \\
&\qquad + \|\mathbf{u}^{n+1} - \mathbf{\Pi}_h\mathbf{u}^{n+1}\|_{\mathbf{L}^2(\Omega)}\Bigg]\Bigg) \\
&\le C\Bigg(1 + h^{-1}\Bigg[h^{r_1+1} + h^{r+1} + h^{r^*} + \Delta t_s \\
&\qquad + \left(\Delta t_p^1\right)^{3/2} + (\Delta t_p)^2\Bigg]\Bigg) \\
&\le \left(\frac{h}{\Delta t_s}\right)^{1/2},
\end{aligned}
\tag{9.73}
$$

for h sufficiently small. A similar bound can be shown for $s^{n,m}$. This completes the proof. □

Corollary 9.4. *Under the assumptions of Theorem* 9.3, *we have*

$$
\begin{aligned}
&\max_{0\le n\le N}\left\{\|p^n - p_h^n\|_{L^2(\Omega)} + \|\mathbf{u}^n - \mathbf{u}_h^n\|_{\mathbf{L}^2(\Omega)}\right\} \\
&\le C\Bigg\{h^{r_1+1}\|s\|_{L^\infty(J;H^{r_1+1}(\Omega))} + h^{r_1+1}\,\|s\|_{H^1(J;H^{r_1+1}(\Omega))} \\
&\quad + h^{r+1}\|\mathbf{u}\|_{\mathbf{L}^\infty(J;\mathbf{H}^{r+1}(\Omega))} + h^{r^*}\|p\|_{L^\infty(J;H^{r^*}(\Omega))} \\
&\quad + \Delta t_s\left\|\frac{\partial s}{\partial t}\right\|_{L^2(\Omega\times J)} + \Delta t_s\left\|\frac{\partial^2 s}{\partial \boldsymbol{\tau}^2}\right\|_{L^2(\Omega\times J)} \\
&\quad + (\Delta t_p)^2\left\|\frac{\partial \mathbf{u}}{\partial t}\right\|_{\mathbf{L}^2(\Omega\times J)} + \left(\Delta t_p^1\right)^{3/2}\left\|\frac{\partial \mathbf{u}}{\partial t}\right\|_{\mathbf{L}^\infty(J;\mathbf{L}^2(\Omega))}\Bigg\}.
\end{aligned}
\tag{9.74}
$$

This corollary can be proven by combining Theorems 9.2 and 9.3 and inequalities in (9.32) (cf. Exercise 9.6). An analogous error estimate holds for $\|\nabla\cdot(\mathbf{u}^n - \mathbf{u}_h^n)\|_{L^2(\Omega)}$, by adding the term $h^{r^*}\|\nabla\cdot\mathbf{u}\|_{L^\infty(J;H^{r^*}(\Omega))}$ to the right-hand side of (9.74) (cf. Exercise 9.7).

It is possible to extend the results (9.41) and (9.74) to three space dimensions. In this case, we must replace $\Delta t_s = o(h)$ by $\Delta t_s = o(h^{3/2})$ in Theorem 9.3, $(h/\Delta t_s)^{1/2}$ by $h^{-1/2}(h^{3/2}/\Delta t_s)^{1/2}$ in (9.52), $|\ln h|^{1/2}$ by $h^{-1/2}\,|\ln h|$ in (9.63), and h^{-1} by $h^{-3/2}$ in (9.73). With these modifications, the proof of Theorem 9.3 remains valid.

9.6 Bibliographical Remarks

For more information on the physical and fluid properties of multiphase flows in porous media, the reader should refer to Peaceman (1977B), Aziz-Settari

(1979), and Chen et al. (2004B). In particular, the book by Chen et al. (2004B) gives a thorough treatment of multiphase flows in porous media using the finite element method. These flows include single phase, two phase, three phase, black oil, compositional, thermal, and chemical flows. For more comparisons between the various formulations for these flows, see Chen-Ewing (1997A) and Chen-Huan (2003). Detailed information on the mixed and characteristic finite element methods can be found in Chaps. 3 and 5, respectively. Finally, the proof of Theorem 9.3 follows Ewing et al. (1984). An error analysis was given in Sect. 9.5 for incompressible flow; a similar analysis can be also carried out for compressible flow (Chen-Ewing, 1997B).

9.7 Exercises

9.1. Derive (9.9) and (9.10) from (9.1)–(9.4) in detail.
9.2. Derive (9.14) from (9.2)–(9.4), (9.7), and (9.12) in detail.
9.3. Derive (9.18) from (9.2), (9.4), (9.7), and (9.16) in detail.
9.4. Use the boundary condition in (9.20) and introduce appropriate spaces to write (9.17) and (9.18) in a mixed variational formulation.
9.5. Verify (9.45).
9.6. Prove Corollary 9.4.
9.7. Apply Theorems 9.2 and 9.3 to establish an error estimate for $\max_{0\le n\le N} \|\nabla\cdot(\mathbf{u}^n-\mathbf{u}_h^n)\|_{L^2(\Omega)}$.
9.8. Define a finite element approximation procedure for the phase formulation of Sect. 9.1.1 similar to that for the global formulation developed in Sects. 9.2 and 9.3, and carry out an error analysis for this procedure analogous to that given in Sect. 9.5.
9.9. Define an approximation procedure for the weighted formulation of Sect. 9.1.2 similar to that for the global formulation developed in Sects. 9.2 and 9.3, and carry out an error analysis for this procedure analogous to that given in Sect. 9.5.

10 Semiconductor Modeling

The mathematical modeling and numerical simulation of charge transport in semiconductors is a very active research area. The most popular model is the *drift-diffusion model* (van Roosbroeck, 1950), which has been widely used in the mathematical modeling of semiconductor devices (Markowich, 1986). As a result, many numerical methods have been developed for this model to simulate efficiently the electric behavior of these devices. This model describes potential distribution, carrier concentrations, and current flow in semiconductor devices. It model materials such as silicon and germanium.

The ongoing miniaturization of semiconductor devices has shifted the focus of research from the drift-diffusion model to more advanced models because the drift-diffusion model does not account well for charge transport in ultra integrated devices. An extension of the drift-diffusion model is the (classical) *hydrodynamic model* (Blotekjaer, 1970). This new model plays an important role in simulating the behavior of charge carriers in submicron semiconductor devices since it exhibits velocity overshoot and ballistic effects missing in the drift-diffusion model. Modern computer technology has made it possible to employ this model to simulate certain highly integrated devices.

Microfabrication technology has advanced rapidly since the dawn of the semiconductor era, with each advance giving a sizable reduction in the size of individual features. Where tens of micrometers were once the common size, devices fabricated with metalorganic chemical-vapor deposition and molecular-beam epitaxy now have features as small as a few nanometers. In addition, electron-beam lithography can be now used to make working field-effect transistors with gates as short as 25 nm (Feynman, 1960). On these spatial scales, quantization effects are quite evident; carriers in a high-electron-mobility transistor travel in a two-dimensional sheet, a result of perpendicular quantization. Fabrication of a gridlike gate extends the quantization to all three dimensions (Ferry-Grondin, 1992). To take into account these quantization effects (such as tunneling effects), the *quantum hydrodynamic model* has been utilized (Wigner, 1932; Ancona-Iafrate, 1989). This model approximates quantum effects in the propagation of electrons in a semiconductor device by adding quantum corrections to the classical hydrodynamic model.

In this chapter, we introduce these three models (Sect. 10.1) and finite element methods for solving them (Sect. 10.2). In Sect. 10.3, we present a numerical example using the hydrodynamic model. Finally, bibliographical information is given in Sect. 10.4.

10.1 Three Semiconductor Models

10.1.1 The Drift-Diffusion Model

The flow of charged carriers in semiconductors is modeled by the *Boltzmann equation* (1872)

$$\frac{\partial f}{\partial t} + \mathbf{u} \cdot \nabla_{\mathbf{x}} f - \frac{e}{m} \mathbf{E} \cdot \nabla_{\mathbf{u}} f = Q(f) \,, \tag{10.1}$$

where $f = f(\mathbf{x}, \mathbf{u}, t)$ is the distribution function of a carrier species, $\mathbf{u}$ is the species' group velocity, e is the electron charge modulus, m is the effective electron mass, $\mathbf{E}$ is the electric field, Q is the time rate of change of f due to collisions, and $\nabla_{\mathbf{x}}$ and $\nabla_{\mathbf{u}}$ represent the gradient operators with respect to $\mathbf{x}$ and $\mathbf{u}$, respectively. The Boltzmann equation (10.1) is derived under the assumption that the traditional Lorentz force field does not have a component induced by an external magnetic field. Based on this equation, the dimension of an M-particle ensemble is $6M$ since $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M)$ and $\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_M)$. For a VLSI device with 10^4 conducting electrons, the dimension of the electron ensemble is 6×10^4, which is prohibitive in numerical simulations. This motivates the introduction of approximate models of the Boltzmann equation.

The models discussed in this chapter are derived from the first three *moments* of the Boltzmann equation:

$$m_0(\mathbf{u}) = 1, \quad \mathbf{m}_1(\mathbf{u}) = m\mathbf{u}, \quad m_2(\mathbf{u}) = \frac{m}{2}|\mathbf{u}|^{\mathbf{2}} \,.$$

We introduce some notation. The concentration n, average velocity $\mathbf{v}$, momentum $\mathbf{p}$, random velocity $\mathbf{c}$, pressure tensor $\mathbf{P}$, and internal energy density e_I are, respectively, defined by

$$\begin{aligned} &n = \int f \, d\mathbf{u}, \quad \mathbf{v} = \frac{1}{n} \int \mathbf{u} f \, d\mathbf{u}, \qquad \mathbf{p} = mn\mathbf{v} \,, \\ &\mathbf{c} = \mathbf{u} - \mathbf{v}, \qquad P_{ij} = m \int c_i c_j f \, d\mathbf{u}, \quad e_I = \frac{1}{2n} \int |\mathbf{c}|^2 f \, d\mathbf{u} \,, \end{aligned}$$

where the integration is performed over the whole $\mathbf{u}$ space. The heat flux $\mathbf{q}$, electron current density $\mathbf{J}$, and energy density w are, respectively, defined by

$$\mathbf{q} = \frac{m}{2} \int \mathbf{c}|\mathbf{c}|^2 f \, d\mathbf{u}, \quad \mathbf{J} = -en\mathbf{v}, \quad w = mn \left(e_I + \frac{1}{2}|\mathbf{v}|^2 \right) \,.$$

The drift-diffusion model can be now obtained by multiplying the Boltzman equation by m_0 and integrating the resulting equation over the velocity:

$$\frac{\partial n_i}{\partial t} + \nabla \cdot \mathbf{J}_i = -R_i \,, \qquad i = 1, 2, \ldots, M \,, \tag{10.2}$$

where we assume that there is an ensemble of M carriers each with the individual concentration n_i, current density $\mathbf{J}_i$, charge e_i, and recombination-generation rate R_i. The classical drift-diffusion model for two carriers, $n_1 = n$ and $n_2 = p$, reduces to

$$\begin{aligned} -\frac{\partial n}{\partial t} + \nabla \cdot \mathbf{J}_n &= R \,, \\ \frac{\partial p}{\partial t} + \nabla \cdot \mathbf{J}_p &= -R \,. \end{aligned} \tag{10.3}$$

The first equation of (10.3) is often called the *electron continuity equation*, while the second is termed the *hole continuity equation*. The constitutive relationships for the current densities are given by

$$\begin{aligned} \mathbf{J}_n &= \mu_n (U_T \nabla n - n \nabla \phi) \,, \\ \mathbf{J}_p &= -\mu_p (U_T \nabla p + p \nabla \phi) \,, \end{aligned} \tag{10.4}$$

where μ_n and μ_p are the field-dependent electron and hole mobilities, U_T is the thermal voltage, and ϕ is the electric potential. The potential ϕ is assumed to obey Poisson's law

$$\mathbf{E} = -\nabla \phi, \quad \nabla \cdot (\epsilon \mathbf{E}) = -e(n - p - \mathcal{C}) \,, \tag{10.5}$$

where ϵ is the dielectric constant and $\mathcal{C}$ is the doping profile. The thermal voltage U_T is related to the temperature T by

$$U_T = \kappa_B T / e \,,$$

where κ_B is the Boltzmann constant. The recombination-generation rate R is modeled by the mass-action law (Markowich, 1986)

$$R = \frac{np - n_0^2}{\tau_n (n + n_0) + \tau_p (p + n_0)} \,,$$

or by the Auger law

$$R_{au} = (C_n n + C_p p)(np - u_0^2) \,,$$

where u_0 is the intrinsic carrier concentration, τ_n and τ_p are the lifetimes, and C_n and C_p are the Auger recombination-generation coefficients. Finally, the mobilities take the form

$$\mu_i = \mu_i^0 \left(1 + \left(\frac{\mu_i^0 |\nabla \phi|}{v_i^0} \right)^{\ell_i} \right)^{-1/\ell_i} \,, \qquad i = n, \; p \,,$$

where μ_i^0 is the field-independent scattering mobility, v_i^0 is the saturation velocity, $\ell_n = 1$ or 2, and $\ell_p = 1$. The unknown variables are n, p, and ϕ. Note that we have an elliptic equation for ϕ in (10.5), and a parabolic equation for each of n and p in (10.3) and (10.4). With appropriate boundary and initial conditions, existence, uniqueness, and regularity of a solution for the transient (parabolic) system of (10.3)–(10.5) for (n, p, ϕ) has been shown (Jerome, 1985; Markowich, 1986). The corresponding stationary system generally admits multiple solutions. The development of efficient numerical methods for solving the drift-diffusion model is still an active area.

10.1.2 The Hydrodynamic Model

As noted earlier, the drift-diffusion model does not take into account two important phenomena: velocity overshoot and ballistic effects existing in submicron semiconductor devices. To include these effects, we introduce the hydrodynamic model:

$$\begin{aligned}
&\frac{\partial n}{\partial t} + \nabla \cdot (n\mathbf{v}) = 0 \,, \\
&\frac{\partial \mathbf{p}}{\partial t} + \mathbf{v}\nabla \cdot \mathbf{p} + \mathbf{p} \cdot \nabla \mathbf{v} + \nabla(n\kappa_B T) = -en\mathbf{E} + \left(\frac{\partial \mathbf{p}}{\partial t}\right)_c \,, \\
&\frac{\partial w}{\partial t} + \nabla \cdot (\mathbf{v}w) + \nabla \cdot (\mathbf{v}n\kappa_B T) = -en\mathbf{v} \cdot \mathbf{E} \\
&\qquad\qquad + \left(\frac{\partial w}{\partial t}\right)_c - \nabla \cdot \mathbf{q} \,,
\end{aligned} \tag{10.6}$$

where these equations are coupled with Poisson's equation defining the electric field $\mathbf{E}$:

$$\begin{aligned}
&\mathbf{E} = -\nabla\phi \,, \\
&\nabla \cdot (\epsilon\mathbf{E}) = e\,(N_D - N_A - n) \,,
\end{aligned} \tag{10.7}$$

with N_D and N_A the densities of donors and acceptors, respectively. The heat flux is expressed by

$$\mathbf{q} = -\kappa\nabla T \,,$$

where κ is the heat conduction coefficient. The "collision" terms in (10.6) are obtained in terms of the momentum and energy relaxation times, $\tau_{\mathbf{p}}$ and τ_w, following Nougier et al. (1981), by

$$\begin{aligned}
&\left(\frac{\partial \mathbf{p}}{\partial t}\right)_c = -\frac{\mathbf{p}}{\tau_{\mathbf{p}}}, && \tau_{\mathbf{p}} = m\frac{\mu_{n0}}{e}\frac{T_0}{T} \,, \\
&\left(\frac{\partial w}{\partial t}\right)_c = -\frac{w - 3n\kappa_B T_0/2}{\tau_w}, && \tau_w = \frac{\tau_{\mathbf{p}}}{2} + \frac{3}{2}\frac{\mu_{n0}}{ev_s^2}\frac{\kappa_B T T_0}{T + T_0} \,,
\end{aligned} \tag{10.8}$$

where T_0 is the ambient temperature, $\mu_{n0} = \mu_{n0}(T_0, N_D, N_A)$ is the low field electron mobility, and $v_s = v_s(T_0)$ is the saturation velocity. Finally, κ is determined by the Wiedemann-Franz law (Blatt, 1968)

$$\kappa = \kappa_0 \frac{\mu_{n0}}{e} n \kappa_B^2 T \left(\frac{T}{T_0} \right)^r . \tag{10.9}$$

A typical value chosen for r is -1.

The three equations in (10.6) can be obtained by multiplying the Boltzmann equation (10.1) by the first three moments m_0, $\mathbf{m}_1$, and m_2, in turn, and by integrating over the velocity space. Note that we have an elliptic equation for ϕ and a hyperbolic system for n and $\mathbf{p}$. With heat conduction, a parabolic equation occurs for w; without this conduction term, it is a hyperbolic equation. The mathematical and numerical theory for the hydrodynamic model is limited.

10.1.3 The Quantum Hydrodynamic Model

As mentioned early, the ongoing miniaturization and integration of semiconductor devices leads to quantum effects. The quantum hydrodynamic model approximates quantum effects in the propagation of electrons in a semiconductor device by adding quantum corrections to the classical hydrodynamic model. The leading $\mathcal{O}(\hbar)$ quantum corrections, where $\hbar$ is an expansion parameter describing the quantum effects, have been remarkably successful in simulating the effects of electron tunneling through potential barriers including single (Grubin-Kreskovsky, 1989) and multiple regions of negative differential resistance in the current-voltage curves of resonant tunneling diodes.

There are three major advantages of using the quantum hydrodynamic model over other models for simulating quantum semiconductors. First, this model is much less computationally intensive than the Wigner function (Kluksdahl et al., 1989) or density matrix methods (Frensley, 1985) and includes the same physics if the expansion parameter $\hbar/(8mTl^2)$ is small (Ancona-Iafrate, 1989), where l is a characteristic length scale of a device. Second, the equations of this model express intuitive classical fluid dynamical quantities (e.g., density, velocity, and temperature). Third, well understood classical boundary conditions can be imposed in simulating quantum devices.

The quantum hydrodynamic model has exactly the same structure as the hydrodynamic model in the previous subsection:

$$\begin{aligned}
&\frac{\partial n}{\partial t} + \nabla \cdot (n\mathbf{v}) = 0 \,, \\
&\frac{\partial \mathbf{p}}{\partial t} + \mathbf{v}\nabla \cdot \mathbf{p} + \mathbf{p} \cdot \nabla \mathbf{v} + \nabla(n\kappa_B T) = -en\mathbf{E} + \left(\frac{\partial \mathbf{p}}{\partial t} \right)_c \,, \\
&\frac{\partial w}{\partial t} + \nabla \cdot (\mathbf{v}w) + \nabla \cdot (\mathbf{v} n \kappa_B T) = -en\mathbf{v} \cdot \mathbf{E} \\
&\qquad\qquad + \left(\frac{\partial w}{\partial t} \right)_c - \nabla \cdot \mathbf{q} \,.
\end{aligned} \tag{10.10}$$

The Poisson equation (10.7) for the electric field applies here. Quantum mechanical effects appear in the energy density and the stress tensor. In the hydrodynamic model, the energy density w and stress tensor $\mathbf{P}$ are defined by

$$w = \frac{3}{2} n \kappa_B T + \frac{1}{2} m n |\mathbf{v}|^2 \ ,$$
$$P_{ij} = -n \kappa_B T \delta_{ij}, \quad i,\ j = 1,2,3 \ ,$$

while in the quantum hydrodynamic model, they are argumented with quantum correction terms, $i,\ j = 1,2,3$:

$$\begin{aligned} w &= \frac{3}{2} n \kappa_B T + \frac{1}{2} m n |\mathbf{v}|^2 - \frac{\hbar^2 n}{24 m} \Delta \log(n) + \mathcal{O}\left(\hbar^4\right) \ , \\ P_{ij} &= -n \kappa_B T \delta_{ij} + \frac{\hbar^2 n}{12 m} \frac{\partial^2}{\partial x_i \partial x_j} \log(n) + \mathcal{O}\left(\hbar^4\right) \ . \end{aligned} \tag{10.11}$$

The quantum hydrodynamic model involves two Schrödinger modes, one parabolic and one elliptic (Chen et al., 1995A). The development of mathematical and numerical theory is also a very active area for this model.

In summary, we have presented the drift-diffusion, classical hydrodynamic, and quantum hydrodynamic models. The first model is derived from the first moment of the Boltzmann equation, while the two other advanced models are obtained from the first three moments of this equation. Moreover, these advanced models take into account the velocity overshoot and ballistic effects. Furthermore, the quantum model includes the quantum effects.

10.2 Numerical Methods

10.2.1 The Drift-Diffusion Model

Recall the drift-diffusion model from Sect. 10.1.1: The electric potential and field satisfy the equation

$$\mathbf{E} = -\nabla \phi, \quad \nabla \cdot (\epsilon \mathbf{E}) = -e(n - p - \mathcal{C}) \ , \tag{10.12}$$

and the electron and hole concentrations satisfy the equations

$$\begin{aligned} -\frac{\partial n}{\partial t} + \nabla \cdot (\mu_n(\mathbf{E})(U_T \nabla n + n\mathbf{E})) &= R(n,p) \ , \\ -\frac{\partial p}{\partial t} + \nabla \cdot (\mu_p(\mathbf{E})(U_T \nabla p - p\mathbf{E})) &= R(n,p) \ . \end{aligned} \tag{10.13}$$

For simplicity, we present the finite element method for homogeneous Neumann or periodic boundary conditions for (10.12) and (10.13). Extensions to other boundary conditions are immediate (cf. Chaps. 3 and 5). The initial conditions are specified by

$$n(\mathbf{x},0) = n_0(\mathbf{x}), \quad p(\mathbf{x},0) = p_0(\mathbf{x}), \qquad \mathbf{x} \in \Omega\ .$$

From the continuity system (10.13) we see that the electron and hole concentration equations depend on the potential only through the electric field, so the mixed finite element method discussed in Chap. 3 is appropriate for the solution of the potential equation. Let Ω be a convex polygonal domain. For $0 < h < 1$, let K_h be a regular partition of Ω into elements, say, tetrahedra, rectangular parallelepipeds, or prisms, with the maximum mesh size h (cf. Sect. 3.4). Associated with the partition K_h, let $\mathbf{V}_h \times W_h \subset \mathbf{H}(\mathrm{div},\Omega)\times L^2(\Omega)$ be a mixed finite element space as defined in Sect. 3.4. In the case of homogeneous Neumann boundary conditions, they are incorporated into $\mathbf{V}_h$.

For each positive integer N, let $0 = t^0 < t^1 < \cdots < t^N = T$ be a partition of $J = (0,T]$ for the potential into subintervals $J^{\aleph} = (t^{\aleph-1}, t^{\aleph}]$, with length $\Delta t^{\aleph} = t^{\aleph} - t^{\aleph-1}$. The time partitions for the potential and concentrations can be different, as in the previous chapter. For simplicity, we take the same time partition for them. Now, the mixed method for the electric potential is given as follows: For any $0 \le \aleph \le N$, find $\mathbf{E}_h^{\aleph} \in \mathbf{V}_h$ and $\phi_h^{\aleph} \in W_h$ such that

$$\begin{aligned} &(\epsilon\nabla\cdot\mathbf{E}_h^{\aleph}, w) = -\left(e\big(n_h^{\aleph} - p_h^{\aleph} - \mathcal{C}^{\aleph}\big), w\right), \quad && w \in W_h\ , \\ &\left(\mathbf{E}_h^{\aleph}, \mathbf{v}\right) - (\phi_h^{\aleph}, \nabla\cdot\mathbf{v}) = 0, && \mathbf{v} \in \mathbf{V}_h\ . \end{aligned} \tag{10.14}$$

Generally, the doping function $\mathcal{C}$ does not depend on time.

The equations for n and p, while formally parabolic, are in fact dominated by the convection terms from physical considerations. Thus the characteristic finite element method developed in Chap. 5 is suitable for the solution of the concentration system (10.13), as it was in (9.24) for the saturation equation in porous media flow. As an example, we describe the MMOC procedure; other procedures can be applied similarly.

For simplicity, we assume that the mobilities μ_n and μ_p are constant. In the case where $\mu_n = \mu_n(\mathbf{E})$ and $\mu_p = \mu_p(\mathbf{E})$ are varying, appropriate extrapolation techniques should be used in the approximation of $\mathbf{E}$ in these two coefficients, as in (9.24) (cf. Exercise 10.1). Using (10.12), system (10.13) can be rewritten:

$$\frac{\partial n}{\partial t} - \mu_n\mathbf{E}\cdot\nabla n - \nabla\cdot(\mu_n U_T\nabla n) = -R(n,p) - \frac{en\mu_n}{\epsilon}(n - p - \mathcal{C})\ ,$$

$$\frac{\partial p}{\partial t} + \mu_p\mathbf{E}\cdot\nabla p - \nabla\cdot(\mu_p U_T\nabla p) = -R(n,p) + \frac{ep\mu_p}{\epsilon}(n - p - \mathcal{C})\ .$$

Let

$$\mathbf{b}_n(\mathbf{x},t) = -\mu_n\mathbf{E}, \quad \psi_n(\mathbf{x},t) = \left(1 + \|\mathbf{b}_n(\mathbf{x},t)\|^2\right)^{1/2}\ ,$$

and let the characteristic direction associated with the operator $\dfrac{\partial n}{\partial t} + \mathbf{b}\cdot\nabla n$ be denoted by $\boldsymbol{\tau}_n(\mathbf{x},t)$, so

$$\frac{\partial}{\partial\boldsymbol{\tau}_n} = \frac{1}{\psi_n(\mathbf{x},t)}\frac{\partial}{\partial t} + \frac{\mathbf{b}_n(\mathbf{x},t)}{\psi_n(\mathbf{x},t)}\cdot\nabla\ .$$

Then the electron concentration equation becomes

$$\psi_n \frac{\partial n}{\partial \boldsymbol{\tau}_n} - \nabla \cdot (\mu_n U_T \nabla n) = -R(n,p) - \frac{en\mu_n}{\epsilon}(n - p - \mathcal{C}) \ . \tag{10.15}$$

Similarly, the hole concentration equation is given by

$$\psi_p \frac{\partial p}{\partial \boldsymbol{\tau}_p} - \nabla \cdot (\mu_p U_T \nabla p) = -R(n,p) + \frac{ep\mu_p}{\epsilon}(n - p - \mathcal{C}) \ , \tag{10.16}$$

with

$$\mathbf{b}_p(\mathbf{x},t) = \mu_p \mathbf{E}, \quad \psi_p(\mathbf{x},t) = \left(1 + |\mathbf{b}_p(\mathbf{x},t)|^2\right)^{1/2} \ .$$

If $\aleph \geq 2$, the linear extrapolation of $\mathbf{E}_h^{\aleph-2}$ and $\mathbf{E}_h^{\aleph-1}$ is

$$E\left(\mathbf{E}_h^{\aleph}\right) = \left(1 + \frac{t^{\aleph} - t^{\aleph-1}}{t^{\aleph-1} - t^{\aleph-2}}\right)\mathbf{E}_h^{\aleph-1} - \frac{t^{\aleph} - t^{\aleph-1}}{t^{\aleph-1} - t^{\aleph-2}}\mathbf{E}_h^{\aleph-2} \ . \tag{10.17}$$

For $\aleph = 0, 1$, define

$$E\left(\mathbf{E}_h^{\aleph}\right) = \mathbf{E}_h^{\aleph} \ .$$

The approximation $E\left(\mathbf{E}_h^{\aleph}\right)$ is first-order accurate in time in the first step and second-order accurate in the later steps.

Let $M_h \subset H^1(\Omega)$ be any of the finite element spaces introduced in Chap. 1. The electron concentration can be computed via the MMOC procedure as follows: For each $1 \leq \aleph \leq N$, find $n_h^{\aleph} \in M_h$ such that

$$\begin{aligned}\left(\frac{n_h^{\aleph} - \check{n}_h^{\aleph-1}}{\Delta t^{\aleph}}, w\right) + \left(\mu_n U_T \nabla n_h^{\aleph}, \nabla w\right) = -\Bigg(R(n_h^{\aleph-1}, p_h^{\aleph-1}) \\ + \frac{en_h^{\aleph}\mu_n}{\epsilon}(n_h^{\aleph-1} - p_h^{\aleph-1} - \mathcal{C}^{\aleph}), w\Bigg), \qquad w \in M_h \ ,\end{aligned} \tag{10.18}$$

where

$$\check{n}_h^{\aleph-1} = n_h^{\aleph-1}\left(\mathbf{x} + \mu_n E\left(\mathbf{E}_h^{\aleph}\right)\Delta t^{\aleph}, t^{\aleph-1}\right) \ .$$

In the same manner, the hole concentration can be calculated as follows: For each $1 \leq \aleph \leq N$, find $p_h^{\aleph} \in M_h$ such that

$$\begin{aligned}\left(\frac{p_h^{\aleph} - \check{p}_h^{\aleph-1}}{\Delta t^{\aleph}}, w\right) + \left(\mu_p U_T \nabla p_h^{\aleph}, \nabla w\right) = -\Bigg(R(n_h^{\aleph-1}, p_h^{\aleph-1}) \\ - \frac{ep_h^{\aleph}\mu_n}{\epsilon}(n_h^{\aleph-1} - p_h^{\aleph-1} - \mathcal{C}^{\aleph}), w\Bigg), \qquad w \in M_h \ ,\end{aligned} \tag{10.19}$$

where

$$\check{p}_h^{\aleph-1} = p_h^{\aleph-1}\left(\mathbf{x} - \mu_p E\left(\mathbf{E}_h^{\aleph}\right)\Delta t^{\aleph}, t^{\aleph-1}\right) \ .$$

The initial approximations n_h^0 and p_h^0 can be defined as the respective appropriate projections of n_0 and p_0 in M_h, for example. Equations (10.14),

(10.18), and (10.19) can be solved as follows: After startup for n_h^0 and p_h^0, we obtain $(\mathbf{E}_h^0, \phi_h^0)$ from (10.14) and then n_h^1 and p_h^1 from (10.18) and (10.19); this process proceeds in a sequential fashion. An error analysis for equations (10.14), (10.18), and (10.19) can be performed in a similar fashion as for (9.21) and (9.24); see Sect. 9.5. In particular, if the finite element spaces $\mathbf{V}_h$, W_h, and M_h satisfy the approximation properties (9.29) and (9.39), we have (Douglas et al., 1986)

$$\begin{aligned}&\|n - n_h\|_{L^\infty(J;L^2(\Omega))} + \|p - p_h\|_{L^\infty(J;L^2(\Omega))} \\ &\quad + \|\phi - \phi_h\|_{L^\infty(J;L^2(\Omega))} + \|\mathbf{E} - \mathbf{E}_h\|_{\mathbf{L}^\infty(J;\mathbf{L}^2(\Omega))} \\ &\leq C\{h^{r_1+1} + h^{r+1} + h^{r^*} + \Delta t\}\ ,\end{aligned}$$

under appropriate assumptions on the solution and data, where $\Delta t = \max\{\Delta t^{\aleph} : 1 \leq \aleph \leq N\}$ and r, r^*, and r_1 are defined as in (9.29) and (9.39).

10.2.2 The Hydrodynamic Model

As mentioned in Sect. 10.1.2, the equations of the hydrodynamic model are mainly hyperbolic in nature. To devise numerical methods for solving these equations, we write them in a conservation law form. We define the vector of unknowns

$$\mathbf{U} = (n, p_{x_1}, p_{x_2}, p_{x_3}, w)^T\ ,$$

where $(\quad)^T$ denotes the transpose of the vector $(\quad)$. Also, we introduce the flux function $\mathbf{F} = (\mathbf{F}_{x_1}, \mathbf{F}_{x_2}, \mathbf{F}_{x_3})$, where

$$\begin{aligned}\mathbf{F}_{x_1}(\mathbf{U}) &= v_{x_1}\mathbf{U} + (0, n\kappa_B T, 0, 0, v_{x_1} n\kappa_B T)^T\ , \\ \mathbf{F}_{x_2}(\mathbf{U}) &= v_{x_2}\mathbf{U} + (0, 0, n\kappa_B T, 0, v_{x_2} n\kappa_B T)^T\ , \\ \mathbf{F}_{x_3}(\mathbf{U}) &= v_{x_3}\mathbf{U} + (0, 0, 0, n\kappa_B T, v_{x_3} n\kappa_B T)^T\ .\end{aligned}$$

Next, we write

$$\mathbf{R}(\mathbf{U}) = \boldsymbol{\xi}_{\mathbf{E}}(\mathbf{U}) + \boldsymbol{\xi}_c(\mathbf{U}) + \boldsymbol{\xi}_{heat}(\mathbf{U})\ ,$$

where

$$\begin{aligned}\boldsymbol{\xi}_{\mathbf{E}}(\mathbf{U}) &= (0, -e\,n\,E_{x_1}, -e\,n\,E_{x_2}, -e\,n\,E_{x_3}, -e\,n\,\mathbf{v}\cdot\mathbf{E})^T\ , \\ \boldsymbol{\xi}_c(\mathbf{U}) &= \left(0, \left(\frac{\partial p_{x_1}}{\partial t}\right)_c, \left(\frac{\partial p_{x_2}}{\partial t}\right)_c, \left(\frac{\partial p_{x_3}}{\partial t}\right)_c, \left(\frac{\partial w}{\partial t}\right)_c\right)^T\ , \\ \boldsymbol{\xi}_{heat}(\mathbf{U}) &= (0, 0, 0, 0, \nabla\cdot(\kappa\,\nabla T))^T\ .\end{aligned}$$

With this notation, the hydrodynamic model (10.6) can be rewritten in the conservation law format

$$\frac{\partial \mathbf{U}}{\partial t} + \nabla\cdot\mathbf{F}(\mathbf{U}) = \mathbf{R}(\mathbf{U})\ . \tag{10.20}$$

The boundary and initial conditions are given by

$$\begin{aligned} &\mathbf{B}\mathbf{U} = \mathbf{g}, && \mathbf{x} \in \Gamma,\ t \in J\ , \\ &\mathbf{U}(\mathbf{x}, t) = \mathbf{U}_0(\mathbf{x}), && \mathbf{x} \in \Omega\ , \end{aligned} \tag{10.21}$$

where $\mathbf{B}$ is a matrix-valued function, $\mathbf{g}$ and $\mathbf{U}_0$ are given, and Γ is the boundary of Ω. System (10.20) is coupled to the Poisson equation (10.7).

As we discussed in Chap. 4, the discontinuous Galerkin (DG) approximation method is a good choice for solving a hyperbolic system. We now give an overview of the discretization for (10.20) using this method. As an example, let K_h be a regular partition of Ω into rectangular parallelepipeds with the maximum mesh size $h > 0$, and define

$$V_h = \{v \in L^\infty(\Omega) : v|_K \text{ is linear}, K \in K_h\}\ .$$

After discretizing system (10.20) in space by the DG method, the resulting discrete equations can be written as the following ODE initial value problem:

$$\begin{aligned} &\frac{d\mathbf{U}_h}{dt} = \mathbf{L}_h(\mathbf{U}_h, \mathbf{g}) + \mathbf{R}_h(\mathbf{U}_h), && t \in J\ , \\ &\mathbf{U}_h(\mathbf{x}, 0) = \mathbf{U}_{0h}(\mathbf{x})\ , \end{aligned} \tag{10.22}$$

where $\mathbf{L}_h$ and $\mathbf{R}_h$ are some approximations of $-\nabla\cdot$ and $\mathbf{R}$, respectively, and $\mathbf{U}_{0h}$ is an approximation of $\mathbf{U}_0$ (e.g., the L^2-projection of $\mathbf{U}_0$ into $(V_h)^3$). The exact solution of the initial value problem (10.22) gives an approximation that is formally second-order accurate in space since linear functions are used on each element $K \in K_h$. Accordingly, a second-order accurate scheme in time should be used to discretize this ODE. Here we utilize a second-order accurate, two stage Runge-Kutta method. To enforce stability of the DG method, a local projection Λ_h is applied to the intermediate values of the Runge-Kutta discretization (Cockburn et al., 2000).

The resulting formally second-order accurate scheme is

- Set $\mathbf{U}_h^0 = \Lambda_h(\mathbf{U}_{0h})$;
- For $\aleph = 0, 1, \ldots, N$, given $\mathbf{U}_h^\aleph$, compute $\mathbf{U}_h^{\aleph+1}$ as follows:
  - set $\mathbf{U}_h^{[0]} = \mathbf{U}_h^\aleph$;
  - compute $\mathbf{U}_h^{[1]}$ and $\mathbf{U}_h^{[2]}$ by

$$\begin{aligned} \mathbf{U}_h^{[1]} &= \Lambda_h\big(\mathbf{U}_h^\aleph + \Delta t^\aleph \mathbf{L}_h(\mathbf{U}_h^{[0]}, \mathbf{g}(t^\aleph)) + \Delta t^\aleph \mathbf{R}_h(\mathbf{U}_h^{[0]})\big)\ , \\ \mathbf{w}_h &= \mathbf{U}_h^{[1]} + \Delta t^\aleph \mathbf{L}_h(\mathbf{U}_h^{[1]}, \mathbf{g}_h(t^{\aleph+1})) + \Delta t^\aleph \mathbf{R}_h(\mathbf{U}_h^{[1]})\big)\ , \\ \mathbf{U}_h^{[2]} &= \Lambda_h\big((\mathbf{U}_h^\aleph + \mathbf{w}_h)/2\big); \end{aligned}$$

  - set $\mathbf{U}_h^{\aleph+1} = \mathbf{U}_h^{[2]}$.

In what follows, we describe in detail the approximation of the divergence operator $-\mathbf{L}_h$, the local projection Λ_h, and the right-hand side $\mathbf{R}_h$.

10.2.2.1 The DG Method

The general definition of the DG method for a single partial differential equation can be found in Chap. 4. To define this method in the present case, we simply apply it component by component.

Let $\mathbf{U} = (U_1, U_2, \ldots, U_5)^T$. Consider the equation for the ith component of system (10.20), multiply it by $v \in V_h$, integrate over each $K \in K_h$, replace the exact solution $\mathbf{U}$ by its approximation $\mathbf{U}_h$, and formally integrate by parts to obtain

$$\begin{aligned} &\frac{d}{dt}\int_K U_{ih}(\mathbf{x},t)v(\mathbf{x})\ d\mathbf{x} + \sum_{e\in\partial K}\int_e \mathbf{F}_i(\mathbf{U}_h(\mathbf{x},t))\cdot\boldsymbol{\nu}_e v(\mathbf{x})\ d\ell \\ &\qquad - \int_K \mathbf{F}_i(\mathbf{U}_h(\mathbf{x},t))\cdot\nabla v(\mathbf{x})\ d\mathbf{x} \qquad (10.23) \\ &= \int_K R_i(\mathbf{U}_h(\mathbf{x},t))v(\mathbf{x})\ d\mathbf{x} \qquad \forall v \in V_h,\ i = 1,2,\ldots,5\ , \end{aligned}$$

where $\boldsymbol{\nu}_e$ is the outward unit normal to e. Note that $\mathbf{F}\cdot\boldsymbol{\nu} = \mathbf{F}_{x_1}\nu_{x_1} + \mathbf{F}_{x_2}\nu_{x_2} + \mathbf{F}_{x_3}\nu_{x_3}$ is a five-dimensional vector whose ith component is $\mathbf{F}_i\cdot\boldsymbol{\nu} = (F_{x_1})_i\,\nu_{x_1} + (F_{x_2})_i\,\nu_{x_2} + (F_{x_3})_i\,\nu_{x_3}$. Also, $\mathbf{F}(\mathbf{U}_h(\mathbf{x},t))\cdot\boldsymbol{\nu}_e$ does not have a precise meaning, since $\mathbf{U}_h$ is discontinuous at the interface $e \in \partial K$. Thus we must replace $\mathbf{F}(\mathbf{U}_h(\mathbf{x},t))\cdot\boldsymbol{\nu}_e$ by a suitably chosen *numerical flux* $\mathbf{h}_{e,K}$, which depends on the two values of $\mathbf{U}_h$ on e, $\mathbf{U}_h(\mathbf{x}^{\mathrm{int}(K)},t)$ and $\mathbf{U}_h(\mathbf{x}^{\mathrm{ext}(K)},t)$ defined by

$$\begin{aligned} &\mathbf{U}_h(\mathbf{x}^{\mathrm{int}(K)},t) = \lim_{\mathbf{x}'\to\mathbf{x},\mathbf{x}'\in K^o} \mathbf{U}_h(\mathbf{x}',t)\ , \\ &\mathbf{U}_h(\mathbf{x}^{\mathrm{ext}(K)},t) = \lim_{\mathbf{x}'\to\mathbf{x},\mathbf{x}'\in K^c} \mathbf{U}_h(\mathbf{x}',t) \quad \text{if } \mathbf{x}\notin\Gamma\ , \\ &\mathbf{B}\mathbf{U}_h(\mathbf{x}^{\mathrm{ext}(K)},t) = \mathbf{g}_h(\mathbf{x},t) \quad \text{if } \mathbf{x}\in\Gamma\ , \end{aligned}$$

where K^o denotes the interior of K and $K^c = \Omega\setminus K$. The choice of the numerical flux $\mathbf{h}_{e,K}$ is crucial since it is through the use of this flux that we introduce the upwinding (or artificial viscosity) which renders the method stable (without destroying its high-order accuracy). In Chap. 4, the numerical fluxes were based on various stabilization (penalty) terms. In this section, we use a local *Lax-Friedrichs* flux to be described in Sect. 10.2.2.4.

Finally, we replace the integrals in (10.23) by quadrature rules:

$$\begin{aligned} &\int_e h_{i,e,R}(\mathbf{x},t)v(\mathbf{x})\ d\ell \simeq \sum_{l=1}^{L_1}\omega_l h_{i,e,R}(\mathbf{x}_l,t)v(\mathbf{x}_l)|e|\ , \\ &\int_K \mathbf{F}_i(\mathbf{U}_h(\mathbf{x},t))\cdot\nabla v(\mathbf{x})\ d\mathbf{x} \simeq \sum_{l=1}^{L_2}\widehat{\omega}_l\mathbf{F}_i(\mathbf{U}_h(\mathbf{x}_l,t))\cdot\nabla v(\mathbf{x}_l)|K|\ , \end{aligned}$$

where

$$\begin{aligned}\mathbf{h}_{e,K}(\mathbf{x},t) &= \mathbf{h}_{e,K}(\mathbf{U}_h(\mathbf{x}^{\mathrm{int}(K)},t),\mathbf{U}_h(\mathbf{x}^{\mathrm{ext}(K)},t))\\ &= (h_{1,e,K},h_{2,e,K},\ldots,h_{5,e,K})\ ,\end{aligned}$$

ω_l and $\widehat{\omega}_l$ are integration weights, L_1 and L_2 are the numbers of integration points, $|e|$ is the area of e, and $|K|$ is the volume of K.

In this way, the weak formulation for approximating the solution to (10.20) is: For each $i = 1, 2, \ldots, 5$,

$$\begin{aligned}&\frac{d}{dt}\int_K U_{ih}(\mathbf{x},t)v(\mathbf{x})\ d\mathbf{x} + \sum_{e\in\partial K}\sum_{l=1}^{L_1}\omega_l h_{i,e,R}(\mathbf{x}_l,t)v(\mathbf{x}_l)|e|\\ &\quad -\sum_{l=1}^{L_2}\widehat{\omega}_l\mathbf{F}_i(\mathbf{U}_h(\mathbf{x}_l,t))\cdot\nabla v(\mathbf{x}_l)|K| \qquad (10.24)\\ &= \sum_{l=1}^{L_2}\widehat{\omega}_l R_i(\mathbf{U}_h(\mathbf{x}_l,t))v(\mathbf{x}_l)|K| \qquad \forall v\in V_h,\ K\in K_h\ .\end{aligned}$$

This weak formulation defines the operators $\mathbf{L}_h$ and $\mathbf{R}_h$.

10.2.2.2 The Degrees of Freedom for $\mathbf{U}_h$

The weak formulation (10.24) is completely independent of the way in which we choose to express our approximate solution $\mathbf{U}_h$. We choose to express $\mathbf{U}_h$ as follows: For $\mathbf{x}$ in the rectangular parallelepiped

$$\begin{aligned}R = &\left(x_{1i}-\frac{h_{1i}}{2},x_{1i}+\frac{h_{1i}}{2}\right)\times\left(x_{2j}-\frac{h_{2j}}{2},x_{2j}+\frac{h_{2j}}{2}\right)\\ &\times\left(x_{3k}-\frac{h_{3k}}{2},x_{3k}+\frac{h_{3k}}{2}\right),\end{aligned} \qquad (10.25)$$

we write

$$\begin{aligned}\mathbf{U}_h(\mathbf{x}) = \bar{\mathbf{U}}_{ijk} &+ \left(\frac{x_1-x_{1i}}{h_{1i}/2}\right)\tilde{\mathbf{U}}_{x_1\,ijk} + \left(\frac{x_2-x_{2j}}{h_{2j}/2}\right)\tilde{\mathbf{U}}_{x_2\,ijk}\\ &+\left(\frac{x_3-x_{3k}}{h_{3k}/2}\right)\tilde{\mathbf{U}}_{x_3\,ijk}\ ;\end{aligned}$$

that is, we choose as degrees of freedom of $\mathbf{U}_h$, its mean on K, $\bar{\mathbf{U}}_{ijk}$ and its variation in the x_l-direction, $\tilde{\mathbf{U}}_{x_l\,ijk}$, $l = 1, 2, 3$.

This choice of the degrees of freedom renders the mass matrix of the weak formulation (10.24) a 4×4 diagonal matrix. More important, this choice greatly facilitates the evaluation of the numerical flux $\mathbf{h}_{e,K}$ (cf. Sect. 10.2.2.4) and the computation of the nonlinear projection Λ_h that then becomes equivalent to three one-dimensional nonlinear projections (cf. Sect. 10.2.2.3).

In order not to make a distinction between boundary and interior faces in the evaluation of the numerical flux $\mathbf{h}_{e,K}$ and in the computation of the nonlinear projection Λ_h, we express the boundary values as follows: Suppose that $\mathbf{x}$ lies on the boundary face $\{x_{1i} + h_{1i}/2\} \times (x_{2j} - h_{2j}/2, x_{2j} + h_{2j}/2) \times (x_{3k} - h_{3k}/2, x_{3k} + h_{3k}/2)$, for example; then we write the boundary values of $\mathbf{U}_h$ as

$$\mathbf{U}_h(\mathbf{x}) = \bar{\mathbf{U}}_{i+1,jk} + \left(\frac{x_2 - x_{2j}}{h_{2j}/2}\right) \tilde{\mathbf{U}}_{x_2\, i+1,jk} + \left(\frac{x_3 - x_{3k}}{h_{3k}/2}\right) \tilde{\mathbf{U}}_{x_3\, i+1,jk} \,.$$

We use similar representations of the boundary values of $\mathbf{U}_h$ on the boundary face $(x_{1i} - h_{1i}/2, x_{1i} + h_{1i}/2) \times \{x_{2j} + h_{2j}/2\} \times (x_{3k} - h_{3k}/2, x_{3k} + h_{3k}/2)$ and on other faces.

10.2.2.3 The Local Projection Λ_h

The local projection Λ_h is used to prevent the appearance of spurious oscillations in the approximate solution. The local averages, $\bar{\mathbf{U}}_{ijk}$, are unchanged to preserve the conservation property of the DG method, but the local variation in the x_l-direction, $\tilde{\mathbf{U}}_{x_l\, ijk}$, $l = 1, 2, 3$, must be controlled to avoid unwanted oscillations. We can obtain some control on the oscillations along the x_1-direction with the following component by component algorithm. For the lth component of the approximate solution, we set

$$\left(\tilde{U}^{(\mathrm{mod})}_{x_1\, ijk}\right)_l = \mathrm{minmod}\left(\left(\tilde{U}_{x_1\, ijk}\right)_l, \left(\Delta_{x_1+}\bar{U}_{ijk}\right)_l, \left(\Delta_{x_1-}\bar{U}_{ijk}\right)_l\right) \,,$$

where Δ_{x_1+} and Δ_{x_1-} denote the standard forward and backward finite difference operators in the x_1-direction and

$$\mathrm{minmod}(a, b_1, b_2, \cdots, b_\imath) = \begin{cases} a, & \text{if } |a| \le \mathcal{M} h_{1i}^2 \,, \\ (\mathrm{sign}\ a) \min\limits_{1 \le \imath_1 \le \imath} \{|a|, |b_{\imath_1}|\} \,, & \\ \quad \text{if sign } a = \text{sign } b_{\imath_1}, 1 \le \imath_1 \le \imath \,, & \\ 0, & \text{otherwise} \,, \end{cases}$$

where $\imath$ is an integer and $\mathcal{M}$ is an upper bound of the second-order x_1-derivative of each of the components of $\mathbf{U}$. To control the oscillation along the x_2- and x_3-directions, a similar algorithm can be used. However, although this algorithm is computationally efficient, it does not take into account the physically relevant directions along which the information travels. Taking these characteristic directions into account results in a better control of the oscillations and in a higher quality of the approximation.

Let us show how to do this to define $\tilde{\mathbf{U}}^{(\mathrm{mod})}_{x_1\, ijk}$. First, we compute the Jacobians

$$\mathbf{J}_{ijk} = \left(\frac{\partial \mathbf{F}}{\partial \mathbf{U}}\right) \cdot (1,0,0)(\bar{\mathbf{U}}_{ijk}) \ ,$$

and obtain their eigenvalues, $\lambda_{ijk}^{(l)}$, and their left and right eigenvectors, $\mathbf{l}_{ijk}^{(l)}$ and $\mathbf{r}_{ijk}^{(l)}$, $l = 1, 2, \ldots, 5$, respectively. The eigenvectors are normalized so that $\mathbf{l}_{ijk}^{(l_1)} \cdot \mathbf{r}_{ijk}^{(l_2)} = \delta_{l_1 l_2}$. Then we project $\tilde{\mathbf{U}}_{x_1\, ijk}$, $\Delta_{x_1+}\bar{\mathbf{U}}_{ijk}$, and $\Delta_{x_1-}\bar{\mathbf{U}}_{ijk}$ into the eigenspace of $\mathbf{J}_{ijk}$:

$$\begin{aligned} a^{(l)} &= \mathbf{l}_{ijk}^{(l)} \cdot \tilde{\mathbf{U}}_{x_1\, ijk}, \qquad l = 1, 2, \ldots, 5 \ , \\ b^{(l)} &= \mathbf{l}_{ijk}^{(l)} \cdot \Delta_{x_1+}\bar{\mathbf{U}}_{ijk}, \quad l = 1, 2, \ldots, 5 \ , \\ c^{(l)} &= \mathbf{l}_{ijk}^{(l)} \cdot \Delta_{x_1-}\bar{\mathbf{U}}_{ijk}, \quad l = 1, 2, \ldots, 5 \ , \end{aligned}$$

and perform the projection (or *slope limiting*) in each characteristic field

$$\tilde{U}_{x_1\, ijk}^{(l)} = \text{minmod}(a^{(l)}, b^{(l)}, c^{(l)}) \ .$$

Next, we project them back to the component space to obtain

$$\tilde{\mathbf{U}}_{x_1\, ijk}^{(\text{mod})} = \sum_{l=1}^{5} \tilde{U}_{x_1\, ijk}^{(l)}\, \mathbf{r}_{ijk}^{(l)} \ .$$

This completes the projection in the x_1-direction. Similar and totally independent procedures can be applied in the x_2- and x_3-directions.

10.2.2.4 The Numerical Flux $\mathbf{h}_{e,K}$

The numerical flux we use is the (componentwise) local Lax-Friedrichs flux. Suppose that K is a rectangular parallelepiped as given in (10.25) and that the quadrature point $\mathbf{x}_l$ lies on the face $\{x_{1i} + h_{1i}/2\} \times (x_{2j} - h_{2j}/2, x_{2j} + h_{2j}/2) \times (x_{3k} - h_{3k}/2, x_{3k} + h_{3k}/2)$, for example. Then we define

$$\begin{aligned} \mathbf{h}_{e,K}(\mathbf{x}_l) = \frac{1}{2}\Big\{ &\Big(\mathbf{F}(\mathbf{U}_h(\mathbf{x}_l^{\text{int}(K)})) \cdot \boldsymbol{\nu}_e + \mathbf{F}(\mathbf{U}_h(\mathbf{x}_l^{\text{ext}(K)}) \cdot \boldsymbol{\nu}_e\Big) \\ &-\alpha_e \Big(\mathbf{U}_h(\mathbf{x}_l^{\text{ext}(K)}) - \mathbf{U}_h(\mathbf{x}_l^{\text{int}(K)})\Big)\Big\} \ , \end{aligned}$$

where

$$\alpha_e = \max\{\, \lambda(\bar{\mathbf{U}}_{ijk}), \lambda(\bar{\mathbf{U}}_{i+1,jk}) \,\} \ ,$$

and

$$\lambda((n, p_{x_1}, p_{x_2}, p_{x_3}, w)^T) = \sqrt{\frac{5}{3}\, T/m} + |\mathbf{v}|$$

is an upper bound for the eigenvalues of the Jacobian of $\mathbf{F} \cdot \mathbf{n}$ evaluated at $(n, p_{x_1}, p_{x_2}, p_{x_3}, w)^T$ for all unit vectors $\mathbf{n}$. Similar expressions hold for other quadrature points on the boundary of K.

A characteristically evaluated local Lax-Friedrichs flux can be also used. However, our experience is that the componentwise evaluated local Lax-Friedrichs flux produces as good results as this more costly flux.

10.2.2.5 The Right-Hand Side $\mathbf{R}(\mathbf{U}_h)$

Finally, we show how to evaluate the function $\mathbf{R}(\mathbf{U}_h) = \boldsymbol{\xi}_{\mathbf{E}}(\mathbf{U}_h) + \boldsymbol{\xi}_c(\mathbf{U}_h) + \boldsymbol{\xi}_{heat}(\mathbf{U}_h)$ for a given $\mathbf{U}_h$. To evaluate $\boldsymbol{\xi}_c(\mathbf{U}_h)$, we simply use (10.8) and the following equations:

$$\mathbf{p} = mn\mathbf{v}, \quad w = \frac{3}{2}n\kappa_B T + \frac{1}{2}mn|\mathbf{v}|^2. \tag{10.26}$$

To evaluate $\boldsymbol{\xi}_{\mathbf{E}}(\mathbf{U}_h)$, we need a numerical method to obtain an approximation $\mathbf{E}_h$ to the electric field $\mathbf{E}$. The equations defining the electric field are (10.7) and the boundary conditions:

$$\begin{aligned} &\mathbf{E} = -\nabla\phi && \text{in } \Omega\,, \\ &\nabla\cdot(\epsilon\mathbf{E}) = e\,(N_D - N_A - n) && \text{in } \Omega\,, \\ &\phi = \phi_D && \text{on } \Gamma_D\,, \\ &\mathbf{E}\cdot\boldsymbol{\nu} = 0 && \text{on } \Gamma_N, \end{aligned} \tag{10.27}$$

where $\bar{\Gamma} = \bar{\Gamma}_D \cup \bar{\Gamma}_N$ and $\Gamma_D \cap \Gamma_N = \emptyset$. These equations can be discretized using the mixed finite element method as in (10.14). As an example, we use the lowest-order Raviart-Thomas mixed method on rectangular parallelepipeds that defines the approximation $\mathbf{E}_h \in \mathbf{V}_h$ and $\phi_h \in W_h$ as the solution of the following equations:

$$\begin{aligned} &(\mathbf{E}_h, \mathbf{w}) - (\phi_h, \nabla\cdot\mathbf{w}) = -\left(\phi_D, \mathbf{w}\cdot\boldsymbol{\nu}\right)_{\Gamma_D} && \forall \mathbf{w} \in \mathbf{V}_h\,, \\ &(\epsilon\nabla\cdot\mathbf{E}_h, v) = (e(N_D - N_A - n_h), v) && \forall v \in W_h\,, \end{aligned}$$

where n_h is the approximate density yielded by (10.24) and

$$\begin{aligned} \mathbf{V}_h = \big\{ &\mathbf{w} \in \mathbf{H}(\mathrm{div};\Omega) : \mathbf{w}|_K = (a_K^1 + a_K^2 x_1, a_K^3 + a_K^4 x_2, a_K^5 + a_K^6 x_3)\,, \\ &\qquad a_K^i \in \mathbb{R},\ K \in K_h;\ \mathbf{w}\cdot\boldsymbol{\nu}|_{\Gamma_N} = 0 \big\}\,, \\ W_h = \big\{ &v \in L^2(\Omega) : v|_K \text{ is a constant},\ K \in K_h \big\}\,. \end{aligned}$$

To evaluate $\boldsymbol{\xi}_{heat}(\mathbf{U}_h)$, we also use the Raviart-Thomas space, although in a very different way. Note that

$$\boldsymbol{\xi}_{heat}(\mathbf{U}) = (0, 0, 0, \nabla\cdot\mathbf{q}))^T\,,$$

where $\mathbf{q}$ is defined by

$$\begin{aligned} &\mathbf{q} = \kappa\,\nabla T && \text{in } \Omega\,, \\ &T = T_D && \text{on } \Gamma'_D\,, \\ &\mathbf{q}\cdot\boldsymbol{\nu} = 0 && \text{on } \Gamma'_N\,, \end{aligned}$$

where $\bar{\Gamma} = \bar{\Gamma}'_D \cup \bar{\Gamma}'_N$, $\Gamma'_D \cap \Gamma'_N = \emptyset$, and Γ'_D and Γ'_N are not necessarily the same as Γ_D and Γ_N. Then we define the approximation $\mathbf{q}_h \in \mathbf{Q}_h$ as the solution of the following problem:

$$(\kappa_h^{-1}\mathbf{q}_h, \mathbf{w}) = -(T_h, \nabla\cdot\mathbf{w}) + (T_D, \mathbf{w}\cdot\boldsymbol{\nu})_{\Gamma_D'} \quad \forall \mathbf{w}\in\mathbf{Q}_h \ ,$$

where κ_h and T_h are given by (10.9) and the second equation of (10.26), respectively, and

$$\mathbf{Q}_h = \big\{\mathbf{w}\in\mathbf{H}(\mathrm{div};\Omega) : \mathbf{w}|_K = (a_K^1 + a_K^2 x_1, a_K^3 + a_K^4 x_2, a_K^5 + a_K^6 x_3) \ , \\ a_K^i \in \mathbb{R},\ K\in K_h;\ \mathbf{w}\cdot\boldsymbol{\nu}|_{\Gamma_N'} = 0\big\} \ .$$

This completes the definition of $\mathbf{R}(\mathbf{U}_h)$ and thus the definition of (10.24). An example of numerical results will be presented in Sect. 10.3.

10.2.3 The Quantum Hydrodynamic Model

Again, to devise numerical methods for solving the quantum hydrodynamic model, we write it in a conservation law format. Let

$$\tilde{w} = w + \frac{\hbar^2 n}{24m}\Delta\log(n) \ ;$$

then, by (10.11), we see that

$$\tilde{w} = \frac{3}{2}n\kappa_B T + \frac{1}{2}mn|\mathbf{v}|^2 \ ,$$

which has the same form as that of the energy band for the hydrodynamic model. Also, after introducing the quantum potential of Bohm (Philippidis et al., 1982)

$$Q = -\frac{\hbar^2}{2m}\frac{1}{\sqrt{n}}\Delta\sqrt{n} \ ,$$

we find that the stress tensor satisfies the relation (cf. Exercise 10.3)

$$-\nabla\cdot\mathbf{P} = \nabla(n\kappa_B T) + \frac{n}{3}\nabla Q \ . \tag{10.28}$$

Using these new definitions, it follows from (10.10) that

$$\begin{aligned}
&\frac{\partial n}{\partial t} + \nabla\cdot(n\mathbf{v}) = 0 \ , \\
&\frac{\partial \mathbf{p}}{\partial t} + \mathbf{v}\nabla\cdot\mathbf{p} + \mathbf{p}\cdot\nabla\mathbf{v} + \nabla(n\kappa_B T) = -en\mathbf{E} + \left(\frac{\partial \mathbf{p}}{\partial t}\right)_c \\
&\qquad\qquad + \mathbf{p}_Q \ , \\
&\frac{\partial \tilde{w}}{\partial t} + \nabla\cdot(\mathbf{v}\tilde{w}) + \nabla\cdot(\mathbf{v}n\kappa_B T) = -en\mathbf{v}\cdot\mathbf{E} + \left(\frac{\partial \tilde{w}}{\partial t}\right)_c \\
&\qquad\qquad + \nabla\cdot(\kappa\nabla T) + \tilde{w}_Q \ ,
\end{aligned} \tag{10.29}$$

where

$$\mathbf{p}_Q = -\frac{n}{3}\nabla Q \ ,$$

$$\tilde{w}_Q = \frac{\hbar^2 n}{24m}\frac{\Delta \log(n)}{\tau_{\tilde{w}}} - \frac{\hbar^2}{24m}\nabla\cdot(n\Delta\mathbf{v}) - \frac{n}{3}\mathbf{v}\cdot\nabla Q \ .$$

In this way, (10.29) are the classical hydrodynamic equations with the addition of the new terms $\mathbf{p}_Q$ and $\tilde{w}_Q$, which come from the quantum correction terms. As a consequence, we only need to treat these new terms for the quantum model if the finite element programs introduced for the hydrodynamic model are applied.

Let $\mathbf{U} = (n, p_{x_1}, p_{x_2}, p_{x_3}, \tilde{w})^T$. Then (10.29) can be written as follows:

$$\frac{\partial \mathbf{U}}{\partial t} + \nabla\cdot\mathbf{F}(\mathbf{U}) = \mathbf{R}(\mathbf{U}) \ , \tag{10.30}$$

where the flux $\mathbf{F} = (\mathbf{F}_{x_1}, \mathbf{F}_{x_2}, \mathbf{F}_{x_3})$ has the following components:

$$\begin{aligned}
\mathbf{F}_{x_1}(\mathbf{U}) &= v_{x_1}\mathbf{U} + (0, n\kappa_B T, 0, 0, v_{x_1} n\kappa_B T)^T \ , \\
\mathbf{F}_{x_2}(\mathbf{U}) &= v_{x_2}\mathbf{U} + (0, 0, n\kappa_B T, 0, v_{x_2} n\kappa_B T)^T \ , \\
\mathbf{F}_{x_3}(\mathbf{U}) &= v_{x_3}\mathbf{U} + (0, 0, 0, n\kappa_B T, v_{x_3} n\kappa_B T)^T \ ,
\end{aligned}$$

and the right-hand side $\mathbf{R}$ is given by

$$\begin{aligned}
\mathbf{R}(\mathbf{U}) &= \boldsymbol{\xi}_{\mathbf{E}}(\mathbf{U}) + \boldsymbol{\xi}_c(\mathbf{U}) + \boldsymbol{\xi}_{heat}(\mathbf{U}) + \boldsymbol{\xi}_Q(\mathbf{U}) \ , \\
\boldsymbol{\xi}_{\mathbf{E}}(\mathbf{U}) &= (0, -e\,n\,E_{x_1}, -e\,n\,E_{x_2}, -e\,n\,E_{x_3}, -e\,n\,\mathbf{v}\cdot\mathbf{E})^T \ , \\
\boldsymbol{\xi}_c(\mathbf{U}) &= \left(0, \left(\frac{\partial p_{x_1}}{\partial t}\right)_c, \left(\frac{\partial p_{x_2}}{\partial t}\right)_c, \left(\frac{\partial p_{x_3}}{\partial t}\right)_c, \left(\frac{\partial \tilde{w}}{\partial t}\right)_c\right)^T \ , \\
\boldsymbol{\xi}_{heat}(\mathbf{U}) &= (0, 0, 0, 0, \nabla\cdot(\kappa\,\nabla T))^T \ , \\
\boldsymbol{\xi}_Q(\mathbf{U}) &= (0, (p_Q)_{x_1}, (p_Q)_{x_2}, (p_Q)_{x_3}, \tilde{w}_Q) \ .
\end{aligned}$$

Thus we see that the numerical techniques developed for the classical hydrodynamic model in Sect. 10.2.2 can be applied to the quantum hydrodynamic model (10.30) (Chen et al., 1995).

10.3 A Numerical Example

A two-dimensional MESFET device $\Omega = (0, 0.6 \quad \mathrm{m}) \times (0, 0.2 \quad \mathrm{m})$ is simulated, using the hydrodynamic model (10.6)–(10.9). The source, drain, and gate are, respectively, the segments $(0, 0.1 \quad \mathrm{m}) \times \{\mathrm{x}_2 = 0.2 \quad \mathrm{m}\}$, $(0.5 \quad \mathrm{m}, 0.6 \quad \mathrm{m}) \times \{\mathrm{x}_2 = 0.2 \quad \mathrm{m}\}$, and $(0.2 \quad \mathrm{m}, 0.4 \quad \mathrm{m}) \times \{\mathrm{x}_2 = 0.2 \quad \mathrm{m}\}$; see Fig. 10.1. The doping N_D is defined by

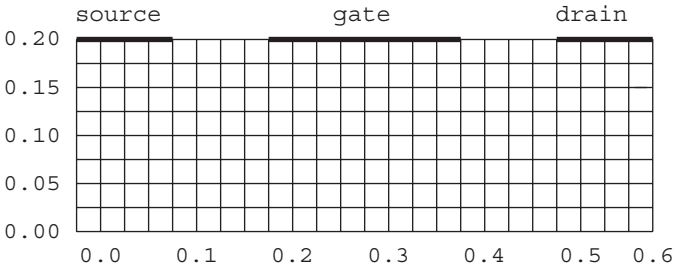

Fig. 10.1. An MESFET semiconductor device

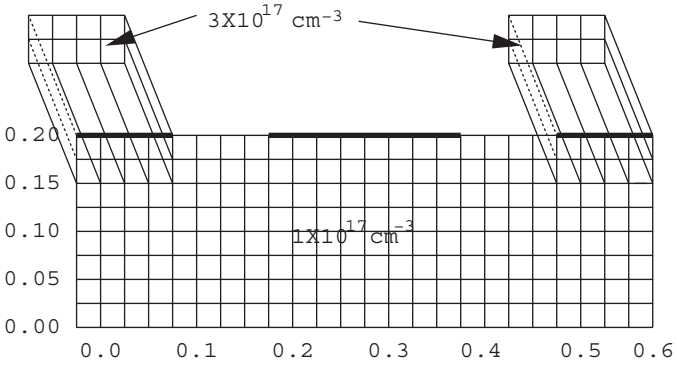

Fig. 10.2. The doping profile N_D

$$N_D = \begin{cases} 3 \times 10^{17}\,\mathrm{cm}^{-3}, & (x_1, x_2) \in [0, 0.1] \times [0.15, 0.2] \\ & \quad \cup [0.5, 0.6] \times [0.15, 0.2]\,, \\ 1 \times 10^{17}\,\mathrm{cm}^{-3}, & \text{elsewhere}\,, \end{cases}$$

and $N_A = 0$ in the whole domain Ω; see Fig. 10.2.

The initial conditions are chosen as $n = N_D$ for the density, $T = T_0 = 300°\mathrm{K}$ for the temperature, and $v_{x_1} = v_{x_2} = 0$ for the velocity. The initial condition for the potential is ϕ_0^*, where $\phi_0^* = \phi_0 - .232$, and

$$\phi_0 = kT_0 \ln\left(\frac{N_D}{n_i}\right) \Big/ e\,,$$

with $k = 0.138 \times 10^{-4}$, $e = 0.1602$, and $n_i = 0.000018$ (for GaAs). We are employing a translation constant 0.232 in ϕ_0 for convenience in our simulations. The boundary conditions are defined as follows:

- **At the source**: $\phi = \phi_0^*$ for the potential, $n = 3 \times 10^{17}\,\mathrm{cm}^{-3}$ for the electron density, $T = 300°\mathrm{K}$ for the temperature, $v_{x_1} = 0$ m/ps for the horizontal velocity, and the homogeneous Neumann boundary condition for the vertical velocity v_{x_2};

- **At the drain**: $\phi = \phi_0^* + 2$ for the potential, $n = 3 \times 10^{17}\,\mathrm{cm}^{-3}$ for the electron density, $T = 300°\mathrm{K}$ for the temperature, $v_{x_1} = 0$ m/ps for the horizontal velocity, and the homogeneous Neumann boundary condition for the vertical velocity v_{x_2};
- **At the gate**: $\phi = \phi_0^* - 0.8$ for the potential, $n = 3.9 \times 10^5\,\mathrm{cm}^{-3}$ for the electron density, $T = 300°\mathrm{K}$ for the temperature, $v_{x_1} = 0$ m/ps for the horizontal velocity, and the homogeneous Neumann boundary condition for the vertical velocity v_{x_2};
- **At all other parts of the boundary**: all variables are subjected to homogeneous Neumann boundary conditions.

We simulate homogeneous Neumann boundary conditions for each of the components of $\mathbf{U}_h$ as follows. Suppose, for example, that the edge $\{x_{1i} + h_{1i}/2\} \times (x_{2j} - h_{2j}/2, x_{2j} + h_{2j}/2)$ lies on a Neumann boundary for, say, the lth component U_{lh}. Then we define the degrees of freedom of U_{lh} at this boundary as

$$(\bar{U}_{i+1,j})_l = (\bar{U}_{ij})_l, \qquad (\tilde{U}_{x_2\, i+1,j})_l = (\tilde{U}_{x_2\, ij})_l\ .$$

Similar expressions are used on the other edges with a Neumann boundary condition. A uniform space mesh of 96×32 is employed for the simulations in which the method in Sect. 10.2.2 is run until the steady state is reached. The numerical results for the density n, velocity $\mathbf{v}$, energy w, temperature T, electric potential ϕ, and electric field $\mathbf{E}$ are displayed in Figs. 10.3–10.9. Notice the sharp transition of the electron density n near the junctions. Also, note that there is a boundary layer for n at the drain, but not at the source. This is reasonable since the drain is an outflow boundary and the source is an inflow boundary. A rapid drop of n at the depletion region occurs near the gate. The normal velocity component at the gate appears to be negligible, while the horizontal component shows an evidence of strong carrier movement

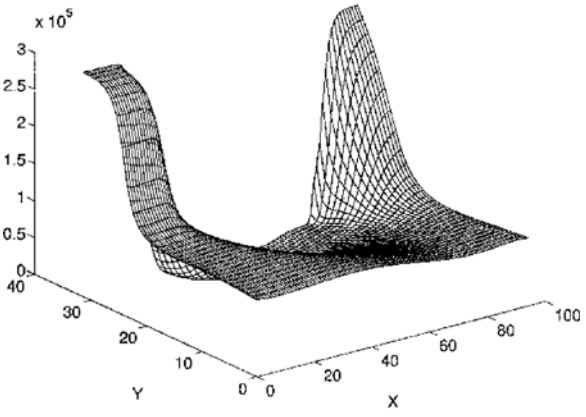

Fig. 10.3. The density n_h

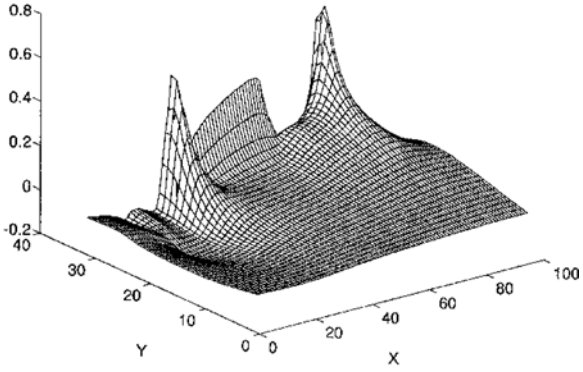

Fig. 10.4. The x_2-component of the velocity

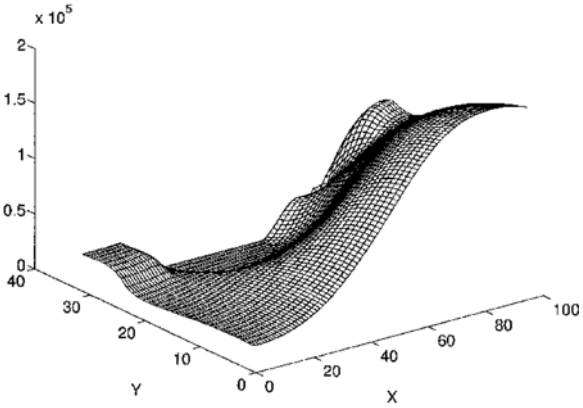

Fig. 10.5. The energy w_h

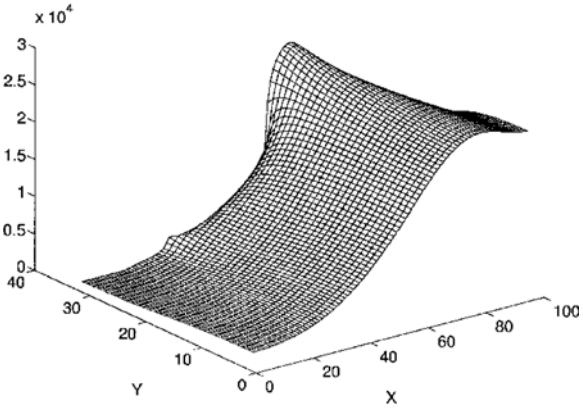

Fig. 10.6. The temperature T_h

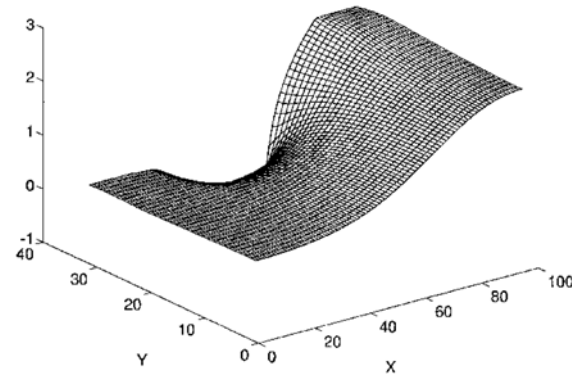

Fig. 10.7. The electric potential ϕ_h

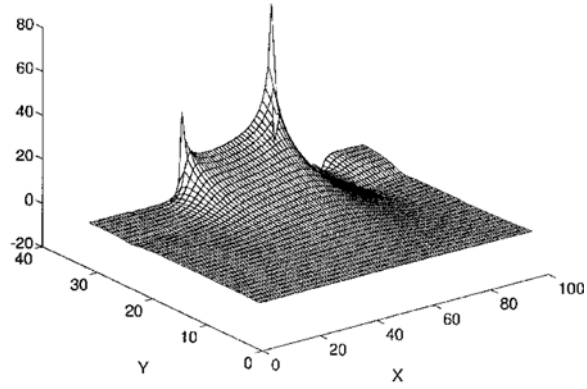

Fig. 10.8. The x_2-component of the electric filed

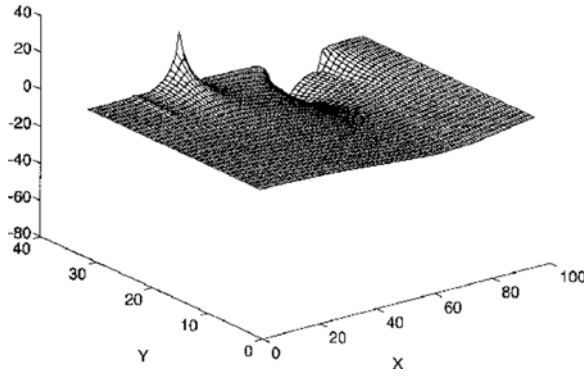

Fig. 10.9. The x_1-component of the electric filed

toward the source beneath the left gate area, and of strong movement toward the drain immediately to the left of the drain junction. Notice the cusps and strong gradients in the components of the velocity. The junction layers and the interface layers are also clearly visible in the energy density w and the potential ϕ. The peaks of the electric field are due to its singularities around the intersections of the Dirichlet and Neumann segments of the boundary.

10.4 Bibliographical Remarks

For more information on the derivation and properties of the semiconductor models presented in Sect. 10.1, the reader should refer to Markowich (1986) and Markowich et al. (1990). The analysis of the MMOC procedure discussed in Sect. 10.2.1 for the drift-diffusion model can be found in Douglas et al. (1986). The development of the approximation procedure for the hydrodynamic model described in Sect. 10.2.2 and the numerical results presented in Sect. 10.3 follow Chen et al. (1995B), and this procedure is based on the Runge-Kutta discontinuous Galerkin method (Cockburn et al., 1990). For a similar approximation procedure and the corresponding numerical results for the quantum hydrodynamic model considered in Sect. 10.2.3, see Chen et al. (1995A).

10.5 Exercises

10.1. Formulate an MMOC procedure for the electron and hole concentration equations of the drift-diffusion model with varying mobilities: $\mu_n = \mu_n(\mathbf{E})$ and $\mu_p = \mu_p(\mathbf{E})$. In this procedure, use linear extrapolation techniques in the approximation of $\mathbf{E}$ in these mobilities (cf. Sect. 9.3).

10.2. Write down a mixed variational formulation for the electric potential and field equations

$$\begin{aligned} &\mathbf{E} = -\nabla\phi && \text{in } \Omega\,, \\ &\nabla\cdot(\epsilon\mathbf{E}) = e\,(N_D - N_A - n) && \text{in } \Omega\,, \\ &\phi = \phi_D && \text{on } \Gamma_D\,, \\ &\mathbf{E}\cdot\boldsymbol{\nu} = 0 && \text{on } \Gamma_N\,, \end{aligned}$$

where $\bar{\Gamma} = \bar{\Gamma}_D \cup \bar{\Gamma}_N$ and $\Gamma_D \cap \Gamma_N = \emptyset$.

10.3. Verify (10.28) in detail.

10.4. In this chapter, we have not presented any theoretical analysis. As a matter of fact, the numerical procedure defined in Sect. 10.2.1 can be analyzed as in Sect. 9.5. Carry out an analysis for this procedure. (If necessary, consult Douglas et al. (1986).)

A Nomenclature

$\mathbf{A}$	coefficient matrix of a system (stiffness matrix)
$\mathbf{a}$	diffusion coefficient
$a(\cdot,\cdot)$	bilinear form
$a_h(\cdot,\cdot)$	mesh-dependent bilinear form
$a_K(\cdot,\cdot)$	restriction of $a(\cdot,\cdot)$ on K
$a_-(\cdot,\cdot)$	bilinear form for symmetric DG
$a_+(\cdot,\cdot)$	bilinear form for nonsymmetric DG
a_{ij}	entries of $\mathbf{A}$
a_{ij}^K	restriction of a_{ij} on K (element)
$\mathbf{B}$	mass matrix
$\mathbf{B}_1$	matrix in an affine mapping
$b(\cdot,\cdot)$	bilinear form
$\mathbf{b}$	convection or advection coefficient
b_e	edge bubble function
b_K	triangle bubble function
c	reaction coefficient
$\mathbf{C}$	coefficient matrix associated with time
d	dimension number ($d = 1$, 2, or 3)
$\mathcal{E}_h^D$	set of edges on Γ_D
$\mathcal{E}_h^N$	set of edges on Γ_N
$\mathcal{E}_h$	set of edges on Γ
F	functional or total potential energy
$\mathbf{F}$	mapping
f	right-hand function or load
$\mathbf{f}$	right-hand vector of a system
f_K	local mean value on K
f_i	ith entry of $\mathbf{f}$
f_α	fractional flow function
$\mathbf{G}$	Jacobian matrix or a mapping

g	boundary datum
g_e	local mean value on e
h	mesh or grid size
h_e	length of edge e
h_k	mesh size at the kth level
h_K	diameter of K (element)
I	interval in $\mathbb{R}$
I_i	subintervals
$\mathbf{I}$	identity matrix or operator
$\check{I}_i(t)$	trace-back of I_i to time t
$\mathcal{I}_i^n$	space-time region following characteristics
J	time interval of interest ($J = (0, T]$)
J^n	nth subinterval of time (t^{n-1}, t^n)
$\mathbf{J}_n$	electron current density
$\mathbf{J}_p$	hole current density
K	element (triangle, rectangle, etc.)
$\hat{K}$	reference element
K_h	triangulation (partition)
$\check{K}(t)$	trace-back of K to time t
$\mathcal{K}^n$	space-time region following characteristics
$L(\cdot)$	linear functional
$L_-(\cdot)$	linear functional for symmetric DG
$L_+(\cdot)$	linear functional for nonsymmetric DG
L_h	space of Lagrange multipliers
L_i	band width of ith row
$\mathbf{L}$	lower triangular matrix
$\mathcal{L}$	linear operator
M	number of grid points (nodes)
$\mathbf{M}$	coefficient matrix arising from mixed method
$\mathbf{m}_i$	vertices of elements
$\mathcal{N}_h$	set of vertices in K_h
p	unknown variable or pressure
p_c	capillary pressure
$\mathbf{p}$	unknown vector of a system
p_h	approximate solution
p_0	initial datum
$\check{p}_h^{n-1}$	value of p_h at $(\check{x}_n, t^{n-1})$: $p_h\left(\check{x}_n, t^{n-1}\right)$
$\tilde{p}_h$	interpolant of p_h
P_r	set of polynomials of total degree $\leq r$
$P_{l,r}$	set of polynomials defined on prisms
p_α	pressure of α-phase
$\mathbf{P}$	pressure tensor
R	reaction coefficient
Re	Reynolds number
R_h	projection operator

$\mathcal{R}_{D,\mathbf{m}}$	local Dirichlet estimator I
$\mathcal{R}_{D,K}$	local Dirichlet estimator II
$\mathcal{R}_H$	hierarchical basis estimator
$\mathcal{R}_K$	residual a-posteriori estimator
$\mathcal{R}_{N,K}$	local Neumann estimator
$\mathcal{R}_Z$	averaging-based estimator
$\mathbf{q}$	heat flux
q_α	source/sink term of α-phase
Q_r	set of polynomials of degree $\leq r$ in each variable
$\mathbf{Q}$	upper triangular matrix
s_α	saturation of α-phase
$\mathbf{t}$	tangential vector
t	time variable
t^n	nth time step
T	final time
u	velocity variable in $\mathbb{R}$
$\mathbf{u}$	velocity variable in $\mathbb{R}^d$
$\mathbf{u}_\alpha$	velocity of α-phase
$\mathbf{U}$	unknown vector for u or $\mathbf{u}$
U_T	thermal voltage
v_-	left-hand limit notation
v_+	right-hand limit notation
V	linear vector space
V'	dual space to V
V_h	finite element space
$\mathbf{V}$	vector space in a pair of mixed spaces
$\mathbf{V}_h$	vector space in a pair of mixed finite element spaces
$\mathbf{V}_h(K)$	restriction of $\mathbf{V}_h$ on K
w_i	integration weight
W	scalar space in a pair of mixed spaces
W_h	scalar space in a pair of mixed finite element spaces
$W_h(K)$	restriction of W_h on K
x	independent variable in $\mathbb{R}$
$\mathbf{x}$	independent variable in $\mathbb{R}^d$: $\mathbf{x} = (x_1, x_2, \ldots, x_d)$
$\check{x}_n$	foot of a characteristic corresponding to x at t^n
Z	subspace of V induced by $b(\cdot,\cdot)$
$Z^\perp$	orthogonal complement of Z
Z^0	polar set of Z
Z_h	discrete counterpart of Z
$\mathrm{cond}(\mathbf{A})$	condition number of $\mathbf{A}$
$\mathbb{R}$	set of real numbers
Ω	open set in $\mathbb{R}^d$ ($d = 2$ or 3)
$\bar{\Omega}$	closure of Ω

Ω_e	union of elements with common edge e
Ω_K	union of elements adjacent to K
$\Omega_{\mathbf{m}}$	union of elements with common vertex $\mathbf{m}$
Γ	boundary of Ω ($\partial\Omega$)
Γ_-	inflow boundary of Γ
Γ_+	outflow boundary of Γ
Γ_D	Dirichlet boundary of Γ
Γ_N	Neumann boundary of Γ
∂K	boundary of K
∇	gradient operator
$\nabla\cdot$	divergence operator (div)
Δ	Laplacian operator
Δ^2	biharmonic operator ($\Delta\Delta$)
Δt	time step size
$\dfrac{\partial}{\partial x_i}$	partial derivative with respect to x_i
$\dfrac{\partial}{\partial t}$	partial derivative with respect to t (time)
$\dfrac{\partial}{\partial \boldsymbol{\nu}}$	normal derivative
$\dfrac{\partial}{\partial \mathbf{t}}$	tangential derivative
$\dfrac{\partial}{\partial \boldsymbol{\tau}}$	directional derivative along characteristics
$\dfrac{D}{Dt}$	material derivative
D^α	partial derivative notation
D_w^α	weak derivative notation
D^r	multilinear form of rth-order derivative
$C^\infty(\Omega)$	space of functions infinitely differentiable
$\mathcal{D}(\Omega)$	subset of $C^\infty(\Omega)$ having compact support in Ω
$C_0^\infty(\Omega)$	same as $\mathcal{D}(\Omega)$
$\operatorname{diam}(K)$	diameter of K
$L^1_{loc}(\Omega)$	integrable functions on any compact set inside Ω
$L^q(\Omega)$	Lebesgue space
$W^{r,q}(\Omega)$	Sobolev spaces
$W_0^{r,q}(\Omega)$	completion of $\mathcal{D}(\Omega)$ with respect to $\|\cdot\|_{W^{r,q}(\Omega)}$
$\|\cdot\|$	norm
$\|\cdot\|_h$	norm on a nonconforming space
$\|\cdot\|_{L^q(\Omega)}$	norm of $L^q(\Omega)$

$\|\cdot\|_{W^{r,q}(\Omega)}$	norm of $W^{r,q}(\Omega)$
$\|\cdot\|_{W^{r,q}(\Omega)}$	seminorm of $W^{r,q}(\Omega)$
$H^r(\Omega)$	same as $W^{r,2}(\Omega)$
$H_0^r(\Omega)$	same as $W_0^{r,2}(\Omega)$
$H^l(K_h)$	piecewise smooth space
$\mathbf{H}(\mathrm{div}, \Omega)$	divergence space
$\boldsymbol{\beta}$	convection or advection coefficient
α	multi-index (a d-tuple): $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$
β_1	measure of smallest angle over $K \in K_h$
β_2	quasi-uniform triangulation constant
π_h	interpolation operator
π_K	restriction of π_h on element K
Π_h	projection operator
$\delta(\mathbf{x} - \mathbf{x}^{(l)})$	Dirac delta function at $\mathbf{x}^{(l)}$
$\boldsymbol{\kappa}$	reservoir permeability
$\kappa_{r\alpha}$	relative permeability of α-phase
μ_α	viscosity of α-phase
ρ	density
ρ_α	density of α-phase
ρ_K	diameter of largest circle inscribed in K
$\boldsymbol{\nu}$	outward unit normal
φ, $\boldsymbol{\varphi}$	interstitial velocity
φ_i	basis function of V_h
$\boldsymbol{\varphi}_i$	basis function of $\mathbf{V}_h$
λ_d	Lagrange multipliers
λ_i	barycentric coordinates ($i = 1, 2, 3$)
λ_α	phase mobility
$\boldsymbol{\lambda}$	degrees of freedom of λ_h
$B_{r+1}(K)$	same as $\lambda_1\lambda_2\lambda_3 P_{r-2}(K)$
$\boldsymbol{\epsilon}$	strain tensor
σ	Poisson's ratio
$\boldsymbol{\sigma}$	stress tensor
τ, $\boldsymbol{\tau}$	characteristic direction
$\tau_{\mathbf{p}}$	momentum relaxation time
τ_w	energy relaxation time
$\hbar$	quantum expansion parameter
$[\![\cdot]\!]$	jump operator notation
$\{\!\{\cdot\}\!\}$	averaging operator notation
$\det(\cdot)$	determinant of a matrix

References

R. A. Adams (1975), Sobolev Spaces, Academic Press, New York.

A. Adini and R. Clough (1961), Analysis of plate bending by the finite element method, NSF Report G. 7337, University of California, Berkeley, CA.

S. Adjerid, J. E. Flaherty, and Y. J. Wang (1993), A posteriori error estimation with finite element methods of lines for one-dimensional parabolic systems, *Numer. Math.* **65**, 1–21.

M. Ainsworth and J. T. Oden (2000), A-posteriori Error Analysis in Finite Element Analysis, Wiley Inter-Science, New York.

M. G. Ancona and G. J. Iafrate (1989), Quantum correction to the equation of state of an electron gas in a semiconductor, *Phys. Rev. B* **39**, 9536–9540.

S. N. Antontsev (1972), On the solvability of boundary value problems for degenerate two-phase porous flow equations, *Dinamika Splošnoĭ Sredy Vyp.* **10**, 28–53, in Russian.

T. Arbogast and Z. Chen (1995), On the implementation of mixed methods as nonconforming methods for second order elliptic problems, *Math. Comp.* **64**, 943–972.

T. Arbogast and M. F. Wheeler (1995), A characteristics-mixed finite element for advection-dominated transport problems, *SIAM J. Numer. Anal.* **32**, 404–424.

D. N. Arnold (1982), An interior penalty finite element method with discontinuous elements, *SIAM J. Numer. Anal.* **19**, 742–760.

D. N. Arnold and F. Brezzi (1985), Mixed and nonconforming finite element methods: implementation, postprocessing and error estimates, *RAIRO Modèl. Math. Anal. Numér.* **19**, 7–32.

D. N. Arnold, F. Brezzi, and J. Douglas, Jr. (1984A), PEERS: A new mixed finite element for plane elasticity, *Japan J. Appl. Math.* **1**, 347–367.

D. N. Arnold, F. Brezzi, and M. Fortin (1984B), A stable finite element for the Stokes equations, *Calcolo* **21**, 337–344.

D. N. Arnold, L. R. Scott, and M. Vogelius (1988), Regular inversion of the divergence operator with Dirichlet boundary conditions on a polygon, *Ann. Scuola Norm. Sup. Pisa Cl. Sci.-Serie IV* **XV**, 169–192.

K. Arrow, L. Hurwicz, and H. Uzawa (1958), Studies in Nonlinear Programming, Stanford University Press, Stanford, CA.

J. P. Aubin (1967), Behavior of the error of the approximate solutions of boundary value problems for linear elliptic operators by Galerkin's and finite difference methods, *Ann. Scuola Norm. Sup. Pisa* **21**, 599–637.

O. Axelsson (1994), Iterative Solution Methods, Cambridge University Press, Cambridge.

K. Aziz and A. Settari (1979), Petroleum Reservoir Simulation, Applied Science Publishers Ltd, London.

I. Babuška and M. R. Dorr (1981), Error estimates for the combined h and p versions of the finite element method, *Numer. Math.* **37**, 257–277.

I. Babuška, J. Osborn, and J. Pitkäranta (1980), Analysis of mixed methods using mesh dependent norms, *Math. Comp.* **35**, 1039–1062.

I. Babuška and W. C. Rheinboldt (1978A), Error estimates for adaptive finite element computations, *SIAM J. Numer. Anal.* **15**, 736–754.

I. Babuška and W. C. Rheinboldt (1978B), A-posteriori error estimates for the finite element method, *Int. J. Num. Meth. Eng.* **12**, 1597–1615.

I. Babuška, A. Miller, and M. Vogelius (1983), Adaptive methods and error estimation for elliptic problems of structural mechanics, in Adaptive Computational Methods for Partial Differential Equations, I. Babuška, et al., eds., SIAM, PA, 35–56.

W. Bangerth and R. Rannacher (2003), Adaptive Finite Element Methods for Differential Equations, Birkhäuser, Basel.

G. A. Baker (1977), Finite element methods for elliptic equations using nonconforming elements, *Math. Comp.* **31**, 45–59.

R. E. Bank (1990), PLTMG: A Software Package for Solving Elliptic Partial Differential Equations, User's Guide 6.0, SIAM, PA.

R. E. Bank, A. H. Sherman, and A. Weiser (1983), Refinement algorithms and data structures for regular local mesh refinement, in Scientific Computing, R. Stepleman, et al., eds., North-Holland, Amsterdam, New York, Oxford, 3–17.

R. E. Bank and K. Smith (1993), A posteriori error estimates based on hierarchical bases, *SIAM J. Numer. Anal.* **30**, 921–935.

R. E. Bank and A. Weiser (1985), Some a posteriori error estimators for elliptic partial differential equations, *Math. Comp.* **44**, 283–301.

J. W. Barrett and K. W. Morton (1984), Approximate symmetrization and Petrov-Galerkin methods for diffusion-convection problems, *Comp. Mech. Appl. Mech. Engrg.* **45**, 97–122.

G. Bazeley, Y. Cheung, B. Irons, and O. Zienkiewicz (1965), Triangular elements in bending conforming and nonconforming solutions, Proceedings of the Conference on Matrix Methods in Structural Mechanics, Wright Patterson A.F.B., Ohio.

J. Bear (1972), Dynamics of Fluids in Porous Media, Dover, New York.

M. J. Berger and J. Oliger (1984), Adaptive mesh refinement for hyperbolic partial differential equations, *J. Comp. Phys.* **53**, 484–512.

C. Bernardi, B. Métivet, and R. Verfürth (1993), Analyse numérique d'indicateurs d'erreur, Report 93025, Université Pierre et Marie Curie, Paris VI.

M. Bieterman and I. Babuška (1982), The finite element method for parabolic equations, a posteriori error estimation, *Numer. Math.* **40**, 339–371.

F. J. Blatt (1968), Physics of Electric Conduction in Solids, McGraw Hill, New York.

K. Blotekjaer (1970), Transport equations for electrons in two-valley semiconductor, *IEEE Trans. Electron Devices ED* **17**, 38–47.

D. Braess (1997), Finite Elements, Theory, Fast Solvers, and Applications in Solid Mechanics, Cambridge University Press, Cambridge.

J. H. Bramble (1966), A second-order finite difference analog of the first biharmonic boundary value problem, *Numer. Math.* **4**, 236–249.

J. H. Bramble (1993), Multigrid Methods, Pitman Research Notes in Math., vol. 294, Longman, London.

J. H. Bramble and S. R. Hilbert (1970), Estimation of linear functionals on Sobolev spaces with application to Fourier transforms and spline interpolation, *SIAM J. Numer. Anal.* **7**, 113–124.

J. H. Bramble, J. E. Pasciak, and A. Vassilev (1997), Analysis of the inexact Uzawa algorithm for saddle point problems, *SIAM J. Numer. Anal.* **34**, 1072–1092.

S. C. Brenner and L. R. Scott (1994), The Mathematical Theory of Finite Element Methods, Springer, New York Berlin Heidelberg.

F. Brezzi, J. Douglas, Jr., R. Durán, and M. Fortin (1987A), Mixed finite elements for second order elliptic problems in three variables, *Numer. Math.* **51**, 237–250.

F. Brezzi, J. Douglas, Jr., M. Fortin, and L. D. Marini (1987B), Efficient rectangular mixed finite elements in two and three space variables, *RAIRO Modèl. Math. Anal. Numér* **21**, 581–604.

F. Brezzi, J. Douglas, Jr., and L. D. Marini (1985), Two families of mixed finite elements for second order elliptic problems, *Numer. Math.* **47**, 217–235.

F. Brezzi and M. Fortin (1991), Mixed and Hybrid Finite Element Methods, Springer, New York Berlin Heidelberg.

A. Brooks and T. J. Hughes (1982), Streamline upwind Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations, *Comp. Mech. Appl. Mech. Engrg.* **32**, 199–259.

D. C. Brown (1982), Alternating-direction iterative schemes for mixed finite element methods for second order elliptic problems, Dissertation, University of Chicago, Illinois.

P. Castillo, B. Cockburn, D. Schotzau, and C. Schwab (2002), Optimal a priori error estimates for the *hp*-version of the local discontinuous Galerkin methods for convection-diffusion problems, *Math. Comp.* **71**, 455-478.

M. A. Celia, T. F. Russell, I. Herrera, and R. E. Ewing (1990), An Eulerian Lagrangian localized adjoint method for the advection-diffusion equation, *Adv. Water Resour.* **13**, 187–206.

G. Chavent and J. Jaffré (1978), Mathematical Models and Finite Elements for Reservoir Simulation, North-Holland, Amsterdam.

H. Chen and Z. Chen (2003), Stability and convergence of mixed discontinuous finite element methods for second-order differential problems, *J.f Numer. Math.* **11**, 253–287.

H. Chen, Z. Chen, and B. Li (2003A), Numerical study of the *hp* version of mixed discontinuous finite element methods for reaction-diffusion problems: the 1D case, *Numer. Meth. Part. Differ. Equ.* **19**, 525–553.

H. Chen, Z. Chen, and B. Li (2003B), The *hp* version of mixed discontinuous finite element methods for advection-diffusion problems, *Int. J. Math Math. Sci.* **53**, 3385–3411.

Z. Chen (1989), On the existence, uniqueness and convergence of nonlinear mixed finite element methods, *Mat. Aplic. Comp.* **8**, 241–258.

Z. Chen (1993A), Analysis of mixed methods using conforming and nonconfirming finite element methods, *RAIRO Model. Math. Anal. Numer.* **27**, 9–34.

Z. Chen (1993B), Projection finite element methods for semiconductor device equations, *Comput. Math. Appl.* **25**, 81–88.

Z. Chen (1996), Equivalence between and multigrid algorithms for nonconforming and mixed methods for second order elliptic problems, *East-West J. Numer. Math.* **4**, 1–33.

Z. Chen (1997), Analysis of expanded mixed methods for fourth order elliptic problems, *Numer. Methods PDEs* **13**, 483–503.

Z. Chen (2000), Formulations and numerical methods for the black-oil model in porous media, *SIAM J. Numer. Anal.* **38**, 489–514.

Z. Chen (2001A), On the relationship of various discontinuous finite element methods for second-order elliptic equations, *East-West J. Numer. Math.* **9**, 99–122.

Z. Chen (2001B), Degenerate two-phase incompressible flow I: Existence, uniqueness and regularity of a weak solution, *J. Diff. Equ.* **171**, 203–232.

Z. Chen (2002A), Degenerate two-phase incompressible flow II: Regularity, stability and stabilization, *J. Diff. Equ.* **186**, 345–376.

Z. Chen (2002B), Characteristic mixed discontinuous finite element methods for advection-dominated diffusion problems, *Comp. Meth. Appl. Mech. Eng.* **191**, 2509–2538.

Z. Chen (2002C), Relationships among characteristic finite element methods for advection-diffusion problems, *J. Korean SIAM* **6**, 1–15.

Z. Chen, B. Cockburn, C. Gardner, and J. W. Jerome (1995A), Quantum hydrodynamic simulation of hysteresis in the resonant tunneling diode, *J. Comp. Phys.* **117**, 274–280.

Z. Chen, B. Cockburn, J. W. Jerome, and C. W. Shu (1995B), Mixed-RKDG finite element methods for the 2-D hydrodynamic model for semiconductor device simulation, *VLSI Des.* **3**, 145–158.

Z. Chen, Y. Cui, and Q. Jiang (2004A), Locking-free nonconforming finite elements for planar linear elasticity, *Dynamical Systems and Differential Equations*, to appear.

Z. Chen and J. Douglas, Jr. (1989), Prismatic mixed finite elements for second order elliptic problems, *Calcolo* **26**, 135–148.

Z. Chen and J. Douglas, Jr. (1991), Approximation of coefficients in hybrid and mixed methods for nonliner parabolic problems, *Mat. Aplic. Comp.* **10**, 137–160.

Z. Chen, M. Espedal, and R. E. Ewing (1995), Continuous-time finite element analysis of multiphase flow in groundwater hydrology, *Appl. Math.* **40**, 203–226.

Z. Chen and R. E. Ewing (1997A), Comparison of various formulations of three-phase flow in porous media, *J. Comp. Phys.* **132**, 362–373.

Z. Chen and R. E. Ewing (1997B), Fully-discrete finite element analysis of multiphase flow in groundwater hydrology, *SIAM J. Numer. Anal.* **34**, 2228–2253.

Z. Chen and R. E. Ewing (2001), Degenerate two-phase incompressible flow III: Sharp error estimates, *Numer. Math.* **90**, 215–240.

Z. Chen and R. E. Ewing (2003), Degenerate two-phase incompressible flow IV: Local refinement and domain decomposition, *J. Sci. Comput.* **18**, 329–360.

Z. Chen, R. E. Ewing, Q. Jiang, and A. M. Spagnuolo (2002), Degenerate two-phase incompressible flow V: Characteristic finite element methods, *J. Numer. Math.* **10**, 87–107.

Z. Chen, R. E. Ewing, Q. Jiang, and A. M. Spagnuolo (2003C), Error analysis for characteristics-based methods for degenerate parabolic problems, *SIAM J. Numer. Anal.* **40**, 1491–1515.

Z. Chen, R. E. Ewing and R. Lazarov (1996), Domain decomposition algorithms for mixed methods for second order elliptic problems, *Math. Comp.* **65**, 467–490.

Z. Chen and G. Huan (2003), Numerical experiments with various formulations for two phase flow in petroleum reservoirs, *Transport in Porous Media* **51**, 89–102.

Z. Chen, G. Huan, and Y. Ma (2004B), Computational Methods for Multiphase Flows in Porous Media, in progress.

Z. Chen and P. Oswald (1998), Multigrid and multilevel methods for nonconforming rotated Q1 elements, *Math. Comp.* **67**, 667–693.

Z. Chen, G. Qin, and R. E. Ewing (2000), Analysis of a compositional model for fluid flow in porous media, *SIAM J. Appl. Math.* **60**, 747–777.

I. Christie, D. F. Griffiths, and A. R. Mitchell (1976), Finite element methods for second order differential equations with significant first derivatives, *Int. J. Num. Eng.* **10**, 1389–1396.

P. G. Ciarlet (1978), The Finite Element Method for Elliptic Problems, North-Holland, Amsterdam.

P. G. Ciarlet (1988), Mathematical Elasticity, vol. I: Three-Dimensional Elasticity, North-Holland, Amsterdam.

P. G. Ciarlet and P.-A. Raviart (1972), The combined effect of curved boundaries and numerical integration in isoparametric finite element methods, in the Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, A. K. Aziz, ed., Academic Press, New York, pp. 409–474.

Ph. Clément (1975), Approximation by finite element functions using local regularization, *RAIRO Anal. Numér.* **2**, 77–84.

B. Cockburn, G. E. Karniadakis, and C. W. Shu (2000), Discontinuous Galerkin methods, Theory, Computation and Application, Lecture Notes in Computational Science and Engineering, Vol. 11, Springer, Berlin Heidelberg New York.

B. Cockburn and C.-W. Shu (1998), The local discontinuous Galerkin method for time-dependent convection-diffusion systems, *SIAM J. Numer. Anal.* **35**, 2440–2463.

B. Cockburn, S. Hou, and C. W. Shu (1990), TVB Runge-Kutta local projection discontinuous Galerkin finite element method for scalar conservation laws IV: the multidimensional case, *Math. Comp.* **54**, 545–581.

J. B. Conway (1985), A Course in Functional Analysis, Springer, New York Berlin Heidelberg.

M. Crouzeix and P. Raviart (1973), Conforming and nonconforming finite element methods for solving the stationary Stokes equations, *RAIRO* **3**, 33–75.

R. Courant (1943), Variational methods for the solution of problems of equilibrium and vibrations, *Bull. Amer. Math. Soc.* **49**, 1–23.

H. K. Dahle, R. E. Ewing, and T. F. Russell (1995), Eulerian-Lagrangian localized adjoint methods for a nonlinear advection-diffusion equation, *Comput. Meth. Appl. Mech. Eng.* **122**, 223–250.

M. Dauge (1988), Elliptic Boundary Value Problems on Corner Domains, Lecture Notes in Math., vol. 1341, Springer, Berlin Heidelberg New York.

C. N. Dawson, T. F. Russell, and M. F. Wheeler (1989), Some improved error estimates for the modified method of characteristics, *SIAM J. Numer. Anal.* **26**, 1487–1512.

L. M. Delves and C. A. Hall (1979), An implicit matching principle for global element calculations, *J. Inst. Math. Appl.* **23**, 223–234.

P. Deuflhard, P. Leinen, and H. Yserentant (1989), Concepts of an adaptive hierarchical finite element code, *IMPACT Comput. Sci. Eng.* **1**, 3–35.

J. C. Diaz, R. E. Ewing, R. W. Jones, A. E. McDonald, I. M. Uhler, and D. U. von Rosenberg (1984), Self-adaptive local grid-refinement for time-dependent, two-dimensional simulation, Finite Elements in Fluids, vol. VI, Wiley, New York, 479–484.

J. Douglas, Jr. (1961), A survey of numerical methods for parabolic differential equations, in Advances in Computers, F. L. Alt, ed., vol. 2, Academic Press, New York, 1–54.

J. Douglas, Jr. (1977), H^1-Galerkin methods for a nonlinear Dirichlet problem, in Proceedings of the Conference "Mathematical Aspects of the Finite Element Methods", Lecture Notes in Math, vol. 606, Springer, Berlin Heidelberg New York, pp. 64–86.

J. Douglas, Jr. and T. Dupont (1976), Interior penalty procedures for elliptic and parabolic Galerkin methods, Lecture Notes in Physics, vol. 58, Springer, Berlin Heidelberg New York, pp. 207–216.

J. Douglas, Jr., R. Durán, and P. Pietra (1987), Formulation of alternating-direction iterative methods for mixed methods in three space, in the Proceedings of the Simposium Internacional de Analisis Numérico, E. Ortiz, ed., Madrid, pp. 21–30.

J. Douglas, Jr., R. E. Ewing, and M. Wheeler (1983), The approximation of the pressure by a mixed method in the simulation of miscible displacement, *RAIRO Anal. Numér.* **17**, 17–33.

J. Douglas, Jr., F. Furtado, and F. Pereira (1997), On the numerical simulation of water flooding of heterogeneous petroleum reservoirs, *Computational Geosciences* **1**, 155–190.

J. Douglas, Jr., I. Gamba, and M. C. J. Squeff (1986), Simulation of the transient behavior of a one-dimensional semiconductor device, *Mat. Aplic. Comp.* **5**, 103–122.

J. Douglas, Jr., D. W. Peaceman, and H. H. Rachford, Jr. (1959), A method for calculating multi-dimensional immiscible displacement, *Trans. SPE AIME* **216**, 297–306.

J. Douglas, Jr., F. Pereira, and L. M. Yeh (2000), A locally conservative Eulerian-Lagrangian numerical method and its application to nonlinear transport in porous media, *Comput. Geosci.* **4**, 1–40.

J. Douglas, Jr. and P. Pietra (1985), A description of some alternating-direction techniques for mixed finite element methods, in Mathematical and Computational Methods in Seismic Exploration and Reservoir Modeling, SIAM, Philadelphia, PA, pp. 37–53.

J. Douglas, Jr. and J. Roberts (1985), Global estimates for mixed methods for second order elliptic problems, *Math. Comp.* **45**, 39–52.

J. Douglas, Jr. and T. F. Russell (1982), Numerical methods for convection dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures, *SIAM J. Numer. Anal.* **19**, 871–885.

T. Dupont and R. Scott (1980), Polynomial approximation of functions in Sobolev spaces, *Math. Comp.* **34**, 441–463.

V. Eijkhout and P. Vassilevski (1991), The role of the strengthened Cauchy-Bujanowsky-Schwarz inequality in multilevel methods, *SIAM Rev.* **33**, 405–419.

H. Elman and G. Golub (1994), Inexact and preconditioned Uzawa algorithms for saddle point problems, *SIAM J. Numer. Anal.* **31**, 1645–1661.

K. Eriksson and C. Johnson (1991), Adaptive finite element methods for parabolic problems I: A linear model problem, *SIAM J. Numer. Anal.* **28**, 43–77.

K. Eriksson and C. Johnson (1995), Adaptive finite element methods for parabolic problems IV: Nonlinear problems, *SIAM J. Numer. Anal.* **32**, 1729–1749.

N. S. Espedal and R. E. Ewing (1987), Characteristic Petrov-Galerkin subdomain methods for two phase immiscible flow, *Comput. Methods Appl. Mech. Eng.* **64**, 113–135.

R. E. Ewing (1986), Efficient adaptive procedures for fluid flow applications, *Comp. Meth. Appl. Mech. Eng.* **55**, 89–103.

R. E. Ewing, R. Lazarov, P. Lu, and P. Vassilevski (1990), Preconditioning indefinite systems arising from the mixed finite element discretization of second-order elliptic systems, in Preconditioned Conjugate Gradient Methods, O. Axelsson and L. Kolotilina, eds., Lecture Notes in Math., vol. 1457, Springer, Berlin Heidelberg New York, pp. 28–43.

R. E. Ewing, T. F. Russell, and M. F. Wheeler (1984), Convergence analysis of an approximation of miscible displacement in porous media by mixed finite elements and a modified method of characteristics, *Comp. Meth. Appl. Mech. Eng.* **47**, 73–92.

R. E. Ewing and J. Wang (1992), Analysis of mixed finite element methods on locally refined grids, *Numer. Math.* **63**, 183–194.

D. Ferry and R. Grondin (1992), Physics of Sub-Micron Devices, New York, Plenum.

R. Feynman (1960), There's plenty of room at the bottom, *Eng. Sci.* **Feb.**, 22–36.

M. Fortin (1977), An analysis of the convergence of mixed finite element methods, *RAIRO Anal. Numér.* **11**, 341–354.

M. Fortin and Soulie (1983), A nonconforming piecewise quadratic finite element on triangles, *Int. J. Numer. Methods Eng.* **19**, 505–520.

B. Fraeijs de Veubeke (1965), Displacement and equilibrium models in the finite element method, Stress Analysis, O. C. Zienkiewicz and G. Holister, eds., Wiley, New York.

B. Fraeijs de Veubeke (1974), Variational principles and the patch test, *Int. J. Numer. Methods Eng.* **8**, 783–801.

W. R. Frensley (1985), Simulation of resonant-tunneling heterostructure devices, *J. Vacuum Sci. Technol.* **B3**, 1261–1266.

A. O. Garder, D. W. Peaceman, and A. L. Pozzi (1964), Numerical calculations of multidimensional miscible displacement by the method of characteristics, *Soc. Pet. Eng. J.* **4**, 26–36.

V. Girault and P.-A. Raviart (1981), Finite Element Approximation of the Navier-Stokes Equations, Springer, Berlin Heidelberg New York.

R. Glowinski (2003), Handbook of Numerical Analysis: Numerical Methods for Fluids, Elsevier.

G. H. Golub and C. F. Van Loan (1996), Matrix Computations, Johns Hopkins University Press, Baltimore and London.

H. Grubin and J. Kreskovsky (1989), Quantum moment balance equations and resonant tunneling structures, *Solid State Electron.* **32**, 1071–1075.

W. Hackbusch (1985), Multigrid Methods and Applications, Springer, Berlin Heidelberg New York.

K. Hellan (1967), Analysis of elastic plates in flexure by a simplified finite element method, *Acta Polytechnica Scandinavia*, Civil Engineering Series, Trondheim **46**.

P. Henrici (1962), Discrete Variable Methods in Ordinary Differential Equations, Wiley, New York.

L. R. Herrmann (1967), Finite element bending analysis for plates, *J. Eng. Mech. Div. ASCE* **93**, 13–26.

R. H. Hoppe and B. Wohlmuth (1997), Adaptive multilevel techniques for mixed finite element discretizations of elliptic boundary value problems, *SIAM J. Numer. Anal.* **34**, 1658–1681.

T. J. R. Hughes, G. Engel, L. Mazzei, and M. G. Larson (2000), A comparison of discontinuous and continuous Galerkin methods based on error estimates, conservation, robustness and efficiency, in Discontinuous Galerkin Methods, Theory, Computation and Applications, B. Cockburn, et al., eds., Lecture Notes in Computational Science and Engineering, vol. 11, Springer, Berlin Heidelberg New York, pp. 135–146.

J. W. Jerome (1985), Consistency of semiconductor modeling: An existence/stability analysis for the stationary van Roosbroeck system, *SIAM J. Appl. Math.* **54**, 565–590.

C. Johnson (1994), Numerical Solutions of Partial Differential Equations by the Finite Element Method, Cambridge University Press, Cambridge.

C. Johnson and J. Pitkaranta (1986), An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation, *Math. Comp.* **46**, 1–26.

C. Johnson and V. Thomée (1981), Error estimates for some mixed finite element methods for parabolic type problems, *RAIRO Anal. Numer.* **15**, 41–78.

W. Kaplan (1991), Advanced Calculus, 4th Ed., Addison Wesley, Publishing Company, Inc.

N. C. Kluksdahl, A. M. Kriman, D. K. Ferry, and C. Ringhofer (1989), Self-consistent study of the resonant tunneling diode, *Phys. Rev. B* **39**, 7720–7735.

V. A. Kondratev (1967), Boundary value problems for elliptic equations with conical or angular point, *Trans. Moscow Math. Soc.* **10**, 227–313.

P. Lascaux and P. LeSaint (1975), Some nonconforming finite elements for the plate bending problem, *RAIRO Anal. Numer.* **9**, 9–53.

P. LeSaint and P. A. Raviart (1974), On a finite element method for solving the neutron transport equation, in *Mathematical Aspects of Finite Elements in Partial Differential Equations*, C. de Boor, ed., Academic Press, 89–145.

K. Li, A. Huang, and Q. Huang (1984), The Finite Element Method and its Application, Xi'an Jiaotong University Press, Xi'an, China, in Chinese.

J. L. Lions and E. Magenes (1972), Non-homogeneous Boundary Value Problems and Applications, Springer, New York Berlin Heidelberg.

P. A. Markowich (1986), The Stationary Semiconductor Equations, Springer, New York Berlin Heidelberg.

P. A. Markowich, C. A. Ringhofer, and C. Schmeiser (1990), Semiconductor Equations, Springer, New York Berlin Heidelberg.

K. Miller and R. N. Miller (1981), Moving finite elements, *SIAM J. Numer. Anal.* **18**, 79–95.

F. Milner (1985), Mixed finite element methods for quasilinear second order elliptic problems, *Math. Comp.* **44**, 303–320.

P. K. Moore (2001), Interpolation error-based a posteriori error estimation for two-point boundary value problems and parabolic equations in one space dimension, *Numer. Math.* **90**, 149–177.

L. Morley (1968), The triangular equilibrium problem in the solution of plate bending problems, *Aero. Quart.* **19**, 149–169.

J. C. Néd'elec (1980), Mixed finite elements in $\mathbb{R}^3$, *Numer. Math.* **35**, 315–341.

J. C. Néd'elec (1986), A new family of mixed finite elements in $\mathbb{R}^3$ *Numer. Math.* **50**, 57–81.

S. P. Neuman (1981), An Eulerian-Lagrangian numerical scheme for the dispersion-convection equation using conjugate-time grids, *J. Comp. Phys.* **41**, 270–294.

J. A. Nitsche (1968), Ein kriterium für die quasi-optimalitat des Ritzchen Verfahrens, *Numer. Math.* **11**, 346–348.

J. A. Nitsche (1971), Über ein variationsprinzip zur lösung von Dirichlet problem bei verwendung von teilräumen, die keinen randbedingungen unterworfen sind, *Abh. Math. Sem. Univ. Hamburg* **36**, 9–15.

J. Nougier, J. Vaissiere, D. Gasquet, J. Zimmermann, and E. Constant (1981), Determination of the transient regime in semiconductor devices using relaxation time approximations, *J. Appl. Phys.* **52**, 825–832.

J. T. Oden, I. Babuška, and C. E. Baumann (1998), A discontinuous hp finite element method for diffusion problems, *J. Comput. Phys.* **146**, 491–519.

J. T. Oden and L. Demkowicz (1988), Advances in adaptive improvements: A survey of adaptive finite element methods in computational mechanics, State-of-the-Art Surveys in Computational Mechanics, A. K. Noor and J. T. Oden, eds., A.S.M.E. Publications, New York.

J. T. Oden, L. Demkowicz, W. Rachowicz, and T. A. Westermann (1989), Toward a universal $h-p$ adaptive finite element strategy, Part 2. A posteriori error estimation, *Comp. Meth. Appl. Mech. Engrg.* **77**, 113–180.

E. R. Oliveira (1971), Optimization of finite element solutions, Proceedings of the Third Conference on Matrix Methods in Structural Mechanics, Wright-Patterson Air Force Base, Ohio, October.

J. M. Ortega and W. C. Rheinboldt (1970), Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, New York.

A. M. Ostrowski (1973), Solution of Equations in Euclidean and Banach Spaces, 3rd Edition, Academic Press, New York.

C. Paige and M. Saunders (1975), Solution of sparse indefinite systems of linear equations, *SIAM Numer. Anal.* **12**, 617–629.

D. W. Peaceman (1977A), Interpretation of well-block pressures in numerical reservoir simulation, SPE 6893, 52nd Annual Fall Technical Conference and Exhibition, Denver.

D. W. Peaceman (1977B), Fundamentals of Numerical Reservoir Simulation, Elsevier, New York.

D. W. Peaceman (1991), Presentation of a horizontal well in numerical reservoir simulation, SPE 21217, presented at 11th SPE Symposium on Reservoir Simulation in Ananheim, California, Feb. 17–20.

C. Philippidis, D. Bohm, and R. D. Kaye (1982), The Aharonov-Bohm effect and the quantum potential, *Il Nuovo Cimento* **71B**, 75–88.

O. Pironneau (1982), On the transport-diffusion algorithm and its application to the Navier-Stokes equations, *Numer. Math.* **38**, 309–332.

A. Quarteroni and A. Valli (1997), Numerical Approximation of Partial Differential Equations, Lecture Notes in Comp. Math., Vol. 23, Springer, Berlin Heidelberg New York.

R. Rannacher and S. Turek (1992), Simple nonconforming quadrilateral Stokes element, *Numer. Meth. Part. Diff. Equ.* **8**, 97–111.

R. Raviart, and J.-M. Thomas (1977), A mixed finite element method for second order elliptic problems, Lecture Notes in Mathematics, vol. 606, Springer, Berlin Heidelberg New York, pp. 292–315.

W. H. Reed and T. R. Hill (1973), Triangular mesh methods for the neutron transport equation, *Technical Report*, LA-UR-73-479, Los Alamos Scientific Laboratory.

W. C. Rheinboldt (1998), Methods for Solving Systems of Nonlinear Equations, 2nd Edition, Society for Industrial and Applied Mathematics, Philadelphia.

W. C. Rheinboldt and C. Mesztenyi (1980), On a data structure for adaptive finite element mesh refinement, *ACM Trans. Math. Softw.* **6**, 166–187.

M. C. Rivara (1984A), Algorithms for refining triangular grids suitable for adaptive and multigrid techniques, *Int. J. Num. Meth. Eng.* **20**, 745–756.

M. C. Rivara (1984B), Design and data structure of fully adaptive, multigrid, finite element software, *ACM Trans. Math. Softw.* **10**, 242–264.

B. Rivière, M. F. Wheeler, and V. Girault (1999), Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems. Part I. *Comput. Geosci.* **3**, 337–360.

J. E. Roberts and J.-M. Thomas (1989), Mixed and hybrid methods, *Handbook of Numerical Analysis*, P. G. Ciarlet and J. L. Lions, eds., vol. II, Finite Element Methods (Part 1), North-Holland, Amsterdam.

R. Rodriguez (1994), Some remarks on Zienkiewicz-Zhu estimator, *Int. J. Numer, Meth. PDE* **10**, 625–635.

W. Rudin (1987), Real and Complex Analysis, 3rd Ed., McGraw-Hill. New York.

T. F. Russell (1990), Eulerian-Lagrangian localized adjoint methods for advection-dominated problems, in Numerical Analysis, Pitman Res. Notes Math. Series, vol. 228, D. F. Griffiths and G. A. Watson, eds., Longman Scientific and Technical, Harlow, England, pp. 206–228.

T. F. Russell and R. V. Trujillo (1990), Eulerian-Lagrangian localized adjoint methods with variable coefficients in multiple dimensions, Gambolati, et al., eds., Comp. Meth. in Surface Hydrology, Springer, Berlin Heidelberg New York, pp. 357–363.

T. F. Russell and M. F. Wheeler (1983), Finite element and finite difference methods for continuous flows in porous media, the Mathematics of Reservoir Simulation, R. E. Ewing, ed., SIAM, Philadelphia, pp. 35–106.

T. Rusten and R. Winther (1992), A preconditioned iterative method for saddle-point problems, *SIAM J. Matrix Anal. Appl.* **13**, 887–904.

M. Sheffield (1970), A non-iterative technique for solving parabolic partial differential equation problems, SPE 2803, 2nd Symposium on Numerical Simulation of Reservoir Performance, Dallas, Texas.

J.W. Sheldon, B. Zondek, and W.T. Cardwell (1959), One-dimensional, incompressible, non-capillary, two-phase fluid flow in a porous medium, *Trans. SPE AIME* **216**, 290–296.

Z.-C. Shi (1987), The F-E-M-test for nonconforming finite elements, *Math. Comp.* **49**, 391–405.

B. Smith, P. Bjorstad, and W. Gropp (1996), Domain Decomposition, Parallel Multilevel Methods for Elliptic Partial Differential Equations, Cambridge University Press, Cambridge.

G. Strang and G. J. Fix (1973), An Analysis of the Finite Element Method, Prentice Hall, Englewood Cliffs, NJ.

B. A. Szabo (1986), Mesh design for the p-version of the finite element method, *Comp. Meth. Appl. Mech. Eng.* **55**, 86–104.

V. Thomée (1984), Galerkin Finite Element Methods for Parabolic Problems, Lecture Notes in Math., vol. 1054, Springer, Berlin Heidelberg New York.

W. V. van Roosbroeck (1950), Theory of flow of electrons and holes in germanium and other semiconductors, *Bell Syst. Techn. J.* **29**, 560–607.

R. Verfürth (1996), A Review of a Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques, Wiley-Teubner, Chichester-Stuttgart.

H. Wang (2000), An optimal-order error estimate for an ELLAM scheme for two-dimensional linear advection-diffusion equations, *SIAM J. Numer. Anal.* **37**, 1338–1368.

H. Wang, R. E. Ewing, and T. F. Russell (1995), Eulerian-Lagrangian localized adjoint methods for convection-diffusion equations and their convergence analysis, *IMA J. Numer. Anal.* **15**, 405–459.

J. Wang and T. Mathew (1994), Mixed finite element methods over quadrilaterals, In the Proceedings of the Third International Conference on Advances in Numerical Methods and Applications, I. T. Dimov, et al., eds., World Scientific, 203–214.

J. J. Westerink and D. Shea (1989), Consistent higher degree Petrov-Galerkin methods for the solution of the transient convection-diffusion equation, *Int. J. Num. Meth. Eng.* **13**, 839–941.

M. F. Wheeler (1973), A priori L_2 error estimates for Galerkin approximation to parabolic partial differential equations, *SIAM J. Numer. Anal.* **10**, 723–759.

M. F. Wheeler (1978), An elliptic collocation-finite element method with interior penalties, *SIAM J. Numer. Anal.* **15**, 152–161.

E. Wigner (1932), On the quantum correction for thermodynamic equilibrium, *Phys. Rev.* **40**, 749–759.

D. Yang (1992), A characteristic mixed method with dynamic finite element space for convection-dominated diffusion problems, *J. Comput. Appl. Math.* **43**, 343–353.

K. Yosida (1971), Functional Analysis, 3rd Edition, Springer, Berlin Heidelberg New York.

O. C. Zienkiewicz and J. Zhu (1987), A simple error estimator and adaptive procedure for practical engineering analysis, *Int. J. Num. Meth. Eng.* **24**, 337–357.

Index

Adaptive algorithm 274, 290–293
Adaptive finite element method 52, 289
Adini element 104
Adjoint linear operator 294
Advection problems 173
Algorithm efficiency 287
Algorithm reliability 275
Antisymmetric operator 311
Approximation error 108, 111
Approximation theory 62
A-priori error estimate 261
Argyris's triangle 35, 44, 45
Arrow-Hurwitz alternating direction algorithms 148
Asymptotically exact 288
Aubin-Nitsche technique 68, 108
Auger law 365
Averaging-based estimators 271, 281, 302

Babuška-Brezzi condition 130, 160
Backward Euler method 51, 55
Backward substitution 71, 72, 76
Banach space 20
Band matrix 74
Band width 18, 74, 85
Barycentric coordinates 37, 277
Basis functions 4, 120
BDDF spaces 137, 139
BDFM spaces 135, 139
BDM spaces 132, 134
Bell's triangle 45
Biharmonic problem 34, 100, 334
Bilinear form 10, 26
Black-oil model 337
Block-centered finite difference 154
Boltzmann equation 364
Bounded bilinear form 27
Bounded linear functional 25
Bramble-Hilbert lemma 110, 316, 318
Bubble pressure 337

C^1 element 44
Céa's lemma 29
Capillary pressure 339
Cauchy's inequality 7, 30, 54, 56, 82
Cauchy's sequence 20
Cauchy-Green strain tensor 306
CD spaces 141, 142
Cell-centered finite difference 154
Center of gravity 39
Central difference scheme 7
Chapeau functions 4
Characteristic 217
Characteristic finite element 215
Characteristic mixed method 216, 242
Characterization curves of displacement 347
Cholesky decomposition 57
Clément interpolation operator 297
Clamped elastic plate 34
Closed range theorem 156
Coercive bilinear form 27
Commuting diagram 163
Compact support 21
Compatibility condition 15, 127, 308, 310
Complete linear space 20
Compositional flow 338
Condition number 54, 76, 77, 146, 191, 221
Conditionally stable 56, 58
Cone condition 110
Conjugate gradient algorithm 70, 76
Conservation of energy 321

Conservation of mass 225, 233, 244, 321, 322
Conservation of momentum 321
Conservation relation 224
Consistency error 108
Consistent term 195
Continuous bilinear form 27
Continuous linear functional 25
Contraction ratio 307
convection-diffusion-reaction problem 32, 215
Convergence 8
Courant number 215, 231
Crank-Nicholson method 56
Criss-cross grid 288
Critical saturation 339
Crouzeix-Raviart element 90, 95
Current densities 365
Cyclic boundary condition 222

Dankwerts boundary condition 15, 128
Darcy's law 339
Data structures 267
Deformation 306
Degrees of freedom 33, 36
DG 173, 175, 208, 373
Diffusion problem 173, 183, 190, 193
Direct method 18
Dirichlet boundary condition 127, 200, 280
Discontinuous finite element method 173
Discontinuous Galerkin method 173
Discontinuous weak formulation 186
Discrete inf-sup condition 130
Displacement 305
Divergence form 321
Divergence free 175, 224, 327, 328
Divergence theorem 9
Drift-diffusion model 363–366, 368
Dual index 26
Dual norm 25
Dual space 25, 26
Duality 26
Duality argument 68, 108
Duality pairing 155

Edge bubble function 277
Efficiency index 287
Elastic bar 2
Elastic membrane 2
Elasticity theory 305
Electric potential 365, 383
Electron continuity equation 365
Element stiffness matrix 18
Element-oriented 19
ELLAM 226–228, 233, 234, 236, 239, 241, 245, 250, 258
Elliptic bilinear form 27
Energy norm 80
Enhanced recovery 338
Equidistribution 275
Equilibrium 305, 306
Equivalence relationship 152
Error estimate 1, 7, 8, 14, 16, 29, 31, 33–35, 47, 54, 55, 92, 94, 102, 104, 109, 122, 126–128, 150, 164, 178, 188–190, 193, 205, 209, 213, 224, 236, 259, 287, 288, 294, 298, 310, 313, 317, 325, 333, 361, 362
Essential boundary condition 15, 127
Essential supremum 20
Eulerian approach 215, 321
Eulerian-Lagrangian localized adjoint method 226
Eulerian-Lagrangian methods 216
Eulerian-Lagrangian mixed discontinuous method 216, 245
Explicit scheme 57
Explicit time approximation 61
Extrapolation 60, 345, 384

Family structure 266
Finite difference 1
Finite element 45
Finite element method 1, 2, 7
Finite element space 3, 35
First boundary condition 34
Five-point stencil scheme 13, 14
Fluid flow in porous media 337
Fluid mechanics 321
Flux boundary condition 229, 231, 232, 259
Forward elimination 71, 72
Forward Euler method 57
Forward tracking 250
Fourier's coefficients 51

Fourth-order problem 34, 87, 98
Fréchet differentiability 294
Fractional flow 340
Fractured reservoirs 337
Fraeijs de Veubeke's element 102
Fully discrete schemes 55
Functionals 25
Fundamental principle of minimum potential energy 3

Galerkin finite element method 4, 325
Galerkin variational form 3
Gaussian elimination 18, 19, 70
General domains 46
Global element method 186
Global matrix 19
Global pressure 342
Global refinement 263
Gradient operator 10
Green edge 264
Green's formula 9
Green's second formula 99
Gronwall's lemma 253, 360
Ground water modeling 337

H^1-conforming method 87
H^2-conforming method 87
Hölder's inequality 20
Hahn-Banach theorem 157
Hanging nodes 263
Harmonical average 154
Hat functions 4
Heat flux 364
Hermite finite element 46
Heterogeneous reservoir 337
Hierarchical basis estimators 271, 283
Hilbert space 26
Hole continuity equation 365
Homeomorphisms 294
Homogeneous deformation 307
H-scheme 262
Hu-Washizu mixed method 309
Hydrodynamic model 363, 366, 371
Hyperbolic problem 173, 233, 250
Hysteresis 339

IMPES 346
Implicit pressure-explicit saturation 346
Implicit scheme 55, 70
Implicit time approximation 60
Inclusion relations 23
Incomplete Cholesky factorization 81
Incompressible flow 322
Inertial effects 323
Inexact Uzawa algorithm 147
Inf-sup condition 130, 157, 160, 331
Inflow boundary 174, 217, 243
Initial transient 52
Injection wells 337
Inner product space 26
Interpolant 7, 62
Interpolant operator 62
Interpolation error 8, 62
Inverse inequality 77
Irregular local refinement 265
Irregularity index 265
Isomorphism 156
Isoparametric finite elements 41, 85
Isotropic material 307

Jumps 183, 246, 273

Kinematics 305

L^2-error estimate 68
L^2-projection 163
Ladyshenskaja-Babuška-Brezzi condition 130, 160
Lagrange finite element 46
Lagrange multipliers 150
Lagrangian approach 321
Lagrangian-Eulerian approach 321
Lamé constants 307
Lamé differential equation 308
Laminar flow 323
Laplacian operator 9
Lax-Friedrichs flux 376
Lax-Milgram lemma 27
Leaves of a tree 266
Lebesgue space 20
Linear elasticity 306
Linear functional 25
Linear Hooke's law 307
Linearization approach 59
Lipschitz continuous 58, 63
Lipschitz domain 63

Load vector 5
Local DG method 212
Local Dirichlet problem
  estimator 277, 279
Local Neumann problem
  estimator 280
Local problem-based estimators 271, 277
Local refinement 17, 266
Localization 180
Locking effects 310
Lower triangular matrix 71
LU-factorization 71

Mass matrix 54, 374
Material derivative 321
Material laws 306
Matrix blocks 337
Matrix norm 57
MESFET device 379
Mesh parameters 11, 31, 89
Method of characteristics 216
MINI element 332
Minimal residual algorithm 145, 147
Minimization problem 3, 10, 27, 28
Minkowski's inequality 21
Mixed boundary condition 14, 128
Mixed discontinuous method 195, 200, 203, 206
Mixed finite element method 117, 119, 127, 128, 143, 145, 158, 166, 167, 216, 242, 330, 338, 369, 377
Mixed finite element spaces 128
Mixed variational form 124, 128, 129
Mixed-hybrid algorithms 150
MMOC 218, 221, 222, 225, 248, 345, 369, 370
Mobility 340, 341
Modified method of
  characteristics 216, 218
  characteristics with adjusted
    advection 225
Modified mixed-hybrid
  algorithm 152
Modified Uzawa algorithm 146
Modulus of elasticity 307
Moments 364
Monotonicity 180
Morley element 35, 100
Multilinear form of derivative 64

Natural boundary condition 15
Navier-Stokes equation 322, 329
Navier-Stokes law 322
Negative norms 25
Nested sequence of spaces 265
Neumann boundary condition 16, 126
Newton law 322
Newton's method 60
Newton-Raphson's method 60
Newtonian fluid 322
No-flow boundary conditions 343
No-slip condition 323
Node-oriented 19
Nodes 11
Nonconforming finite element
  method 87, 90, 310, 313
Nonconforming spaces 87, 90, 92, 106
Nondivergence form 218, 222, 321
Nonlinear transient problem 58, 87, 105, 117, 248, 292
Nonsymmetric DG 188, 190
Nonsymmetric interior penalty DG 189, 213
Normal derivative 10
Normed linear space 20
Norms 23, 25, 65
Numerical flux 373, 376

Oil recovery 337
Oil recovery curve 348
Operator splitting method 216
Orthogonal complement 156
Orthonormal system 51
Outflow boundary 174, 242, 381

P_1-nonconforming element 92
Parabolic problem 50
Parallel triangulation 288
PEERS 311
Periodic boundary condition 222
Permeabilities 339
Petroleum reservoirs 337
Petrov-Galerkin method 215
Phase formulation 340
Pivots 72
Planar strain 311
Planar stress 319

Plate problem 100
Poincaré's inequality 23, 24
Poisson locking 310
Poisson's equation 16
Poisson's ratio 307
Positive definite matrix 6
Preconditioning technique 80
Pressure equation 341, 349
Primary recovery 337
Principle of virtual work 3
Prismatic elements 97, 98
Production wells 337
Programming 16
Projection operators 162
P-scheme 262
Pure displacement 307
Pure traction 307

Quadrature rules 49
Quantum corrections 363
Quantum hydrodynamic model 363
Quantum potential 378
Quasi-uniform triangulation 54, 332, 353

Reaction-diffusion-advection problem 215
Recombination-generation rate 365
Rectangular elements 92
Rectangular parallelepipeds 42
Reduced Argyris triangle 35, 45
Refinement rule 264
Regular local refinement 263
Regular triangulation 31, 68
Regularity on functions 30
Regularity on solution 68, 69, 92
Relaxation times 366
Residual estimators 271
Residual saturation 339
Reynolds number 323
Riesz representation theorem 27
Ritz finite element method 4
Ritz variational form 3
Robin boundary condition 15, 128
Root of a tree 266
Rotated Q_1 element 93, 96
R-scheme 263
RT spaces 130, 133
RTN spaces 136, 138, 140
Runge-Kutta method 372

Saddle point problem 119, 120, 146, 155, 245
Saturation assumption 287
Saturation equation 338, 341
Saturation relation 338
Scalar product 2, 7, 109
Scaling argument 112
Schwarz's inequality 20
Second boundary condition 15, 126
Secondary recovery 337
Semi-discrete scheme 53
Semiconductor modeling 363
Seminorms 23
Sink/source term 339
Slave nodes 263
Slope limiting 376
Sobolev norm 22
Sobolev spaces 23
Solenoidal 175, 224
Solid mechanics 305
Solution regularity 68, 69
Sparse matrix 6, 16
Square integrable functions 20
Stability 51, 55, 170
Stability condition 57, 62, 130
Stabilization parameter 179
Stabilized DG methods 178, 210
Stationary problem 2, 9, 14, 54, 118, 123, 126, 128, 270
Stiffness matrix 5, 18, 77, 82
Stokes equation 322, 323
Strain 306
Strang's second lemma 107
Stream function 132, 325
Stream-function vorticity formulation 322
Streamline 178
Streamline diffusion method 179
Strengthened Cauchy-Schwarz inequality 286
Stress tensor 322
Symmetric bilinear form 27
Symmetric DG method 186
Symmetric interior penalty DG 187
Symmetric term 195

Tertiary recovery 338

Test functions 3, 233, 235
Tetrahedra 42
Third boundary condition 15, 128
Total potential energy of the plate 100
Total velocity 341
Trace of a tensor 310
Trace theorem 110
Transformation formula 64
Transient problem 50, 289
Transport diffusion method 216
Tree structure 266
Triangle bubble functions 277
Triangle inequality 21
Triangular elements 35
Triangulation 11
Tridiagonal matrix 6, 70
Trilinear form 330
Turbulence 323
Turbulent flow 323
Two-phase flow 338
Two-point boundary value problem 2

Unconditionally stable 56
Unrefinements 266
Upper triangular matrix 71
Upwind finite difference 176, 177
Uzawa algorithm 145, 146
Uzawa alternating-direction algorithms 148

Variational problem 27
Virtual parabolic problem 148
Virtually optimal estimate 189
Volume locking 310

Water flooding 337
Weak derivatives 21
Weak form 3
Weighted formulation 342
Weighted pressure 342
Wiedemann-Franz law 366
Wilson nonconforming element 94, 96

Young modulus 307

Zienkiewicz element 103