PURE AND APPLIED MATHEMATICS

*A Series of Monographs and Textbooks*

# GENERALIZED DIFFERENCE METHODS FOR DIFFERENTIAL EQUATIONS

## Numerical Analysis of Finite Volume Methods

Ronghua Li
Zhongying Chen
Wei Wu

# GENERALIZED DIFFERENCE METHODS FOR DIFFERENTIAL EQUATIONS

## Numerical Analysis of Finite Volume Methods


**Ronghua Li**
*University of Jilin*
*Changchun, China*

**Zhongying Chen**
*Zhongshan University*
*Guangzhou, China*

**Wei Wu**
*Dalian University of Technology*
*Dalian, China*



MARCEL DEKKER, INC. NEW YORK • BASEL

# GENERALIZED DIFFERENCE METHODS FOR DIFFERENTIAL EQUATIONS

## MONOGRAPHS AND TEXTBOOKS IN PURE AND APPLIED MATHEMATICS

1. *K. Yano,* Integral Formulas in Riemannian Geometry (1970)
2. *S. Kobayashi,* Hyperbolic Manifolds and Holomorphic Mappings (1970)
3. *V. S. Vladimirov,* Equations of Mathematical Physics (A. Jeffrey, ed.; A. Littlewood, trans.) (1970)
4. *B. N. Pshenichnyi,* Necessary Conditions for an Extremum (L. Neustadt, translation ed.; K. Makowski, trans.) (1971)
5. *L. Narici et al.,* Functional Analysis and Valuation Theory (1971)
6. *S. S. Passman,* Infinite Group Rings (1971)
7. *L. Dornhoff,* Group Representation Theory. Part A: Ordinary Representation Theory. Part B: Modular Representation Theory (1971, 1972)
8. *W. Boothby and G. L. Weiss, eds.,* Symmetric Spaces (1972)
9. *Y. Matsushima,* Differentiable Manifolds (E. T. Kobayashi, trans.) (1972)
10. *L. E. Ward, Jr.,* Topology (1972)
11. *A. Babakhanian,* Cohomologicsl Methods in Group Theory (1972)
12. *R. Gilmer,* Multiplicative Ideal Theory (1972)
13. *J. Yeh,* Stochastic Processes and the Wiener Integral (1973)
14. *J. Barros-Neto,* Introduction to the Theory of Distributions (1973)
15. *R. Larsen,* Functional Analysis (1973)
16. *K. Yano and S. Ishihara,* Tangent and Cotangent Bundles (1973)
17. *C. Procesi,* Rings with Polynomial Identities (1973)
18. *R. Hermann,* Geometry, Physics, and Systems (1973)
19. *N. R. Wallach,* Harmonic Analysis on Homogeneous Spaces (1973)
20. *J. Dieudonné,* Introduction to the Theory of Formal Groups (1973)
21. *I. Vaisman,* Cohomology and Differential Forms (1973)
22. *B.-Y. Chen,* Geometry of Submanifolds (1973)
23. *M. Marcus,* Finite Dimensional Multilinear Algebra (in two parts) (1973, 1975)
24. *R. Larsen,* Banach Algebras (1973)
25. *R. O. Kujala and A. L. Vitter, eds.,* Value Distribution Theory: Part A; Part B: Deficit and Bezout Estimates by Wilhelm Stoll (1973)
26. *K. B. Stolarsky,* Algebraic Numbers and Diophantine Approximation (1974)
27. *A. R. Magid,* The Separsble Galois Theory of Commutative Rings (1974)
28. *B. R. McDonald,* Finite Rings with Identity (1974)
29. *J. Satake,* Linear Algebra (S. Koh et al., trans.) (1975)
30. *J. S. Golan,* Locelization of Noncommutative Rings (1975)
31. *G. Klambauer,* Mathematicsl Analysis (1975)
32. *M. K. Agoston,* Algebraic Topology (1976)
33. *K. R. Goodearl,* Ring Theory (1976)
34. *L. E. Mansfield,* Linear Algebra with Geometric Applications (1976)
35. *N. J. Pullman,* Matrix Theory and Its Applications (1976)
36. *B. R. McDonald,* Geometric Algebra Over Local Rings (1976)
37. *C. W. Groetsch,* Generalized Inverses of Linear Operators (1977)
38. *J. E. Kuczkowski and J. L. Gersting,* Abstract Algebra (1977)
39. C. O. *Christenson and W. L. Voxman,* Aspects of Topology (1977)
40. *M. Nagata,* Field Theory (1977)
41. *R. L. Long,* Algebraic Number Theory (1977)
42. *W. F. Pfeffer,* Integrals and Measures (1977)
43. *R. L. Wheeden and A. Zygmund,* Measure and Integral (1977)
44. *J. H. Curtiss,* Introduction to Functions of a Complex Variable (1978)
45. *K. Hrbacek and T. Jech,* Introduction to Set Theory (1978)
46. *W. S. Massey,* Homology and Cohomology Theory (1978)
47. *M. Marcus,* Introduction to Modern Algebra (1978)
48. *E. C. Young,* Vector and Tensor Analysis (1978)
49. *S. B. Nadler, Jr.,* Hyperspaces of Sets (1978)
50. *S. K. Segal,* Topics in Group Kings (1978)
51. *A. C. M. van Rooij,* Non-Archimedean Functional Analysis (1978)
52. *L. Corwin and R. Szczarba,* Calculus in Vector Spaces (1979)
53. *C. Sadosky,* Interpolation of Operators and Singular Integrals (1979)
54. *J. Cronin,* Differential Equations (1980)
55. *C. W. Groetsch,* Elements of Applicable Functional Analysis (1980)

56. *I. Vaisman*, Foundations of Three-Dimensional Euclidean Geometry (1980)
57. *H. I. Freedan*, Deterministic Mathematical Models in Population Ecology (1980)
58. *S. B. Chae*, Lebesgue Integration (1980)
59. *C. S. Rees et al.*, Theory and Applications of Fourier Analysis (1981)
60. *L. Nachbin*, Introduction to Functional Analysis (R. M. Aron, trans.) (1981)
61. *G. Orzech and M. Orzech*, Plane Algebraic Curves (1981)
62. *R. Johnsonbaugh and W. E. Pfaffenberger*, Foundations of Mathematical Analysis (1981)
63. *W. L. Voxman and R. H. Goetschel*, Advanced Calculus (1981)
64. *L. J. Corwin and R. H. Szczarba*, Multivariable Calculus (1982)
65. *V. I. Istrățescu*, Introduction to Linear Operator Theory (1981)
66. *R. D. Järvinen*, Finite and Infinite Dimensional Linear Spaces (1981)
67. *J. K. Beem and P. E. Ehrlich*, Global Lorentzian Geometry (1981)
68. *D. L. Armacost*, The Structure of Locally Compact Abelian Groups (1981)
69. *J. W. Brewer and M. K. Smith, eds.*, Emmy Noether: A Tribute (1981)
70. *K. H. Kim*, Boolean Matrix Theory and Applications (1982)
71. *T. W. Wieting*, The Mathematical Theory of Chromatic Plane Ornaments (1982)
72. *D. B. Gauld*, Differential Topology (1982)
73. *R. L. Faber*, Foundations of Euclidean and Non-Euclidean Geometry (1983)
74. *M. Carmeli*, Statistical Theory and Random Matrices (1983)
75. *J. H. Carruth et al.*, The Theory of Topological Semigroups (1983)
76. *R. L. Faber*, Differential Geometry and Relativity Theory (1983)
77. *S. Barnett*, Polynomials and Linear Control Systems (1983)
78. *G. Karpilovsky*, Commutative Group Algebras (1983)
79. *F. Van Oystaeyen and A. Verschoren*, Relative Invariants of Rings (1983)
80. *I. Vaisman*, A First Course in Differential Geometry (1964)
81. *G. W. Swan*, Applications of Optimal Control Theory in Biomedicine (1984)
82. *T. Petrie and J. D. Randall*, Transformation Groups on Manifolds (1984)
83. *K. Goebel and S. Reich*, Uniform Convexity, Hyperbolic Geometry, and Nonexpansive Mappings (1984)
84. *T. Albu and C. Năstăsescu*, Relative Finiteness in Module Theory (1984)
85. *K. Hrbacek and T. Jech*, Introduction to Set Theory: Second Edition (1984)
86. *F. Van Oystaeyen and A. Verschoren*, Relative Invariants of Rings (1984)
87. *B. R. McDonald*, Linear Algebra Over Commutative Rings (1984)
88. *M. Namba*, Geometry of Projective Algebraic Curves (1984)
89. *G. F. Webb*, Theory of Nonlinear Age-Dependent Population Dynamics (1985)
90. *M. R. Bremner et al.*, Tables of Dominant Weight Multiplicities for Representations of Simple Lie Algebras (1985)
91. *A. E. Fekete*, Real Linear Algebra (1985)
92. *S. B. Chae*, Holomorphy and Calculus in Normed Spaces (1985)
93. *A. J. Jerri*, Introduction to Integral Equations with Applications (1985)
94. *G. Karpilovsky*, Projective Representations of Finite Groups (1985)
95. *L. Narici and E. Beckenstein*, Topological Vector Spaces (1985)
96. *J. Weeks*, The Shape of Space (1985)
97. *P. R. Gribik and K. O. Kortanek*, Extremal Methods of Operations Research (1985)
98. *J.-A. Chao and W. A. Woyczynski, eds.*, Probability Theory and Harmonic Analysis (1986)
99. *G. D. Crown et al.*, Abstract Algebra (1986)
100. *J. H. Carruth et al.*, The Theory of Topological Semigroups, Volume 2 (1986)
101. *R. S. Doran and V. A. Belfi*, Characterizations of $C^*$-Algebras (1986)
102. *M. W. Jeter*, Mathematical Programming (1986)
103. *M. Altman*, A Unified Theory of Nonlinear Operator and Evolution Equations with Applications (1986)
104. *A. Verschoren*, Relative Invariants of Sheaves (1987)
105. *R. A. Usmani*, Applied Linear Algebra (1987)
106. *P. Blass and J. Lang*, Zariski Surfaces and Differential Equations in Characteristic $p > 0$ (1987)
107. *J. A. Reneke et al.*, Structured Hereditary Systems (1987)
108. *H. Busemann and B. B. Phadke*, Spaces with Distinguished Geodesics (1987)
109. *R. Harte*, Invertibility and Singularity for Bounded Linear Operators (1988)
110. *G. S. Ladde et al.*, Oscillation Theory of Differential Equations with Deviating Arguments (1987)
111. *L. Dudkin et al.*, Iterative Aggregation Theory (1987)
112. *T. Okubo*, Differential Geometry (1987)
113. *D. L. Stancl and M. L. Stancl*, Real Analysis with Point-Set Topology (1987)

114. *T. C. Gard*, Introduction to Stochastic Differential Equations (1988)
115. *S. S. Abhyankar*, Enumerative Combinatorics of Young Tableaux (1988)
116. *H. Strade and R. Farnsteiner*, Modular Lie Algebras and Their Representations (1988)
117. *J. A. Huckaba*, Commutative Rings with Zero Divisors (1988)
118. *W. D. Wallis*, Combinatorial Designs (1988)
119. *W. Więsław*, Topological Fields (1988)
120. *G. Karpilovsky*, Field Theory (1988)
121. *S. Caenepeel and F. Van Oystaeyen*, Brauer Groups and the Cohomology of Graded Rings (1989)
122. *W. Kozlowski*, Modular Function Spaces (1988)
123. *E. Lowen-Colebunders*, Function Classes of Cauchy Continuous Maps (1989)
124. *M. Pavel*, Fundamentals of Pattern Recognition (1989)
125. *V. Lakshmikantham et al.*, Stability Analysis of Nonlinear Systems (1989)
126. *R. Sivaramakrishnan*, The Classical Theory of Arithmetic Functions (1989)
127. *N. A. Watson*, Parabolic Equations on an Infinite Strip (1989)
128. *K. J. Hastings*, Introduction to the Mathematics of Operations Research (1989)
129. *B. Fine*, Algebraic Theory of the Bianchi Groups (1989)
130. *D. N. Dikranjan et al.*, Topological Groups (1989)
131. *J. C. Morgan II*, Point Set Theory (1990)
132. *P. Biler and A. Witkowski*, Problems in Mathematical Analysis (1990)
133. *H. J. Sussmann*, Nonlinear Controllability and Optimal Control (1990)
134. *J.-P. Florens et al.*, Elements of Bayesian Statistics (1990)
135. *N. Shell*, Topological Fields and Near Valuations (1990)
136. *B. F. Doolin and C. F. Martin*, Introduction to Differential Geometry for Engineers (1990)
137. *S. S. Holland, Jr.*, Applied Analysis by the Hilbert Space Method (1990)
138. *J. Okninski*, Semigroup Algebras (1990)
139. *K. Zhu*, Operator Theory in Function Spaces (1990)
140. *G. B. Price*, An Introduction to Multicomplex Spaces and Functions (1991)
141. *R. B. Darst*, Introduction to Linear Programming (1991)
142. *P. L. Sachdev*, Nonlinear Ordinary Differential Equations and Their Applications (1991)
143. *T. Husain*, Orthogonal Schauder Bases (1991)
144. *J. Foran*, Fundamentals of Real Analysis (1991)
145. *W. C. Brown*, Matrices and Vector Spaces (1991)
146. *M. M. Rao and Z. D. Ren*, Theory of Orlicz Spaces (1991)
147. *J. S. Golan and T. Head*, Modules and the Structures of Rings (1991)
148. *C. Small*, Arithmetic of Finite Fields (1991)
149. *K. Yang*, Complex Algebraic Geometry (1991)
150. *D. G. Hoffman et al.*, Coding Theory (1991)
151. *M. O. González*, Classical Complex Analysis (1992)
152. *M. O. González*, Complex Analysis (1992)
153. *L. W. Baggett*, Functional Analysis (1992)
154. *M. Sniedovich*, Dynamic Programming (1992)
155. *R. P. Agarwal*, Difference Equations and Inequalities (1992)
156. *C. Brezinski*, Biorthogonality and Its Applications to Numerical Analysis (1992)
157. *C. Swartz*, An Introduction to Functional Analysis (1992)
158. *S. B. Nadler, Jr.*, Continuum Theory (1992)
159. *M. A. Al-Gwaiz*, Theory of Distributions (1992)
160. *E. Perry*, Geometry: Axiomatic Developments with Problem Solving (1992)
161. *E. Castillo and M. R. Ruiz-Cobo*, Functional Equations and Modelling in Science and Engineering (1992)
162. *A. J. Jerri*, Integral and Discrete Transforms with Applications and Error Analysis (1992)
163. *A. Charlier et al.*, Tensors and the Clifford Algebra (1992)
164. *P. Biler and T. Nadzieja*, Problems and Examples in Differential Equations (1992)
165. *E. Hansen*, Global Optimization Using Interval Analysis (1992)
166. *S. Guerre-Delabrière*, Classical Sequences in Banach Spaces (1992)
167. *Y. C. Wong*, Introductory Theory of Topological Vector Spaces (1992)
168. *S. H. Kulkarni and B. V. Limaye*, Real Function Algebras (1992)
169. *W. C. Brown*, Matrices Over Commutative Rings (1993)
170. *J. Loustau and M. Dillon*, Linear Geometry with Computer Graphics (1993)
171. *W. V. Petryshyn*, Approximation-Solvability of Nonlinear Functional and Differential Equations (1993)
172. *E. C. Young*, Vector and Tensor Analysis: Second Edition (1993)
173. *T. A. Bick*, Elementary Boundary Value Problems (1993)

174. *M. Pavel*, Fundamentals of Pattern Recognition: Second Edition (1993)
175. *S. A. Albeverio et al.*, Noncommutative Distributions (1993)
176. *W. Fulks*, Complex Variables (1993)
177. *M. M. Rao*, Conditional Measures and Applications (1993)
178. *A. Janicki and A. Weron*, Simulation and Chaotic Behavior of α-Stable Stochastic Processes (1994)
179. *P. Neittaanmäki and D. Tiba*, Optimal Control of Nonlinear Parabolic Systems (1994)
180. *J. Cronin*, Differential Equations: Introduction and Qualitative Theory, Second Edition (1994)
181. *S. Heikkilä and V. Lakshmikantham*, Monotone Iterative Techniques for Discontinuous Nonlinear Differential Equations (1994)
182. *X. Mao*, Exponential Stability of Stochastic Differential Equations (1994)
183. *B. S. Thomson*, Symmetric Properties of Real Functions (1994)
184. *J. E. Rubio*, Optimization and Nonstandard Analysis (1994)
185. *J. L. Bueso et al.*, Compatibility, Stability, and Sheaves (1995)
186. *A. N. Michel and K. Wang*, Qualitative Theory of Dynamical Systems (1995)
187. *M. R. Darnel*, Theory of Lattice-Ordered Groups (1995)
188. *Z. Naniewicz and P. D. Panagiotopoulos*, Mathematical Theory of Hemivariational Inequalities and Applications (1995)
189. *L. J. Corwin and R. H. Szczarba*, Calculus in Vector Spaces: Second Edition (1995)
190. *L. H. Erbe et al.*, Oscillation Theory for Functional Differential Equations (1995)
191. *S. Agaian et al.*, Binary Polynomial Transforms and Nonlinear Digital Filters (1995)
192. *M. I. Gil'*, Norm Estimations for Operation-Valued Functions and Applications (1995)
193. *P. A. Grillet*, Semigroups: An Introduction to the Structure Theory (1995)
194. *S. Kichenassamy*, Nonlinear Wave Equations (1996)
195. *V. F. Krotov*, Global Methods in Optimal Control Theory (1996)
196. *K. I. Beidar et al.*, Rings with Generalized Identities (1996)
197. *V. I. Arnautov et al.*, Introduction to the Theory of Topological Rings and Modules (1996)
198. *G. Sierksma*, Linear and Integer Programming (1996)
199. *R. Lasser*, Introduction to Fourier Series (1996)
200. *V. Sima*, Algorithms for Linear-Quadratic Optimization (1996)
201. *D. Redmond*, Number Theory (1996)
202. *J. K. Beem et al.*, Global Lorentzian Geometry: Second Edition (1996)
203. *M. Fontana et al.*, Prüfer Domains (1997)
204. *H. Tanabe*, Functional Analytic Methods for Partial Differential Equations (1997)
205. *C. Q. Zhang*, Integer Flows and Cycle Covers of Graphs (1997)
206. *E. Spiegel and C. J. O'Donnell*, Incidence Algebras (1997)
207. *B. Jakubczyk and W. Respondek*, Geometry of Feedback and Optimal Control (1998)
208. *T. W. Haynes et al.*, Fundamentals of Domination in Graphs (1998)
209. *T. W. Haynes et al.*, Domination in Graphs: Advanced Topics (1998)
210. *L. A. D'Alotto et al.*, A Unified Signal Algebra Approach to Two-Dimensional Parallel Digital Signal Processing (1998)
211. *F. Halter-Koch*, Ideal Systems (1998)
212. *N. K. Govil et al.*, Approximation Theory (1998)
213. *R. Cross*, Multivalued Linear Operators (1998)
214. *A. A. Martynyuk*, Stability by Liapunov's Matrix Function Method with Applications (1998)
215. *A. Favini and A. Yagi*, Degenerate Differential Equations in Banach Spaces (1999)
216. *A. Illanes and S. Nadler, Jr.*, Hyperspaces: Fundamentals and Recent Advances (1999)
217. *G. Kato and D. Struppa*, Fundamentals of Algebraic Microlocal Analysis (1999)
218. *G. X.-Z. Yuan*, KKM Theory and Applications in Nonlinear Analysis (1999)
219. *D. Motreanu and N. H. Pavel*, Tangency, Flow Invariance for Differential Equations, and Optimization Problems (1999)
220. *K. Hrbacek and T. Jech*, Introduction to Set Theory, Third Edition (1999)
221. *G. E. Kolosov*, Optimal Design of Control Systems (1999)
222. *N. L. Johnson*, Subplane Covered Nets (2000)
223. *B. Fine and G. Rosenberger*, Algebraic Generalizations of Discrete Groups (1999)
224. *M. Väth*, Volterra and Integral Equations of Vector Functions (2000)
225. *S. S. Miller and P. T. Mocanu*, Differential Subordinations (2000)
226. *R. Li et al.*, Generalized Difference Methods for Differential Equations: Numerical Analysis of Finite Volume Methods (2000)

Additional Volumes in Preparation

*H. Li and F. Van Oystaeyen*, A Primer of Algebraic Geometry

*R. P. Agarwal*, Difference Equations and Inequalities: Theory, Methods, and Applications: Second Edition, Revised and Expanded

*A. B. Kharazishvili*, Strange Functions in Real Analysis

*Y. Talpaert*, Differential Geometry

*D. Jagerman*, Difference Equations with Applications to Queues

# GENERALIZED DIFFERENCE METHODS FOR DIFFERENTIAL EQUATIONS

## Numerical Analysis of Finite Volume Methods


**Ronghua Li**
*University of Jilin*
*Changchun, China*

**Zhongying Chen**
*Zhongshan University*
*Guangzhou, China*

**Wei Wu**
*Dalian University of Technology*
*Dalian, China*



MARCEL DEKKER

MARCEL DEKKER, INC. NEW YORK • BASEL

**Library of Congress Cataloging-in-Publication Data**

Li, Ronghua.
Generalized difference methods for differential equations: numerical analysis of finite volume methods / Ronghua Li, Zhongying Chen, Wei Wu.
p. cm. — (Monographs and textbooks in pure and applied mathematics; 226)
Includes bibliographical references and index.
ISBN 0-8247-0330-8 (alk. paper)
1. Finite differences. 2. Differential equations, Partial—Numerical solutions. I. Chen, Zhongying. II. Wu, Wei. III. Title. IV. Series.
QA431.L575 2000
515'.353—dc21 99-051331

This book is printed on acid-free paper.

**Headquarters**
Marcel Dekker, Inc.
270 Madison Avenue, New York, NY 10016
tel: 212-696-9000; fax: 212-685-4540

**Eastern Hemisphere Distribution**
Marcel Dekker AG
Hutgasse 4, Postfach 812, CH-4001 Basel, Switzerland
tel: 41-61-261-8482; fax: 41-61-261-8896

**World Wide Web**
http://www.dekker.com

The publisher offers discounts on this book when ordered in bulk quantities. For more information, write to Special Sales/Professional Marketing at the headquarters address above.



Current printing (last digit)
10 9 8 7 6 5 4 3 2 1

**PRINTED IN THE UNITED STATES OF AMERICA**

# Preface

Finite difference (on rectangular networks) and finite element methods are the two most important classes of numerical methods for partial differential equations. The finite difference method is particularly preferred for hyperbolic equations, especially quasi-linear ones which admit discontinuous solutions. The main defects of the difference method are: the considerable geometrical error of the approximation of curved domains by rectangular grids; the lack of a united and effective approach to deal with natural and internal boundary conditions; the difficulty to construct difference schemes with high accuracy, unless we allow the difference equation to relate more nodal points (which will in turn further increase the difficulty in dealing with boundary conditions). In 1953, R.H. MacNeal used integral interpolation (or integral balance) methods to establish difference schemes on irregular networks. These schemes reduce the geometrical error and, in particular, provide a united and effective approach to handle natural and internal boundary conditions, marking a significant advance in the development of difference methods. But in the following two decades, MacNeal's method did not attract much attention, perhaps because people had turned their attention to finite element methods. Among the few people doing research in the field at that time, I would like to mention A.M. Winslow (1967) and the engineering and mechanics group at Dalian Institute of Technology (1973). They employed the linear finite elements to construct difference schemes on arbitrary triangulations (besides the circumcenter dual grid discussed by MacNeal, they also considered the barycenter dual grid), and applied them to the computation of the electromagnetic fields and the stress of elastic bodies. Since the late seventies, there have been series of papers on difference methods on irregular networks in the former Soviet Union and the former East Germany. Basically they followed MacNeal's approach to construct difference schemes, and adopted the framework of classical difference methods to establish *a priori* estimates (especially the extremum principle), convergences and error estimates. These results were included in B. Heinrich's monograph: Finite Difference Methods on Irregular Networks, ISNM 82, 1987.

Although the difference method on irregular networks has successfully reduced the geometrical error and overcome the difficulty in dealing with natural boundary conditions, it has not resolved the problem of constructing high accuracy difference schemes, and hence cannot match finite element methods in this respect. Besides, its error estimates require too many restrictions and are usually not optimal. Therefore, the theory of finite difference methods is still not as perfect as that of finite element methods. In 1978, the first author of this book utilized finite element spaces and generalized characteristic functions on dual elements, i.e., the common terms of the local Taylor expansions, to rewrite integral interpolation methods in a form of generalized Galerkin methods, and thus obtained a generalization of difference methods on irregular networks, that is, the so-called *generalized difference methods* (GDM for short). Since then, extensive research has been carried out on the theory and application of GDM, such as constructing linear and high order difference schemes for elliptic, parabolic and hyperbolic equations, establishing optimal error estimates in Sobolev norms, and applying GDM to underground fluids, electromagnetic fields and other practical problems. Both the theoretical observations and the computational experiments show that GDM enjoy not only the simplicity of difference methods but also the accuracy of finite element methods. To elaborate, the advantages of GDM are summarized as follows:

1. The grid is flexible (allowing, e.g., triangular and quadrilateral grids), the geometrical error is small, and the natural boundary conditions are easy to deal with.

2. The computational effort is greater than in (classical) finite

difference methods and less than finite element methods, while the accuracy is higher than with finite difference methods and nearly the same as with finite element methods. (The orders of the error estimates of GDM are the same as those of finite element methods, but practical computations show that finite element methods perform slightly better than GDM, perhaps due to the different magnitude of the constants in the error estimates.)

3. The mass conservation law is maintained, which is fairly desirable for, e.g., fluid and underground fluid computations.

4. The theory of GDM is almost as perfect as that of finite element methods. On the other hand, a special case of first order GDM leads to a general and united theory for difference methods on regular and irregular networks.

5. The variational form of GDM (the generalized Galerkin form) is helpful to connect the theories and algorithms of finite element and finite difference methods.

Therefore, GDM are meaningful generalization of difference methods and their further development seems promising. In 1994, the first two authors of this work wrote a book *Generalized Difference Methods for Differential Equations*, published (in Chinese) by Jilin University Press, in which is summarized the research of Chinese researchers on this topic in the preceding ten years.

At the end of the seventies and the beginning of the eighties, some computational fluid researchers (e.g., S. V. Patankar and A. Jameson among others) proposed to apply the difference method on irregular networks to the computation of compressible and incompressible fluid equations. Due to its many advantages, in particular its inheritance of the mass conservation law, this method developed rapidly, and by the end of the seventies, it had become one of the most efficient methods for fluid computation. This method appeared under many different names in the literature. But by the end of the eighties, people usually called it the finite volume method, or finite control volume method, indicating that it is a discrete approximation of the control equation in an integral form. This method is basically equivalent to the generalized difference method with piecewise constant and piecewise linear elements. Using the finite volume method to construct numerical schemes for nonlinear conservative equations actually amounts to generalizing the classical difference schemes (such as Godunov or TVD schemes) to arbitrary grids (including triangular and tetrahedral grids). Not until the end of the eighties, did the numerical analysts get involved in the research of the finite volume method, and by now they have taken it as one of their favorite topics. For the convenience of international communication, we rewrote the Chinese edition of this book in English, and supplemented it with some new materials and recent important references. The original name *Generalized Difference Methods for Differential Equations* survives, but we have added the subtitle *Numerical Analysis of Finite Volume Methods*. In this way we expect to indicate that, on one hand, the generalized difference method is an extension and a development of the finite volume method, and on the other hand, it also provides from another angle a theoretical basis for the finite volume method.

This book is divided into eight chapters and arranged as follows.

Some preliminary materials are gathered in Chapter 1, such as a discussion of Sobolev spaces and the basic results of variational problems and their approximations. In particular, an abstract framework of the generalized difference method is provided in this chapter for later use.

Chapters 2 (except §7) and 3 discuss GDM for one- and two-dimensional second order linear elliptic equations, construct the generalized difference schemes with first-, second- and third-order elements, and establish some fairly comprehensive $H^1$ and $L^2$ error estimates, including certain superconvergence estimates. These results are basically parallel to those of finite element methods, but usually more difficult to prove.

GDM are extended to second-order nonlinear elliptic equations and biharmonic equations in Chapter 4 (and §7 of Chapter 2). As the orders of the partial differential equations increase, the nonconforming feature of GDM becomes more evident, making it more difficult to construct schemes and estimate errors. We introduce in this chapter the GDM based on mixed variational principles (§§2-3) and certain modified variational principles (§§3-4). The corresponding error estimates are also presented.

The GDM for parabolic equations are treated in Chapter 5 in a way similar to the corresponding finite element methods.

The GDM for hyperbolic equations, especially the first order systems, are considered in Chapter 6. The GDM for elliptic equations cannot be directly extended here. Instead we modify a discontinuous finite element method to obtain a generalized upwind scheme with high accuracy. The convergence order is shown. §4 of this chapter discusses briefly the finite volume method for nonlinear conservative equations, and the corresponding references are provided.

Chapter 7 presents the GDM for convection-dominated diffusion equations. The basic idea is to use GDM to discretize the diffusion term, and upwind or high accuracy upwind schemes to the convection term. For the sake of comparison, we also outline in §1 of this chapter the characteristic difference method proposed by Douglas and Russel in the early eighties.

Chapter 8 is devoted to the applications of GDM to plane elasticity problems, electromagnetic fields, groundwater contaminations, Stokes equations, coupled sound-heat flows, and the regularized long wave equations. By virtue of the variational form of GDM, we are also able to extend the hierarchical basis methods for finite element equations to difference equations.

A *Bibliography and Comments* section is attached to the end of each chapter. A complete (to the best of our knowledge) bibliography is provided at the end of the book, which is divided into three groups: A in Chinese, B in English (including a few papers in German and French), and C in Russian.

We would like to thank Jilin University Press for its kind permission for the English edition of this book. Particular recognition is due to Prof. Yuesheng Xu of the North Dakota State University for his enthusiastic encouragement and support. We are grateful to the referees of this manuscript for their careful reading and many valuable suggestions. The research of the authors is partly supported by the National Natural Science Foundation of China. The third author also thanks The Third World Academy for financial support.

Ronghua Li

# Contents

# Chapter 1

# PRELIMINARIES

## 1.1 Sobolev Spaces

Sobolev spaces and their interpolations are basic tools for numerical solutions of partial differential equations. The main related results are outlined in this section, cf. [A-19] and [B-1] for details.

### 1.1.1 Smooth approximations. Fundamental lemma of variational methods

Let $R^n$ be an $n$-dimensional Euclidean space and $\Omega$ a region in $R^n$. $L^p(\Omega)$ $(1 \le p < \infty)$ denotes the set of all the functions defined on $\Omega$ of which the $p$-th powers are integrable, and $L^\infty(\Omega)$ all essentially bounded (i.e. bounded except on a zero measure set) measurable functions. $L^p(\Omega)$ becomes a Banach space if supplied with a norm

$$\|u\|_p = \begin{cases} \left(\int_\Omega |u(x)|^p \mathrm{d}x\right)^{1/p}, \ 1 \le p < \infty, \\ \text{ess} \sup\limits_{x\in\Omega} |u(x)| \equiv \inf\limits_{me=0} \sup\limits_{x\in\Omega - e} |u(x)|, \ p = \infty. \end{cases}$$

Here $me$ denotes the Lebesgue measure of the set $e$. Denote by $C^m(\Omega)$ the set of $m$-th continuously differentiable functions defined on $\Omega$, and $C^\infty(\Omega)$ of infinitely differentiable functions. $C^0(\Omega)$ is simplified as $C(\Omega)$. The closure of the set $\{x \in \Omega : u(x) \neq 0\}$ is called the support of the function $u$ and denoted by supp $u$. $C_0^m(\Omega)$ and $C_0^\infty(\Omega)$

are subsets of $C^m(\Omega)$ and $C^\infty(\Omega)$, respectively, containing functions with compact supports in $\Omega$.

Take any function $j(x)$ satisfying the following conditions:

(i) $j(x) \in C_0^\infty(R^n)$;

(ii) $j(x) \geq 0,\ j(x) = 0$ when $|x| > 1$;

(iii) $\displaystyle\int_{R^n} j(x)\mathrm{d}x = 1.$

For example, we can set

$$j(x) = \begin{cases} \dfrac{1}{\gamma}\exp\Big(-\dfrac{|x|^2}{1-|x|^2}\Big), & |x| < 1, \\ 0, & |x| \geq 1, \end{cases}$$

where

$$\gamma = \int_{|x|<1} \exp\Big(-\frac{|x|^2}{1-|x|^2}\Big)\mathrm{d}x.$$

**Definition 1.1.1** *An integral operator* $J_\epsilon$

$$J_\epsilon u(x) = \int_{R^n} j_\epsilon(x,y)u(y)\mathrm{d}y$$

*with a kernel*

$$j_\epsilon(x,y) = \frac{1}{\epsilon^n} j\Big(\frac{x-y}{\epsilon}\Big) \quad (\epsilon > 0)$$

*is called a smoothing operator, and* $J_\epsilon u$ *an averaging function.*

The following theorem summarizes the main properties of the averaging function.

**Theorem 1.1.1** (Average approximation theorem) *For any function* $u \in L^p(\Omega)$ $(1 \leq p < \infty)$, *we define its value to be zero outside of* $\Omega$. *Then we have*

(i) $J_\epsilon u \in C^\infty(R^n) \cap L^p(\Omega)$, *and* $J_\epsilon u \in C_0^\infty(R^n)$ *when* supp $u$ *is bounded;*

(ii) $\|J_\epsilon u\|_p \leq \|u\|_p$;

(iii) $\lim\limits_{\epsilon \to 0} \|J_\epsilon u - u\|_p = 0$;

Theorem 1.1.1 indicates that the functions in $L^p(\Omega)$ $(1 \le p < \infty)$ can be approximated by sufficiently smooth functions. In other words, $C^\infty(R^n)$ is dense in $L^p(\Omega)$ for $1 \le p < \infty$. Furthermore, we have the following theorem.

**Theorem 1.1.2** *If $1 \le p < \infty$, then $C_0^\infty(\Omega)$ is dense in $L^p(\Omega)$.*

The following theorem can be proved by the average approximation theorem.

**Theorem 1.1.3** (Fundamental lemma of variational methods) *If $u \in L^p(\Omega)$ $(1 \le p < \infty)$ satisfies*

$$\int_\Omega u\phi \mathrm{d}x = 0, \ \forall \phi \in C_0^\infty(\Omega),$$

*then $u = 0$ almost everywhere on $\Omega$.*

**Proof** In fact, $j_\epsilon(x,y) \in C_0^\infty(\Omega)$ for any $0 < \epsilon < \delta$ and $x \in \Omega_\delta = \{x \in \Omega : \mathrm{dist}(x, \partial\Omega) > \delta\}$. So we have

$$J_\epsilon u(x) = \int_\Omega j_\epsilon(x,y)u(y)\mathrm{d}y = 0, \ \forall x \in \Omega_\delta.$$

It follows from Theorem 1.1.1 that

$$\|u\|_{L^p(\Omega_\delta)} = \|u - J_\epsilon u\|_{L^p(\Omega_\delta)} \le \|u - J_\epsilon u\|_{L^p(\Omega)} \to 0 \ (\text{as } \epsilon \to 0).$$

This leads to the desired conclusion. □

### 1.1.2 Generalized derivatives and Sobolev spaces

Write the partial derivative of the function $u$ as

$$\mathrm{D}^\alpha u = \frac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}},$$

where $\alpha = (\alpha_1, \cdots, \alpha_n)$ is an $n$-index, $\alpha_1, \cdots, \alpha_n$ are non-negative integers and $|\alpha| = \alpha_1 + \cdots + \alpha_n$.

**Definition 1.1.2** *Let $L^1_{\rm loc}(\Omega)$ be the local Lebesgue integrable function space and $u \in L^1_{\rm loc}(\Omega)$. If there exists a $v \in L^1_{\rm loc}(\Omega)$ such that*

$$\int_\Omega v\phi \mathrm{d}x = (-1)^{|\alpha|} \int_\Omega u \mathrm{D}^\alpha \phi \mathrm{d}x, \ \forall \phi \in C_0^\infty(\Omega),$$

*then we call $v$ an $|\alpha|$-th generalized derivative of $u$ and write $v = \mathrm{D}^\alpha u$.*

By the fundamental lemma of variational methods, a generalized derivative must be unique as long as it exists. It is easy to show that if a classical derivative of $u$ exists and belongs to $L^2(\Omega)$, then its generalized derivative also exists and is identical with the classical derivative. Hence the generalized derivative is indeed a generalization of the classical one.

Generalized derivatives enjoy the following properties:
(i) $\mathrm{D}^\alpha(au + bv) = a\mathrm{D}^\alpha u + b\mathrm{D}^\alpha v$ ($a, b$ are constants),
(ii) $\mathrm{D}^{\alpha+\beta} u = \mathrm{D}^\alpha(\mathrm{D}^\beta u)$,
(iii) $\mathrm{D}(uv) = v\mathrm{D}u + u\mathrm{D}v \quad \left(\mathrm{D} = \dfrac{\partial}{\partial x_k}\right)$,
(iv) $\mathrm{D}^\alpha u = 0$ for all $\alpha$ with $|\alpha| = m$, if and only if $u$ equals to an $(m-1)$-th polynomial almost everywhere.

**Definition 1.1.3** *Let $m$ be a non-negative integer and $1 \le p \le \infty$. Set*

$$W^{m,p}(\Omega) \equiv \{u \in L^p(\Omega) : \ \mathrm{D}^\alpha u \in L^p(\Omega), \ \forall \alpha, \ 0 \le |\alpha| \le m\},$$

*and supply it with a norm $\|\cdot\|_{m,p}$*

$$\|u\|_{m,p} = \begin{cases} \Big(\sum\limits_{0\le|\alpha|\le m} \|\mathrm{D}^\alpha u\|_p^p\Big)^{1/p}, & \text{for } 1 \le p < \infty, \\ \max\limits_{0\le|\alpha|\le m} \|\mathrm{D}^\alpha u\|_\infty, & \text{for } p = \infty. \end{cases}$$

*Define $W_0^{m,p}(\Omega)$ as the closure of $C_0^\infty(\Omega)$ with respect to the norm $\|\cdot\|_{m,p}$. The normed linear spaces $W^{m,p}(\Omega)$ and $W_0^{m,p}(\Omega)$ are called Sobolev spaces on $\Omega$.*

In particular when $p = 2$ we write $W^{m,p}(\Omega)$ and $W_0^{m,p}(\Omega)$ as $H^m(\Omega)$ and $H_0^m(\Omega)$ , respectively. It is an easy matter to see that $W^{0,p}(\Omega) = L^p(\Omega)$ and $H^0(\Omega) = L^2(\Omega)$.

$W^{m,p}(\Omega)$ and $W_0^{m,p}(\Omega)$ are obviously Banach spaces, and $H^m(\Omega)$ and $H_0^m(\Omega)$ are Hilbert spaces equipped with an inner product

$$(u,v)_m = \sum_{0\le|\alpha|\le m} (\mathrm{D}^\alpha u, \mathrm{D}^\alpha v)_{L^2(\Omega)}, \ u,v \in H^m(\Omega).$$

The norm $\|\cdot\|_{m,p}$ is written as $\|\cdot\|_{m,p,\Omega}$ when the region needs to be specified, and as $\|\cdot\|_m$ when $p=2$ and there is no danger of confusion. We can also introduce a $|\cdot|_{m,p}$ semi-norms

$$|u|_{m,p} = \begin{cases} \Big(\sum_{|\alpha|=m} \|\mathrm{D}^\alpha u\|_{0,p}^p\Big)^{1/p}, & \text{for } 1\le p<\infty, \\ \max_{|\alpha|=m} \|\mathrm{D}^\alpha u\|_{0,\infty}, & \text{for } p=\infty. \end{cases}$$

The following theorem on equivalent norms can be proved by the compact imbedding theorem given later on.

**Theorem 1.1.4** (Equivalent norm theorem) *Suppose $\Omega \subset R^n$ is a bounded L-region; $m\ge 1$; $1\le p\le\infty$; and $l_1,\cdots,l_N$ are bounded linear functionals on $W^{m,p}(\Omega)$ and they are not simultaneously equal to zero on any nonzero polynomial of degree less than or equal to $m-1$. Then the functional*

$$\|u\| \equiv |u|_{m,p} + \sum_{j=1}^N |l_j(u)|$$

*on $W^{m,p}(\Omega)$ is an equivalent norm, that is, there exist constants $\alpha,\beta>0$ such that*

$$\alpha\|u\| \le \|u\|_{m,p} \le \beta\|u\|, \ \forall u \in W^{m,p}(\Omega).$$

**Remark** By an $L$-region $\Omega$ we mean that $\Omega$ has a local Lipschitz boundary, that is, there is a neighborhood $U_x$ for each point $x$ on the boundary of $\Omega$ such that $\partial\Omega \cap U_x$ can be expressed as a Lipschitz continuous function with respect to certain local Cartesian coordinates.

By virtue of the above theorem and the trace theorem (Theorem 1.1.9 below) we know that $\|u\| = |u|_{1,p} + \left|\int_{\partial\Omega} u\mathrm{d}s\right|$ is an equivalent norm for $W^{1,p}(\Omega)$. So there exists $\beta>0$ such that

$$|u|_{0,p} \le \beta|u|_{1,p}, \ \forall u \in W_0^{1,p}(\Omega).$$

Using this and the inductive method leads to the following theorem.

**Theorem 1.1.5** *Let $\Omega \in R^n$ be a bounded L-region and $m \geq 0$, $1 \leq p < \infty$, then $|u|_{m,p}$ is an equivalent norm for $W_0^{m,p}(\Omega)$.*

Some important properties of Sobolev spaces are given in the following theorem.

**Theorem 1.1.6** *Let $\Omega \in R^n$ be a region and $m \geq 1$. Then we have the following:*

(i) *$W^{m,p}(\Omega)$ $(1 \leq p < \infty)$ is separable.*

(ii) *$W^{m,p}(\Omega)$ $(1 < p < \infty)$ is reflexive and uniformly convex.*

(iii) *$\{u \in C^\infty(\Omega) : \|u\|_{m,p} < \infty\}$ is dense in $W^{m,p}(\Omega)$ $(1 \leq p < \infty)$; so $C^\infty(\Omega)$ is dense in $W^{m,p}(\Omega)$ $(1 \leq p < \infty)$; and $C^\infty(\overline{\Omega})$ is dense in $W^{m,p}(\Omega)$ $(1 \leq p < \infty)$ when $\Omega$ is a bounded L-region.*

Property (iii) enables us to make an equivalent definition when $\Omega$ is an $L$-region:

$W^{m,p}(\Omega) \equiv$ the completion of $C^\infty(\overline{\Omega})$ under the norm $\|\cdot\|_{m,p}$.

Next we introduce the Sobolev spaces with negative index. For $1 < p < \infty$, let $p' = p/(p-1)$ be its conjugate index. Write

$$\langle u, v\rangle = \int_\Omega u(x)v(x)\mathrm{d}x.$$

For any $v \in L^{p'}(\Omega)$ we define a bounded linear functional $L_v$ on $W_0^{m,p}(\Omega)$

$$L_v(u) = \langle u, v\rangle, \ \forall u \in W_0^{m,p}(\Omega),$$

and a corresponding norm

$$\|L_v\| = \sup_{u \in W_0^{m,p}(\Omega), \|u\|_{m,p} \leq 1} |L_v(u)|.$$

It can be verified that $V = \{L_v : \ v \in L^{p'}(\Omega)\}$ is dense in $(W_0^{m,p}(\Omega))'$, and hence its closure $\overline{V} = (W_0^{m,p}(\Omega))'$. (The notation $B'$ denotes the dual space of a Banach space $B$.)

**Definition 1.1.4** *Let $1 < p < \infty$, $p' = p/(p-1)$, $v \in L^{p'}(\Omega)$. Define a negative norm of $v$ by $\|v\|_{-m,p'}$*

$$\|v\|_{-m,p'} \equiv \sup_{u \in W_0^{m,p}(\Omega), \|u\|_{m,p} \le 1} |\langle u, v\rangle|.$$

*Correspondingly we define the Sobolev spaces with negative index:*
*$W^{-m,p'}(\Omega)$ $\equiv$the completion of $L^{p'}(\Omega)$ in the norm $\|\cdot\|_{-m,p'}$.*
*When $p' = 2$ we write $H^{-m}(\Omega) = W^{-m,p'}(\Omega)$.*

Notice that $L^{p'}(\Omega)$ and $V$ are isometrically isomorphic, so we have the identification

$$W^{-m,p'}(\Omega) = (W_0^{m,p}(\Omega))'.$$

### 1.1.3 Imbedding and trace theorems

The next two theorems reveal some more profound properties of Sobolev spaces.

**Definition 1.1.5** *Let $X$ and $Y$ be two normed linear spaces. We say that $X$ is imbedded in $Y$, written as $X \to Y$, if*
(i) *$X \subset Y$,*
(ii) *The identification operator $I$ mapping $x \in X$ to $Ix \in Y$ is continuous, i.e., there exists a constant $M > 0$ such that*

$$\|Ix\|_Y \le M\|x\|_X, \ \forall x \in X.$$

*$I$ is called an imbedding operator and $M$ an imbedding constant.*

**Theorem 1.1.7** (Sobolev imbedding theorem) *Suppose that $\Omega \subset R^n$ is a bounded L-region, that $m, k$ are nonnegative constants and that $1 \le p < \infty$; then*
$W^{m+k,p}(\Omega) \to W^{k,q}(\Omega)$ for $1 \le q \le np/(n-mp)$ and $m < n/p$;
$W^{m+k,p}(\Omega) \to W^{k,q}(\Omega)$ for $1 \le q < \infty$ and $m = n/p$;
$W^{m+k,p}(\Omega) \to C^k(\overline{\Omega})$ for $m > n/p$.
*In particular,*
$W^{m,p}(\Omega) \to L^q(\Omega)$ for $1 \le q \le np/(n-mp)$ and $m < n/p$;
$W^{m,p}(\Omega) \to L^q(\Omega)$ for $1 \le q < \infty$ and $m = n/p$;
$W^{m,p}(\Omega) \to C(\overline{\Omega})$ for $m > n/p$.

**Theorem 1.1.8** (Compact imbedding theorem) *Under the assumption of Theorem 1.1.7, the following imbedding operators are compact:*
$W^{m+k,p}(\Omega) \to W^{k,q}(\Omega)$ for $1 \le q < np/(n-mp)$ and $m < n/p$;
$W^{m+k,p}(\Omega) \to W^{k,q}(\Omega)$ for $1 \le q < \infty$ and $m = n/p$;
$W^{m+k,p}(\Omega) \to C^k(\overline{\Omega})$ for $m > n/p$.

It should be pointed out that the elements of $W^{m,p}(\Omega)$ are in fact equivalent classes. Almost equal functions are said to be equivalent and classified into an equivalent class. $W^{m,p}(\Omega) \to C(\overline{\Omega})$ means that any $u \in W^{m,p}(\Omega)$ must be equivalent to a function in $C(\overline{\Omega})$, i.e., the equivalent class $u \in W^{m,p}(\Omega)$ contains an element belonging to $C(\overline{\Omega})$, and that there exists a constant $M$ such that

$$\|u\|_{C(\overline{\Omega})} \le M\|u\|_{m,p,\Omega}, \ \forall u \in W^{m,p}(\Omega).$$

Now let us consider the boundary value of the functions of $H^m(\Omega)$, i.e., the trace $u|_{\partial\Omega}$ of $u$. Suppose a bounded region $\Omega$ possesses an $m$-th smooth boundary $\partial\Omega$. Since $\partial\Omega$ has zero measure in $R^n$, it is meaningless to talk in the usual sense about the value of $u$ on the boundary $\partial\Omega$. Some precise and reasonable definition must be introduced. The idea is to employ the density of $C^m(\overline{\Omega})$ in $H^m(\Omega)$ to generalize the definition.

**Definition 1.1.6** *Assume that $\Omega \subset R^n$ is a bounded region with an $m$-th smooth boundary $\partial\Omega$ and that $u \in C^m(\overline{\Omega})$. The linear operator $(\gamma_0, \gamma_2, \cdots, \gamma_{m-1})$ is called a trace operator, where*

$$\gamma_j u = \frac{\partial^j u}{\partial n^j}\Big|_{\partial\Omega}, \ 0 \le j \le m-1,$$

*and $\frac{\partial^j}{\partial n^j}$ denotes the $j$-th directional derivative on the outer normal direction of $\partial\Omega$.*

**Lemma 1.1.1** *For the above mentioned region we have a constant $C > 0$ such that*

$$\|\gamma_j u\|_{0,\partial\Omega} \le C\|u\|_{j+1,\Omega}, \ \forall u \in C^m(\overline{\Omega}), \ 0 \le j \le m-1.$$

For $u \in H^m(\Omega)$ we can choose a sequence $\{u_k\} \subset C^m(\overline{\Omega})$ such that $\|u - u_k\|_{m,\Omega} \to 0$ as $k \to \infty$. It follows from the lemma that $\{\gamma_j u_k\}$ is a Cauchy sequence in $L^2(\partial\Omega)$. So there is a limit $v_j \in L^2(\partial\Omega)$. $v_j$ is obviously independent of the choice of $\{u_k\}$. Thus, we can define the trace of $u \in H^m(\Omega)$ on $\partial\Omega$ as

$$\gamma_j u = v_j = \lim_{k\to\infty} \gamma_j u_k.$$

**Theorem 1.1.9** *Suppose that $\Omega \subset R^n$ is a bounded region with an m-th smooth boundary and that $u \in H^m(\Omega)$. Then there exists a constant $C > 0$ independent of $u$ such that*

$$\|\gamma_j u\|_{0,\partial\Omega} \le C\|u\|_{j+1,\Omega}, \ \forall 0 \le j \le m-1.$$

*In particular,*

$$\|u\|_{0,\partial\Omega} \le C\|u\|_{1,\Omega}.$$

*The last inequality (the imbedding $H^1(\Omega) \to L^2(\partial\Omega)$) only requires $\partial\Omega$ to be a Lipschitz continuous surface.*

Finally we point out that it follows from the definition of the trace operator that

$$H_0^m(\Omega) = \left\{u \in H^m(\Omega) : \gamma_j u = \frac{\partial^j u}{\partial n^j}\Big|_{\partial\Omega} = 0, \ 0 \le j \le m-1\right\}.$$

### 1.1.4 Finite element spaces

Essentially a numerical method for differential equations means discretizations of the infinite dimensional function spaces and approximations of the original equations by equations in finite dimensional spaces. The interpolation is a basic method to construct these finite dimensional spaces, and the error estimates of the approximate solutions and the true solutions often rely on the error estimates of the interpolate approximations. Our interpolate approximation problem involves Sobolev spaces, with the approximating finite dimensional spaces being piecewise polynomial spaces (finite element spaces) subject to certain constraints. The related concepts and results constitute the so-called interpolation theory of Sobolev spaces.

In this subsection, let us first introduce the finite element spaces. Let $T_h$ be a decomposition of a region $\overline{\Omega}$, dividing $\overline{\Omega}$ into finite bounded closed sets $K'$s which possess Lipschitz continuous boundaries, share no common inner points and have nonempty interior. So $\overline{\Omega} = \bigcup_{K\in T_h} K$. Here $K$ is called an element of $T_h$ and $h$ stands for the largest element diameter.

**Definition 1.1.7** *A finite dimensional space $V_h$ is called a finite element space with respect to the decomposition $T_h$ if we have the following:*

(i) *For each $K \in T_h$, the set $P_K \equiv \{p : \ p = v_h|_K, \ \forall v_h \in V_h\}$ is a family of polynomials. And there exists a set of freedoms $\sum_K = \{l_i, 1 \le i \le N\}$ (namely a set of linearly independent linear functionals, often presented as a group of parameters $\{\alpha_i, 1 \le i \le N\}$, c.f. the examples below), which is $P_K$-uniquely solvable: for any given $\{\alpha_i, 1 \le i \le N\}$ there exists a unique function $p \in P_K$ satisfying*

$$l_i(p) = \alpha_i, \ 1 \le i \le N;$$

(ii) *The functions of $V_h$ possess certain smoothness on $\Omega$, e.g., $V_h \subset C^m(\overline{\Omega})$ ($m$ is a non-negative integer).*

*The triple $\{K, P_K, \sum_K\}$ specifies a finite element space.*

We observe the following fact:

$$V_h = \{v_h \in C^m(\overline{\Omega}) : \ v_h|_K \in P_K, \ K \in T_h\} \subset H^{m+1}(\Omega).$$

Next we give some examples of finite elements.

**Triangulation.** Suppose $\overline{\Omega} \subset R^2$ can be decomposed into finite triangles such that different triangles have no overlap interior region, and a vertex of any triangle does not belong to the interior of a side of any other triangle. All such triangles form a decomposition of $\overline{\Omega}$, called a triangulation and denoted by $T_h = \{K\}$. The element $K$ with vertexes $a_1, a_2$ and $a_3$ can be expressed as

$$K = \Big\{(x,y) : \ (x,y) = \sum_{i=1}^{3} \lambda_i a_i, \ 0 \le \lambda_i \le 1 \ (1 \le i \le 3), \sum_{i=1}^{3} \lambda_i = 1\Big\}.$$

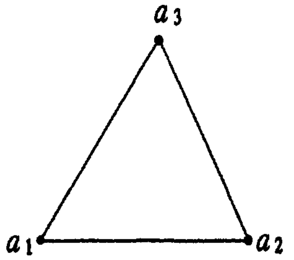



**Fig. 1.1.1**

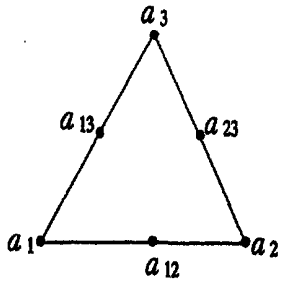



**Fig. 1.1.2**

$\lambda_i = \lambda_i(x, y)$ $(i = 1, 2, 3)$ are called area coordinates of the point $(x, y)$. Denote by $\mathcal{P}_k(K)$ the set of polynomials on $K$ of degrees less than or equal to $k$.

**Example 1** Lagrange linear element.

$K = \Delta a_1 a_2 a_3$ (c.f. Fig.1.1.1);

$P_K = \mathcal{P}_1(K)$, $\dim P_K = 3$;

$\sum_K = \{p(a_i) : i = 1, 2, 3\}$.

Any $p \in P_K$ is determined uniquely by its values on the vertexes $a_1, a_2$ and $a_3$:

$$p = \sum_{i=1}^{3} p(a_i)\lambda_i.$$

It is easy to see that the corresponding finite element space $V_h \subset C(\overline{\Omega})$, and hence $V_h \subset H^1(\Omega)$.

**Example 2** Lagrange quadratic element.

$K = \Delta a_1 a_2 a_3$, and the midpoints $a_{ij} = \frac{1}{2}(a_i + a_j)$ (see Fig.1.1.2);

$P_K = \mathcal{P}_2(K)$, $\dim P_K = 6$;

$\sum_K = \{p(a_i), 1 \le i \le 3; p(a_{ij}), 1 \le i < j \le 3\}$.

Any $p \in P_K$ is determined uniquely by its values on the vertexes and the midpoints:

$$p = \sum_{i=1}^{3} \lambda_i(2\lambda_i - 1)p(a_i) + \sum_{i<j} 4\lambda_i\lambda_j p(a_{ij}).$$

The corresponding finite element space $V_h \subset C(\overline{\Omega}) \cap H^1(\Omega)$.

**Example 3** Hermite cubic element.

$K = \triangle a_1a_2a_3$, and the barycenter $a_{123} = \frac{1}{3}(a_1 + a_2 + a_3)$;

$P_K = \mathcal{P}_3(K)$, $\dim P_K = 10$;

$\sum_K = \left\{p(a_i), \dfrac{\partial p(a_i)}{\partial x}, \dfrac{\partial p(a_i)}{\partial y}, 1 \le i \le 3; p(a_{123})\right\}.$

Any $p \in P_K$ has the following expression:

$$\begin{aligned} p = & \sum_{i=1}^{3}(-2\lambda_i^3 + 3\lambda_i^2 - 7\lambda_1\lambda_2\lambda_3)p(a_i) + 27\lambda_1\lambda_2\lambda_3 p(a_{123}) \\ & + \sum_{i \ne j} \lambda_i\lambda_j(2\lambda_i + \lambda_j - 1)\mathrm{D}p(a_i) \cdot (a_j - a_i), \end{aligned}$$

where $\mathrm{D}p(a) = \left(\dfrac{\partial p(a)}{\partial x}, \dfrac{\partial p(a)}{\partial y}\right)$. The corresponding finite element space $V_h \subset C(\overline{\Omega}) \cap H^1(\Omega)$.

**Example 4** Restricted Hermite cubic element (Zienkiewicz element).

$K = \triangle a_1a_2a_3$;

$P_K = \left\{p \in \mathcal{P}_3(K) : p(a_{123}) = \dfrac{1}{3}\sum_{i=1}^{3} p(a_i) - \dfrac{1}{6}\sum_{i=1}^{3} \mathrm{D}p(a_i) \cdot (a_i - a_{123})\right\},$

$\dim P_K = 9$, $P_K \supset \mathcal{P}_2(K)$;

$\sum_K = \left\{p(a_i), \dfrac{\partial p(a_i)}{\partial x}, \dfrac{\partial p(a_i)}{\partial y}, 1 \le i \le 3\right\}.$

The expression of $p \in P_K$ can be obtained by inserting the constraint condition on $p(a_{123})$ in the definition of $P_K$ into the second term of

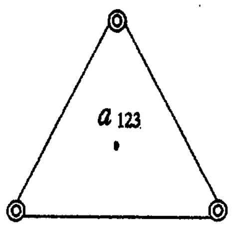



**Fig. 1.1.3**

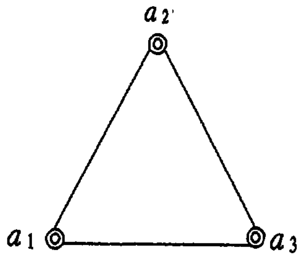



**Fig. 1.1.4**

the expression in Example 3. The corresponding finite element space $V_h \subset C(\overline{\Omega}) \cap H^1(\Omega)$.

**Rectangular grid.** Suppose the region $\overline{\Omega} \subset R^2$ can be divided into a sum of finite number of rectangles with each side of the rectangles being parallel to the axes of coordinates. Any two different rectangles either are disjoint or share a common side or vertex. All the rectangles constitute a grid of $\overline{\Omega}$, called a rectangular grid. The affine mapping.

$$\xi = (x - x_i)/\Delta x, \ \eta = (y - y_i)/\Delta y$$

maps the rectangle

$$K = \{(x, y) : x_i \le x \le x_i + \Delta x, y_i \le y \le y_i + \Delta y\}$$

onto a unit square $[0, 1; 0, 1]$. Denote by $\mathcal{Q}_k$ the set of polynomials of $x$ and $y$ with degrees less than or equal to $k$. Note $\mathcal{P}_k \subset \mathcal{Q}_k \subset \mathcal{P}_{2k}$.

**Example 5** Lagrange bilinear element (c.f. Fig. 1.1.5).

$K =$ rectangle $a_1 a_2 a_3 a_4$;

$P_K = \mathcal{Q}_1(K), \ \dim P_K = 4$;

$\sum_K = \{p(a_i), 1 \le i \le 4\}$;

$\forall p \in P_K, \ p = \sum_{i=1}^{4} p(a_i)\mu_i,$

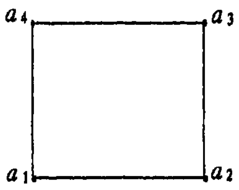



Fig. 1.1.5

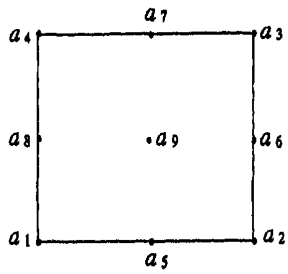



Fig. 1.1.6

where

$$\mu_1 = (1-\xi)(1-\eta),\ \mu_2 = \xi(1-\eta),\ \mu_3 = \xi\eta,\ \mu_4 = (1-\xi)\eta.$$

The corresponding finite element space $V_h \subset C(\overline{\Omega}) \cap H^1(\Omega)$.

**Example 6** Lagrange bi-quadratic element (Fig. 1.1.6).

$K =$ rectangle $a_1a_2a_3a_4$, midpoints $a_5, a_6, a_7, a_8$, barycenter $a_9$;

$P_K = Q_2(K)$, $\dim P_K = 9$;

$\sum_K = \{p(a_i), 1 \le i \le 9\}$;

$\forall p \in P_K,\ p = \sum_{i=1}^{9} p(a_i)\mu_i,$

where

$$\begin{aligned}
\mu_1 &= (2\xi-1)(\xi-1)(2\eta-1)(\eta-1),\\
\mu_2 &= \xi(2\xi-1)(2\eta-1)(\eta-1),\\
\mu_3 &= \xi(2\xi-1)\eta(2\eta-1),\\
\mu_4 &= (2\xi-1)(\xi-1)\eta(2\eta-1),\\
\mu_5 &= 4\xi(1-\xi)(2\eta-1)(\eta-1),\\
\mu_6 &= 4\xi(2\xi-1)\eta(1-\eta),\\
\mu_7 &= 4\xi(1-\xi)\eta(2\eta-1),
\end{aligned}$$

$$\mu_8 = 4(2\xi - 1)(\xi - 1)\eta(1 - \eta),$$
$$\mu_9 = 16\xi(1 - \xi)\eta(1 - \eta).$$

The corresponding finite element space $V_h \subset C(\overline{\Omega}) \cap H^1(\Omega)$.

**Example 7** Hermite bi-cubic element (Bogner-Fox-Schmidt rectangular element).

$K =$ rectangle $a_1 a_2 a_3 a_4$;

$P_K = \mathcal{Q}_3(K)$, $\dim P_K = 16$;

$$\sum\nolimits_K = \left\{ p(a_i), \frac{\partial p(a_i)}{\partial x}, \frac{\partial p(a_i)}{\partial y}, \frac{\partial^2 p(a_i)}{\partial x \partial y}, 1 \le i \le 4 \right\}.$$

The corresponding finite element space $V_h \subset C^1(\overline{\Omega}) \cap H^2(\Omega)$.

**Example 8** Hermite incomplete cubic element (Adini rectangular element).

$K =$ rectangle $a_1 a_2 a_3 a_4$;

$$P_K = \{p = p_3(x,y) + \alpha x^3 y + \beta x y^3 : p_3(x,y) \in \mathcal{P}_3(K), \alpha, \beta \in \mathrm{R}\},\ \dim P_K = 12;$$

$$\sum\nolimits_K = \left\{ p(a_i), \frac{\partial p(a_i)}{\partial x}, \frac{\partial p(a_i)}{\partial y}, 1 \le i \le 4 \right\}.$$

The corresponding finite element space $V_h \subset C(\overline{\Omega}) \cap H^1(\Omega)$.

In the above examples, the set of freedoms $\sum_K$ is composed of the following "interpolation functionals":

$$l_i^{(\alpha)} : p \to \mathrm{D}^\alpha p(a_i),$$

where $a_i$'s are nodes of the finite element, $\alpha = (\alpha_1, \alpha_2)$, and the orders of the derivatives $|\alpha| = 0, 1, 2$. Other kinds of freedoms can also be considered. For more examples of finite elements see, e.g., [A-2] and [B-17].

### 1.1.5 Interpolation error estimates in Sobolev spaces

The main results of the interpolation theory in Sobolev spaces are given in this subsection.

**Definition 1.1.8** *For a given finite element $\{K, P_K, \sum_K\}$, we call $\Pi_K v$ a $P_K$-interpolation of $v \in C^s(K)$ (where $s$ is the highest order of the partial derivatives in $\sum_K$), if*

$$\Pi_K v \in P_K,$$

$$l(\Pi_K v) = l(v),\ \forall l \in \textstyle\sum_K .$$

*$\Pi_K : C^s(K) \to P_K$ is then called a $P_K$-interpolation operator. Let $V_h$ be the finite element space related to the grid $T_h = \{K\}$. We define $\Pi_h v$, the $V_h$-interpolation of $v \in C^s(\overline{\Omega})$, by*

$$\Pi_h v \in V_h,$$

$$\Pi_h v|_K = \Pi_K v,\ \forall K \in T_h.$$

*$\Pi_h : C^s(\overline{\Omega}) \to V_h$ is referred to as a $V_h$-interpolation operator.*

Now our main task is to provide the error estimates, in the norm $\|\cdot\|_{m,q,K}$, of $v \in W^{k+1,p}(K)$ and its $P_K$-interpolation $\Pi_K v$, under the imbedding conditions $W^{k+1,p}(K) \to C^s(K)$ and $W^{k+1,p}(K) \to W^{m,q}(K)$. First we show a relationship between the norm and semi-norm of quotient spaces.

**Theorem 1.1.10** *Take a quotient norm*

$$\|\dot{v}\|_{k+1,p,\Omega} \equiv \inf_{p \in P_k} \|v + p\|_{k+1,p,\Omega},\ \dot{v} \in W^{k+1,p}(\Omega)/\mathcal{P}_k,$$

*in the quotient space $W^{k+1,p}(\Omega)/\mathcal{P}_k$, where $v$ is any element in the equivalent class $\dot{v}$. Then there exists a constant $C = C(\Omega)$ such that*

$$\|\dot{v}\|_{k+1,p,\Omega} \leq C|\dot{v}|_{k+1,p,\Omega},\ \forall \dot{v} \in W^{k+1,p}(\Omega)/\mathcal{P}_k.$$

An easy consequence of Theorem 1.1.10 is the following abstract estimation for linear operators invariant with polynomials of degree at most $k$.

**Theorem 1.1.11** *Let $\Omega$ be a bounded open set with a Lipschitz continuous boundary. If*

(i) $W^{k+1,p}(\Omega) \to W^{m,q}(\Omega)$,

(ii) $\Pi$ *is a bounded linear operator from* $W^{k+1,p}(\Omega)$ *to* $W^{m,q}(\Omega)$, *and is invariant on* $\mathcal{P}_k$*:*

$$\Pi p = p, \ \forall p \in \mathcal{P}_k,$$

*then there exists a constant* $C = C(\Omega)$ *such that*

$$|u - \Pi v|_{m,q,\Omega} \le C\|I - \Pi\| \cdot |v|_{k+1,p,\Omega}, \ \forall v \in W^{k+1,p}(\Omega).$$

The constants $C$ and $\|I - \Pi\|$ in the above estimate depend on the region $\Omega$ and the operator $\Pi$ respectively. To obtain an error estimate for a family of finite elements, we need to relate these finite elements with a special finite element through an affine mapping.

**Definition 1.1.9** *Two finite elements* $\{\hat{K}, \hat{P}, \hat{\Sigma}\}$ *and* $\{K, P, \Sigma\}$ *are said to be (affine) equivalent, if there exists an invertible affine mapping* $F: \ \hat{x} \in \hat{K} \to x = F(\hat{x}) \in K$ *satisfying*

$$\begin{aligned} &K = F(\hat{K}), \\ &P = \{p = \hat{p} \circ F^{-1}, \hat{p} \in \hat{P}\}, \\ &\Sigma = \{l_i : p \to \hat{l}_i(p), \ \forall p = \hat{p} \circ F^{-1}, \hat{l}_i \in \hat{\Sigma}\}. \end{aligned}$$

*A family of finite elements is referred to as an affine family, if all the finite elements in the family are (affine) equivalent to a certain finite element, called the reference element of the family. An affine family is said to be regular if there exists a constant* $\sigma$ *such that for all* $K$

$$h_K/\rho_K \le \sigma, \ \text{and}\ h_K \to 0,$$

*where* $h_K = \text{diam}(K)$, $\rho_K = \sup\{\text{diam}(S) : \textit{the ball}\ S \subset K\}$.

The next result reveals a relationship between the Sobolev seminorms of a function before and after an affine mapping.

**Theorem 1.1.12** *Let* $\hat{\Omega}$ *and* $\Omega$ *be two affine equivalent open sets in* $R^n$, *i.e., there is an invertible affine mapping*

$$F: \ \hat{x} \in \hat{\Omega} \to F(\hat{x}) = B\hat{x} + b \in \Omega,$$

*such that*

$$\Omega = F(\hat{\Omega}),$$

*where $B$ is an $n \times n$ nonsingular matrix and $b$ an $n$-dimensional vector. Then there exists a constant $C = C(m, n, p)$ independent of $\Omega$ and $F$ such that for any $v \in W^{m,p}(\Omega)$*

$$|\hat{v}|_{m,p,\hat{\Omega}} \leq C\|B\|^m |\det B|^{-1/p} |v|_{m,p,\Omega},$$

*where $\hat{v}(\hat{x}) = v(F(\hat{x}))$. Conversely, for any $\hat{v} \in W^{m,p}(\hat{\Omega})$*

$$\|v\|_{m,p,\Omega} \leq C\|B^{-1}\|^m |\det B|^{1/p} |\hat{v}|_{m,p,\hat{\Omega}},$$

*where $v(x) = \hat{v}(F^{-1}(x))$. Besides,*

$$\|B\| \leq h_\Omega/\rho_{\hat{\Omega}}, \ \|B^{-1}\| \leq h_{\hat{\Omega}}/\rho_\Omega,$$

$$|\det B| = \text{meas}(\Omega)/\text{meas}(\hat{\Omega}),$$

*where $h_\Omega = \text{diam}\Omega$, $\rho_\Omega = \sup\{\text{diam}S :$ the ball $S \subset \Omega\}$.*

Transfer, by virtue of Theorem 1.1.12, the estimate for the interpolation error $v - \Pi_K v$ of the finite element $\{K, P, \Sigma\}$ to the reference element $\{\hat{K}, \hat{P}, \hat{\Sigma}\}$, and apply the abstract estimate given in Theorem 1.1.11 to the reference element, then we have the following important result.

**Theorem 1.1.13** *For a given regular family of finite elements, suppose the reference element $\{\hat{K}, \hat{P}, \hat{\Sigma}\}$ satisfies*

$$W^{k+1,p}(\hat{K}) \to C^s(\hat{K}),$$

$$W^{k+1,p}(\hat{K}) \to W^{m,q}(\hat{K}),$$

$$\mathcal{P}_k \subset \hat{P} \subset W^{m,q}(\hat{K}),$$

*where $s$ is the highest order of the derivatives in $\hat{\Sigma}$; $m, k$ are nonnegative integers; and $1 \leq p, q \leq \infty$. Then there exists a constant $C$ independent of $K$ such that for every finite element $K$ in the family and any function $v \in W^{k+1,p}(K)$*

$$|v - \Pi_K v|_{m,p,K} \leq C(h_K^n)^{\frac{1}{q}-\frac{1}{p}} h_K^{k+1-m} |v|_{k+1,p,K}.$$

*In particular, if $p = q = 2$ then*

$$|v - \Pi_K v|_{m,K} \leq C h_K^{k+1-m} |v|_{k+1,K}.$$

Finally we present an inverse property of the finite element. To this end we need further assumptions for the grid.

**Definition 1.1.10** *A family of grids $\{T_h\}$ is said to be quasi-uniform if there are constants $\sigma$ and $\gamma$ such that*

$$h_K/\rho_K \le \sigma,\ h/h_K \le \gamma,\ \forall K \in T_h,\ h > 0.$$

**Theorem 1.1.14** (Inverse property) *Let a family of finite elements with quasi-uniform grids be given. Assume that $V_h$ is a finite element space related to a grid $T_h$, that $l$, $m$ are nonnegative integers, that $1 \le r,\ q \le \infty$, and that*

$$l \le m,\ \hat{P} \subset W^{l,r}(\hat{K}) \cap W^{m,q}(\hat{K}).$$

*Then there exists a constant $C = C(\sigma, \gamma, l, m, r, q)$ such that for all $v_h \in V_h$*

$$\Big(\sum_{K\in T_h} |v_h|^q_{m,q,K}\Big)^{1/q} \le \frac{C}{(h^n)^{\max\{0,\frac{1}{r}-\frac{1}{q}\}} h^{m-l}} \Big(\sum_{K\in T_h} |v_h|^r_{l,r,K}\Big)^{1/r}.$$

*Here we make a convention for $q = \infty$ that*

$$\Big(\sum_{K\in T_h} |v_h|^q_{m,q,K}\Big)^{1/q} = \max_{K\in T_h} |v_h|_{m,\infty,K}.$$

*In particular, when $r = q = 2$ we have*

$$\Big(\sum_{K\in T_h} |v_h|^2_{m,K}\Big)^{1/2} \le Ch^{l-m}\Big(\sum_{K\in T_h} |v_h|^2_{l,K}\Big)^{1/2}.$$

## 1.2 Variational Problems and Their Approximations

### 1.2.1 Abstract variational form

Let $H$ be a real Hilbert space, equipped with an inner product $(\cdot,\cdot)$ and the corresponding norm $\|\cdot\|$. $V$ is a subspace of $H$ satisfying

$\overline{V} = H$ according to the norm $\|\cdot\|$. $V$ becomes a Hilbert space with respect to an inner product $[\cdot,\cdot]$ and its related norm $|\cdot|$. The imbedding of $V$ in $H$ is continuous, that is, there exists a constant $\gamma > 0$ such that

$$\|v\| \leq \gamma|v|, \ \forall v \in V.$$

Suppose $A$ is a linear operator from a dense linear subspace $D(A)$ of $V$ to the dual space $V'$ of $V$, satisfying $\overline{D(A)} = V$ (in the norm $|\cdot|$). For $f \in V'$ consider an operator equation

$$Au = f. \tag{1.2.1}$$

In many cases, equation (1.2.1) does not necessarily have a solution in $D(A)$. Thus we need to extend the operator $A$ in some sense and then to discuss the generalized solution of the problem. To this end, we construct a bilinear form:

$$a(u,v) \equiv \langle Au, v\rangle = Au(v), \ \forall u,v \in D(A)$$

(where the notation $\langle\cdot,\cdot\rangle$ denotes the dual pair of $V' \times V$), and make the following basic hypothesis:

(H) $a(u,v)$ is a bounded bilinear form on $D(A) \times D(A)$, i.e., there exists a constant $M$ such that

$$|a(u,v)| \leq M|u||v|, \ \forall u,v \in D(A).$$

Now we consider the problem in the space $V$. For any $u \in D(A)$, $\langle Au, v\rangle$ is a bounded linear functional of $v$ on $V$. By the Riesz representation theorem there is a unique element $\mathcal{A}u \in V$ satisfying

$$\langle Au, v\rangle = [\mathcal{A}u, v], \ u \in D(A), \ v \in V.$$

Similarly for each $f \in H$ we have a unique $Rf \in V$ such that

$$\langle f, v\rangle = [Rf, v], \ v \in V.$$

**Proposition 1.2.1** *Let (H) hold. Then $\mathcal{A} : D(A) \to V$ can be uniquely extended into a bounded linear operator $T : V \to V$, and* *the corresponding bilinear form can be (uniquely and continuously) extended onto* $V \times V$*:*

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

*and we have*

$$|a(u,v)| \le M|u| \ |v|, \ \forall u, v \in V.$$

**Proof** For any $u \in V$, if $u \in D(A)$, then set $Tu = \mathcal{A}u$; If $u \notin D(A)$, set $Tu = \lim_{j\to\infty} \mathcal{A}u_j$ (in $V$), where $\{u_j\} \subset D(A)$, $\lim_{j\to\infty} u_j = u$ (in $V$). Such a sequence $\{u_j\}$ does exist since $\overline{D(A)} = V$ (according to $|\cdot|$), and $\{\mathcal{A}u_j\}$ converges in $V$ by condition (H). The limit $Tu$ is obviously independent of the choice of $\{u_j\}$. It is easy to check that the bilinear form remains to be bounded after the extension, and hence $T$ is a bounded linear operator:

$$|Tu| = \sup_{v\in V, |v|\le 1} |[Tu, v]| \le M|u|, \ \forall u \in V.$$

Finally, this continuous extension is apparently unique. This completes the proof. □

Now in place of (1.2.1), we consider the operator equation

$$Tu = Rf,$$

or equivalently

$$[Tu, v] = [Rf, v], \ \forall v \in V,$$

that is,

$$a(u,v) = \langle f, v\rangle, \ \forall v \in V. \tag{1.2.2}$$

**Definition 1.2.1** *The solution* $u \in V$ *of the equation (1.2.2) is called the Galerkin generalized (or weak) solution of the original equation (1.2.1), and the solution of (1.2.1) in* $D(A)$ *the classical solution.*

The following conclusion is obvious.

**Theorem 1.2.1** *If* $u \in D(A)$ *is a classical solution of (1.2.1), then it is also a Galerkin generalized solution. Conversely, if* $u$ *is a Galerkin generalized solution of (1.2.1) and* $u \in D(A)$*, then it is also a classical solution.*

The above result is called a variational principle in Galerkin form, and equation (1.2.2) a variational problem in Galerkin form with respect to (1.2.1). In mechanics, $v$ stands for the virtual displacement, $a(u,v)-\langle f,v\rangle$ the virtual work, and (1.2.2) the virtual work equation. So we also call Theorem 1.2.1 a virtual work principle.

Assume that $a(\cdot,\cdot)$ is symmetric. Let us introduce a quadratic functional

$$J(v)=\frac{1}{2}a(v,v)-\langle f,v\rangle,\ \forall v\in V,$$

and consider the following functional minimization problem: Find $u\in V$ such that

$$J(u)=\inf_{v\in V}J(v). \tag{1.2.3}$$

Many practical problems (e.g. the elastic problems) can be deduced into this form.

**Definition 1.2.2** *The solution $u\in V$ of the problem (1.2.3) is referred to as a Riesz generalized (or weak) solution of the original equation (1.2.1).*

**Theorem 1.2.2** *Assume that $a(\cdot,\cdot)$ is symmetric and positive definite:*

$$a(u,v)=a(v,u),\ \forall u,v\in V, \tag{1.2.4}$$

$$a(v,v)\geq\alpha|v|^2,\ \forall v\in V\ (\alpha \text{ a positive constant}). \tag{1.2.5}$$

*Then the problems (1.2.3) and (1.2.2) are equivalent. So under the above assumptions, if $u\in D(A)$ is a classical solution of equation (1.2.1), then it is also a Riesz geneneralized solution; and conversely, if $u$ is a Riesz generalized solution of (1.2.1) and $u\in D(A)$, then it is a classical solution.*

**Remark** It is the condition (1.2.5) that is called the $V$-elliptic condition, also often referred to as a coercive, or positive definite condition.

**Proof** We consider for $u, v \in V$

$$\begin{aligned}\phi(t) &\equiv J(u+tv)\\ &= \frac{1}{2}a(u+tv, u+tv) - \langle f, u+tv\rangle\\ &= J(u) + t[a(u,v) - \langle f, v\rangle] + \frac{t^2}{2}a(v,v).\end{aligned}$$

The symmetry condition (1.2.4) is used in the last equality. It follows from the positive definite condition (1.2.5) that $u$ is the minimum function of the functional $J(u)$ if and only if $\phi'(0) = 0$ i.e.

$$a(u,v) = \langle f, v\rangle, \ \forall v \in V.$$

The second half of the theorem results from Theorem 1.2.1. This completes the proof. □

Theorem 1.2.2 is called a Riesz variational principle, and problem (1.2.3) a Riesz variational problem corresponding to (1.2.1). In mechanics, the quadratic functional $J(u)$ represents the energy of the system. The above conclusion illustrates that in all the possible displacements satisfying the given boundary constraints, the displacement that makes the balance minimizes the total potential energy. Therefore, Theorem 1.2.2 is also called the minimum potential energy principle.

The virtual work principle is more general and has wider applications than the potential energy principle. It applies not only to symmetric and positive definite problems (corresponding to conservative field equations) but also to asymmetric and nonpositive definite problems (nonconservative field equations).

### 1.2.2 Green's formulas and variational problems

For differential equations, the continuous extension of the operators and the bilinear forms as well as the deduction of the variational forms mentioned in the above subsection are realized by integrations in parts or by the use of Green's formulas.

Let $\Omega$ be a bounded open region in $R^N$ with a Lipschitz continuous boundary $\Gamma = \partial\Omega$, and $n = (n_1, \cdots, n_N)$ the unit outer normal vector.

Then the following Green's formula holds:

$$-\int_\Omega \frac{\partial u}{\partial x_i} v\mathrm{d}x = \int_\Omega u\frac{\partial v}{\partial x_i}\mathrm{d}x - \int_\Gamma uvn_i\mathrm{d}s,\ \forall u, v \in H^1(\Omega). \quad (1.2.6)$$

Replacing $u$ by $\frac{\partial u}{\partial x_i}$ and summing for $i$ lead to the first Green's formula:

$$-\int_\Omega \Delta uv\mathrm{d}x = \int_\Omega \sum_{i=1}^N \frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_i}\mathrm{d}x - \int_\Gamma \frac{\partial u}{\partial n} v\mathrm{d}s,$$

$$\forall u \in H^2(\Omega),\ v \in H^1(\Omega). \quad (1.2.7)$$

Exchange the positions of $u$ and $v$ to get another equality, and subtract the above equality from it, then we have the second Green's formula:

$$\int_\Omega (v\Delta u - u\Delta v)\mathrm{d}x = \int_\Gamma \Big(\frac{\partial u}{\partial n}v - \frac{\partial v}{\partial n}u\Big)\mathrm{d}s,\ \forall u, v \in H^2(\Omega). \quad (1.2.8)$$

If we replace $u$ above by $\Delta u$, then we obtain another Green's formula:

$$\int_\Omega v\Delta^2 u\mathrm{d}x = \int_\Omega \Delta u\Delta v\mathrm{d}x + \int_\Gamma \Big(\frac{\partial \Delta u}{\partial n}v - \Delta u\frac{\partial v}{\partial n}\Big)\mathrm{d}s,$$

$$\forall u \in H^4(\Omega), v \in H^2(\Omega). \quad (1.2.9)$$

It is easy to show for $N = 2$ that

$$\int_\Omega \Big(2\frac{\partial^2 u}{\partial x_1\partial x_2}\frac{\partial^2 v}{\partial x_1\partial x_2} - \frac{\partial^2 u}{\partial x_1^2}\frac{\partial^2 v}{\partial x_2^2} - \frac{\partial^2 u}{\partial x_2^2}\frac{\partial^2 v}{\partial x_1^2}\Big)\mathrm{d}x_1\mathrm{d}x_2$$

$$= \int_\Gamma \Big(-\frac{\partial^2 u}{\partial \tau^2}\frac{\partial v}{\partial n} + \frac{\partial^2 u}{\partial \tau\partial n}\frac{\partial v}{\partial \tau}\Big)\mathrm{d}s,\ \forall u \in H^3(\Omega), v \in H^2(\Omega). \quad (1.2.10)$$

Here $\tau = (\tau_1, \tau_2)$ is the unit tangent vector along $\Gamma$, $\frac{\partial}{\partial \tau}$ the derivative along the tangent direction, and

$$\frac{\partial^2 u}{\partial \tau^2} = \mathrm{D}^2 u\cdot(\tau, \tau) = \sum_{i,j=1}^2 \tau_i\tau_j\frac{\partial^2 u}{\partial x_i\partial x_j},$$

$$\frac{\partial^2 u}{\partial \tau\partial n} = \mathrm{D}^2 u\cdot(\tau, n) = \sum_{i,j=1}^2 \tau_i n_j\frac{\partial^2 u}{\partial x_i\partial x_j}.$$

As an example of variational problems, let us consider the mixed boundary value problem of the Poisson equation:

$$\begin{cases} -\Delta u = f(x,y),\ (x,y) \in \Omega, & (1.2.11a) \\ u|_{\Gamma_1} = 0, & (1.2.11b) \\ \Big(\dfrac{\partial u}{\partial n} + \alpha u\Big)\Big|_{\Gamma_2} = 0, & (1.2.11c) \end{cases}$$

where $\Omega$ is a bounded plane region. Its boundary $\Gamma$ is a piecewise smooth simple closed curve, divided into two disjoint parts $\Gamma_1$ and $\Gamma_2$. $\alpha > 0$ and $f \in L^2(\Omega)$.

Set

$$V = H_E^1(\Omega) = \{v : v \in H^1(\Omega), v|_{\Gamma_1} = 0\},$$

$$a(u,v) = \iint_\Omega \Big(\frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y}\Big)\mathrm{d}x\mathrm{d}y + \int_{\Gamma_2} \alpha uv\mathrm{d}s,$$

$$\langle f, v\rangle = \iint_\Omega fv\mathrm{d}x\mathrm{d}y.$$

Multiply (1.2.11a) by $v \in V$, integrate it on $\Omega$, and employ Green's formula (1.2.7) and the boundary conditions (1.2.11b,c) to obtain

$$\iint_\Omega -v\Delta u\mathrm{d}x\mathrm{d}y = \iint_\Omega \Big(\frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y}\Big)\mathrm{d}x\mathrm{d}y + \int_{\Gamma_2} \alpha uv\mathrm{d}s.$$

Now the variational problem corresponding to problem (1.2.11) becomes: Find $u \in V$ such that

$$a(u,v) = \langle f, v\rangle,\ \forall v \in V. \tag{1.2.12}$$

Here $a(u,v)$ is a continuous extension of $(-\Delta u, v)$ thanks to Green's formula. This kind of continuous extension on $V \times V$ is unique and hence (1.2.12) is identical with the variational form mentioned in the above subsection.

Note that $|\cdot|_1$ is an equivalent norm of $H_E^1$, and that $a(\cdot,\cdot)$ is symmetric and positive definite:

$$a(u,v) = a(v,u),\ \forall u,v \in V,$$

$$a(v,v) \geq |v|_1^2,\ \forall v \in V.$$

As an example of high order equations, we consider the first boundary value problem of the biharmonic equation:

$$\begin{cases} \Delta^2 u = \dfrac{\partial^4 u}{\partial x^4} + 2\dfrac{\partial^4 u}{\partial x^2 \partial y^2} + \dfrac{\partial^4 u}{\partial y^4} = f, \ (x,y) \in \Omega, & (1.2.13\text{a}) \\ u|_\Gamma = 0, & (1.2.13\text{b}) \\ \dfrac{\partial u}{\partial n}\Big|_\Gamma = 0, & (1.2.13\text{c}) \end{cases}$$

where $f \in L^2(\Omega)$.

It follows from Green's formula (1.2.9) that

$$\iint_\Omega v\Delta^2 u \mathrm{d}x\mathrm{d}y = \iint_\Omega \Delta u \Delta v \mathrm{d}x\mathrm{d}y, \ v \in H_0^2(\Omega).$$

Write

$$V = H_0^2(\Omega),$$
$$a(u,v) = \iint_\Omega \Delta u \Delta v \mathrm{d}x\mathrm{d}y,$$
$$\langle f, v\rangle = \iint_\Omega f v \mathrm{d}x\mathrm{d}y.$$

Then the variational problem related to (1.2.13) becomes: Find $u \in V$ such that

$$a(u,v) = \langle f, v\rangle, \ \forall v \in V. \tag{1.2.14}$$

These kind of problems arise particularly in fluid mechanics.

Obviously $a(\cdot,\cdot)$ is a symmetric bilinear form. As for its positive definiteness, we first note

$$a(v,v) = |\Delta v|_{0,\Omega}^2 = \iint_\Omega \Big[\Big(\frac{\partial^2 v}{\partial x^2}\Big)^2 + 2\Big(\frac{\partial^2 v}{\partial x \partial y}\Big)^2 + \Big(\frac{\partial^2 v}{\partial y^2}\Big)^2\Big]\mathrm{d}x\mathrm{d}y.$$

Apply twice Green's formula (1.2.6), then we have

$$\begin{aligned} \iint_\Omega \Big(\frac{\partial^2 v}{\partial x \partial y}\Big)^2 \mathrm{d}x\mathrm{d}y &= -\iint_\Omega \frac{\partial v}{\partial x}\frac{\partial^3 v}{\partial x \partial y^2}\mathrm{d}x\mathrm{d}y \\ &= \iint_\Omega \frac{\partial^2 v}{\partial x^2}\frac{\partial^2 v}{\partial y^2}\mathrm{d}x\mathrm{d}y, \ \forall v \in C_0^\infty(\Omega). \end{aligned}$$

The density of $C_0^\infty(\Omega)$ in $H_0^2(\Omega)$ implies that

$$\iint_\Omega \Big(\frac{\partial^2 v}{\partial x \partial y}\Big)^2 \mathrm{d}x\mathrm{d}y = \iint_\Omega \frac{\partial^2 v}{\partial x^2}\frac{\partial^2 v}{\partial y^2}\mathrm{d}x\mathrm{d}y,\ \forall v \in H_0^2(\Omega).$$

So we have

$$a(v,v) = |\Delta v|_{0,\Omega}^2 = |v|_{2,\Omega}^2,\ \forall v \in H_0^2(\Omega).$$

Now the positive definiteness of $a(\cdot,\cdot)$ follows since $|\cdot|_{2,\Omega}$ is an equivalent norm on $H_0^2(\Omega)$.

By Green's formula (1.2.10) and boundary conditions (1.2.13b,c), the above bilinear form can be rewritten as

$$\begin{aligned} & a(u,v) \\ &= \iint_\Omega \Big[\Delta u \Delta v + (1-\sigma)\Big(2\frac{\partial^2 u}{\partial x \partial y}\frac{\partial^2 v}{\partial x \partial y} - \frac{\partial^2 u}{\partial x^2}\frac{\partial^2 v}{\partial y^2} - \frac{\partial^2 u}{\partial y^2}\frac{\partial^2 v}{\partial x^2}\Big)\Big]\mathrm{d}x\mathrm{d}y \\ &= \iint_\Omega \Big[\sigma\Delta u \Delta v + (1-\sigma)\Big(2\frac{\partial^2 u}{\partial x \partial y}\frac{\partial^2 v}{\partial x \partial y} + \frac{\partial^2 u}{\partial x^2}\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\frac{\partial^2 v}{\partial y^2}\Big)\Big]\mathrm{d}x\mathrm{d}y, \end{aligned}$$

where $\sigma$ is the Poinsson ratio satisfying $0 < \sigma < \frac{1}{2}$. This corresponds to the variational form for the clamped plate problem and it also applies to the plate bending with other boundary constraints. The bilinear form is again symmetric and positive definite:

$$a(v,v) = \sigma|\Delta v|_{0,\Omega}^2 + (1-\sigma)|v|_{2,\Omega}^2.$$

### 1.2.3 Well-posedness of variational problems

The following important and widely used result concerns the well-posedness of variational problems, namely the solution's existence, uniqueness and continuous dependence on the right-hand side term.

**Theorem 1.2.3** (Lax-Milgram) *Let $V$ be a real Hilbert space and $a(\cdot,\cdot)$ a bilinear form defined on $V \times V$, satisfying the following conditions:*

(i) *Boundedness: There exists a constant $M > 0$ such that*

$$|a(u,v)| \le M|u||v|,\ \forall u, v \in V.$$

(ii) *Positive definiteness: There exists a constant* $\alpha > 0$ *such that*

$$|a(v,v)| \geq \alpha |v|^2, \ \forall v \in V.$$

*Then, the variational problem (1.2.2) has a unique solution* $u \in V$ *for any given* $f \in V'$ *and*

$$|u| \leq \frac{1}{\alpha} \|f\|_{V'}.$$

The next theorem is a generalization of Theorem 1.2.3.

**Theorem 1.2.4** (Babuska) *Let* $U$ *and* $V$ *be two real Hilbert spaces, supplied with inner products* $[\cdot,\cdot]_U$, $[\cdot,\cdot]_V$ *and norms* $|\cdot|_U$, $|\cdot|_V$ *respectively.* $a(\cdot,\cdot)$ *is a bilinear form on* $U \times V$, *satisfying the following conditions:*

(i) *Boundedness: There exists a constant* $M > 0$ *such that*

$$|a(u,v)| \leq M |u|_U |v|_V, \ \forall u \in U, \ v \in V. \tag{1.2.15}$$

(ii) *Weak positive definiteness: There exists a constant* $\alpha > 0$ *such that*

$$\inf_{\substack{u \in U \\ |u|_U = 1}} \sup_{\substack{v \in V \\ |v|_V = 1}} |a(u,v)| \geq \alpha, \tag{1.2.16}$$

$$\sup_{u \in U} |a(u,v)| > 0, \ \forall v \in V, v \neq 0. \tag{1.2.17}$$

*Then, for any given* $f \in V'$ *there exists a unique* $u \in U$ *such that*

$$a(u,v) = \langle f, v \rangle, \ \forall v \in V, \tag{1.2.18}$$

*and*

$$|u|_U \leq \frac{1}{\alpha} \|f\|_{V'}. \tag{1.2.19}$$

**Proof** For any $u \in U$, the boundedness guarantees that $a(u,\cdot)$ is a bounded linear functional on $V$. So by the Riesz representation theorem there is a $Tu \in V$ such that

$$a(u,v) = [Tu, v]_V, \ \forall v \in V.$$

Similarly for $f \in V'$ there exists an $Rf \in V$ such that

$$\langle f, v\rangle = [Rf, v]_V, \ \forall v \in V.$$

Therefore, (1.2.18) is equivalent to a bounded linear operator equation:

$$Tu = Rf. \tag{1.2.20}$$

The weak positive definiteness implies

$$\inf_{\substack{u\in U\\ |u|_U=1}} |Tu|_V \geq \alpha > 0,$$

which indicates that the operator $T$ has a bounded inverse operator $T^{-1}$ satisfying

$$\|T^{-1}\| \leq \alpha^{-1}.$$

In particular, the range of $T$ is a closed linear subspace of $V$.

Now we claim that the range of $T$ is indeed the whole space $V$. If this is not true, then by virtue of the projection theorem there must be a $v_0 \in V, v_0 \neq 0$ satisfying

$$[Tu, v_0] = a(u, v_0) = 0, \ \forall u \in U.$$

But this contradicts (1.2.17). Therefore, (1.2.18) or (1.2.20) must have a unique solution $u = T^{-1}Rf$. Finally (1.2.19) holds since

$$\begin{aligned} |u|_U &\leq \alpha^{-1}|Rf|_V = \alpha^{-1} \sup_{\substack{v\in V\\ |v|_V=1}} |[Rf, v]_V| \\ &= \alpha^{-1} \sup_{\substack{v\in V\\ |v|_V=1}} |\langle f, v\rangle| = \alpha^{-1}\|f\|_{V'}. \end{aligned}$$

This completes the proof. □

### 1.2.4 Approximation methods. A necessary and sufficient condition for approximate-solvability

Next we discuss the approximation methods. Notice that the abstract variational problem is equivalent to a bounded linear operator equation. So in this subsection, we discuss in a more general setting the approximation methods for bounded linear operator equations and present a necessary and sufficient condition for the unique approximate-solvability.

Suppose $U$ and $V$ are two reflexive Banach spaces. We denote both the norms by $\|\cdot\|$. Let $T$ be a bounded linear operator from $U$ to $V$. We consider the approximation methods for the operator equation

$$Tu = g, \tag{1.2.21}$$

where $g \in V$ is given.

The essence of the finite dimensional approximation is to discretize the problem and replace the original equation by an approximate equation in a finite dimensional space. First we need to discretize the solution space and the range space, and accordingly to construct an approximation of the original equation. Different strategies of discretization and construction lead to different numerical methods. Generally, we will choose in a certain way finite dimensional subspaces $U_n$ and $V_n$ of $U$ and $V$, and the mappings $P_n : U \to U_n$ and $Q_n : V \to V_n$, respectively. Then set $T_n = Q_nT|_{U_n}$ and take

$$T_nu_n = Q_ng \ \ (u_n \in U_n) \tag{1.2.22}$$

as the approximation equation of (1.2.21).

$P_n$ and $Q_n$ are linear operators for most of the approximation methods. If we choose $P_n$ and $Q_n$ as linear projection operators, then we end up with the so-called projection methods. They become the usual Galerkin methods if $U = V$, $U_n = V_n$. If $T$ is a differential or integral operator, and $U_n$ is a spline function space (or a piecewise polynomial space), then we have the finite element method. If $U$ and $V$ are function spaces on $\Omega$, and $Q_n : \ V \to V_n$ is an interpolation operator:

$$Q_nv(x) = \sum_{i=1}^{N} v(x_i)\psi_i(x), \ v \in V,$$

where $x_1, \cdots, x_N \in \Omega$, and $\{\psi_1 \cdots, \psi_N\}$ is a basis of $V_n$, then the approximation equation (1.2.22) reads

$$\sum_{i=1}^{N}(Tu_n)_{x=x_i}\psi_i(x) = \sum_{i=1}^{N}(g)_{x=x_i}\psi_i(x),$$

which is equivalent to

$$(Tu_n)_{x=x_i} = (g)_{x=x_i},\ i = 1, 2, \cdots, N. \qquad (1.2.23)$$

This is the so-called collocation method. If $U$ and $V$ are Hilbert spaces; $(\cdot,\cdot)$ is the inner product of $V$; $U_n = \text{span}\{\phi_1, \cdots, \phi_N\}$ and $V_n = \text{span}\{\psi_1, \cdots, \psi_N\}$; $P_n$ and $Q_n$ are orthogonal projection operators; then the approximation equation (1.2.22) is equivalent to

$$(Tu_n - g, \psi_i) = 0\ , i = 1, 2, \cdots, N, \qquad (1.2.24)$$

which is the Petrov-Galerkin method, also called the generalized Galerkin method. There are many different choices for $\{\phi_i\}$ and $\{\psi_i\}$. Choosing $\psi_i = M\phi_i$ for some suitably chosen linear operator $M$ leads to a moment method. If in particular $\psi_i = T\phi_i$, then it becomes the least square method, since in this case (1.2.24) is equivalent to the problem of finding $u \in U_n$ to minimize $\|Tu_n - g\|$. The case of $U = V$, $\psi_i = \phi_i$ corresponds to the Galerkin method.

Now we turn to discuss the approximate-solvability. Let $\{U_n\}$ and $\{V_n\}$ be sequences of finite dimensional subspaces of $U$ and $V$, respectively. $P_n : U \to U_n$ and $Q_n : V \to V_n$ are respectively the linear projection operators, satisfying $P_nP_m = P_n$ and $Q_nQ_m = Q_n$ for $n \leq m$ and the following properties:

(1) $U_n \subset U_{n+1},\ V_n \subset V_{n+1},\ n = 1, 2, \cdots;$

(2) $\overline{\bigcup_{n=1}^{\infty} U_n} = U,\ \overline{\bigcup_{n=1}^{\infty} V_n} = V;$

(3) $\|P_n\| \leq C,\ \|Q_n\| \leq C,\ n = 1, 2, \cdots$. ($C$ is a constant.)

**Proposition 1.2.2** *Under the above assumptions we have*

(i) $\forall u \in U,\ P_nu \to u,$ as $n \to \infty$;

(ii) $\forall v \in V,\ Q_nv \to v,$ as $n \to \infty$;

(iii) $\forall l \in U',\ P_n'l \to l,$ as $n \to \infty$;

(iv) $\forall l \in V'$, $Q_n' l \to l$, as $n \to \infty$;
*where $U'$ and $V'$ are the dual spaces of $U$ and $V$, and $P_n'$ and $Q_n'$ are the dual operators of $P_n$ and $Q_n$, respectively.*

**Proof** To show (i), we note from conditions (1) and (2) that for any $u \in U$ there exists a $u_n \in U_n$ $(n = 1, 2, \cdots)$ satisfying

$$\|u - u_n\| \to 0, \ n \to \infty.$$

Then it follows from the triangular inequality and condition (3) that

$$\begin{aligned} \|P_n u - u\| &\le \|P_n(u - u_n)\| + \|u_n - u\| \\ &\le (C+1)\|u_n - u\| \to 0, \ n \to \infty. \end{aligned}$$

Now we deal with (iii). Write $W_n = P_n' U'$. It results from $P_n P_m = P_n$ when $n \le m$ that $P_m' P_n' = P_n'$. Hence $W_n \subset W_m$ $(n \le m)$, and in particular $W_n \subset W_{n+1}$.

By the reflexivity of the space $U$, if $\overline{\bigcup_{n=1}^{\infty} W_n} = U'$ does not hold, then there exists a $u_0 \in U$, $u_0 \ne 0$ such that

$$\langle l, u_0 \rangle = 0, \ \forall l \in \bigcup_{n=1}^{\infty} W_n.$$

So for any $l \in U'$ and $n \ge 1$ we have

$$\langle l, P_n u_0 \rangle = \langle P_n' l, u_0 \rangle = 0,$$

which implies $P_n u_0 = 0$. But $P_n u_0 \to u_0$ as $n \to \infty$. So $u_0 = 0$, yielding a contradiction. Therefore we must have $\overline{\bigcup_{n=1}^{\infty} W_n} = U'$.

It follows from condition (3) that $\|P_n'\| \le C$. Thus, we can show (iii) as in the proof to (i).

(ii) and (iv) can be similarly proved. This completes the proof.□

**Proposition 1.2.3** *Under the above conditions we have ($\rightharpoonup$ stands for the weak convergence):*

(i) *If $\{u_j\} \subset U$ and $u_j \rightharpoonup u \in U$ $(j \to \infty)$, then $Tu_j \rightharpoonup Tu \in V$.*

(ii) *If $u \in U$, then $T_n P_n u \rightharpoonup Tu$ $(n \to \infty)$.*

(iii) *If $\{u_j\} \subset U$, $u_j \in U_{n_j}$ $(j = 1, 2 \cdots)$, $n_j \to \infty\,(j \to \infty)$; and $u_j \rightharpoonup u \in U$ $(j \to \infty)$, then $T_{n_j} u_j \rightharpoonup Tu$ $(j \to \infty)$.*

**Proof** (i) For any $l \in V'$ and $u_j \rightharpoonup u$ $(j \to \infty)$ we have

$$\langle l, Tu_j \rangle = \langle T'l, u_j \rangle \to \langle T'l, u \rangle = \langle l, Tu \rangle,$$

which means $Tu_j \rightharpoonup Tu$ $(j \to \infty)$.

(ii) For $u \in U$ it follows from Proposition 1.2.2 that

$$\begin{aligned} &\|T_n P_n u - Tu\| \\ &\leq \|Q_n T(P_n u - u)\| + \|Q_n Tu - Tu\| \\ &\leq C\|T\|\|P_n u - u\| + \|Q_n Tu - Tu\| \to 0 \ (n \to \infty). \end{aligned}$$

(iii) We know by (i) that $Tu_j \rightharpoonup Tu$ $(j \to \infty)$. By Proposition 1.2.2 we have for any $l \in V'$ that $Q'_{n_j} l \to l$ $(j \to \infty)$. This gives

$$\langle l, T_{n_j} u_j \rangle = \langle Q'_{n_j} l, Tu_j \rangle \to \langle l, Tu \rangle \ (j \to \infty).$$

This implies $T_{n_j} u_j \rightharpoonup Tu$ $(j \to \infty)$ and completes the proof. □

**Definition 1.2.3** *Equation (1.2.21) is said to be uniquely approximate-solvable, if there exists an integer $N > 0$ such that for $n \geq N$ equation (1.2.22) possesses a unique solution $u_n \in U_n$, and $\{u_n\}$ converges in $U$ as $n \to \infty$, of which the limit $u \in U$ is the unique solution of (1.2.21).*

**Theorem 1.2.5** *Let $U$ and $V$ be reflexive Banach spaces, $T : U \to V$ a bounded linear operator, and $U_n, V_n, P_n, Q_n$ as above. Then, a sufficient and necessary condition for equation (1.2.21) to be uniquely approximate-solvable for any given $g \in V$ is that there exists an integer $N > 0$ and a constant $\alpha > 0$ such that*

$$\|Q_n Tu\| \geq \alpha \|u\|, \ \forall u \in U_n, \ n \geq N, \tag{1.2.25}$$

*or equivalently*

$$\varliminf_{n \to \infty} \inf_{\substack{u \in U_n \\ u \neq 0}} \left\{ \frac{\|Q_n Tu\|}{\|u\|} \right\} > 0.$$

*In this case, we have the following error estimate for $u_n$*

$$\|u - u_n\| \leq \left(1 + \frac{C}{\alpha}\|T\|\right) \inf_{w \in U_n} \|u - w\|. \tag{1.2.26}$$

**Proof** Necessity. That for any $g \in V$ equation (1.2.22) always has a unique solution $u_n \in U_n$ means that there exists an inverse operator $T_n^{-1} : V_n \to U_n$ and the range of $T_n$ is $V_n$. Moreover, $T_n^{-1}$ is bounded by the inverse operator theorem. For the operator $T_n^{-1}Q_n : V \to U_n \subset U$, we note

$$\forall g \in V,\ T_n^{-1}Q_n g = u_n \to u \ \ (n \to \infty),$$

where $u_n$ and $u$ are the solutions of (1.2.22) and (1.2.21) respectively. So it follows from the resonance theorem that the sequence of operators $\{T_n^{-1}Q_n\}$ is uniformly bounded, namely, there exists a constant $\beta > 0$ such that

$$\|T_n^{-1}Q_n\| \le \beta \ (n \le N).$$

So (1.2.25) is valid:

$$\|T_n u\| \ge \frac{1}{\beta}\|u\|,\ \forall u \in U_n,\ n \ge N.$$

Sufficiency. Condition (1.2.25) implies that when $n \ge N$, for any $g \in V$ equation (1.2.22) has a unique solution $u_n \in U_n$ and

$$\|u_n\| \le \frac{C}{\alpha}\|g\|.$$

The reflexivity of the space $U$ gives the existence of a weakly convergent subsequence $\{u_{n_j}\}$ satisfying

$$u_{n_j} \in U_{n_j},\ u_{n_j} \rightharpoonup u \in U \ (j \to \infty).$$

By (iii) of Proposition 1.2.3 we have

$$T_{n_j}u_{n_j} \rightharpoonup Tu \ \ (j \to \infty).$$

On the other hand

$$T_{n_j}u_{n_j} = Q_{n_j}g \to g \ \ (j \to \infty).$$

Hence $u$ is the solution to (1.2.21).

Let us show $u_n \to u$ $(j \to \infty)$. In fact, by virtue of (1.2.25) for any $\tilde{u} \in U_n$

$$\begin{aligned} \|u_n - u\| &\leq \|u_n - \tilde{u}\| + \|\tilde{u} - u\| \\ &\leq \frac{1}{\alpha}\|Q_n T(u_n - \tilde{u})\| + \|\tilde{u} - u\| \\ &= \frac{1}{\alpha}\|Q_n T(u - \tilde{u})\| + \|u - \tilde{u}\| \\ &\leq \Big(\frac{C}{\alpha}\|T\| + 1\Big)\|u - \tilde{u}\|. \end{aligned}$$

This gives (1.2.26). Setting $\tilde{u} = P_n u$ yields $u_n \to u$ $(n \to \infty)$.

Finally we claim that condition (1.2.25) guarantees the uniqueness of the solution to (1.2.21). In fact, if $w \in U$ satisfies $Tw = 0$, then it follows from (1.2.25) and (ii) of Proposition 1.2.3 that

$$\|P_n w\| \leq \frac{1}{\alpha}\|T_n P_n w\| \to 0 \ (n \to \infty),$$

which gives $w = 0$. This completes the proof. □

### 1.2.5 Galerkin methods

Let us recall the framework of the abstract variational problem in §1.2.1. Suppose $H$ and $V$ are the Hilbert spaces mentioned there, $A$ is the linear operator from a dense subset $D(A)$ of $V$ to $V'$, and $f \in V'$. Consider the operator equation

$$Au = f. \tag{1.2.27}$$

We assume the bilinear form $a(u, v) = \langle Au, v\rangle$ is bounded on $D(A) \times D(A)$:

$$|a(u, v)| \leq M|u||v|, \ \forall u, v \in V,$$

which enables us to continuously extend $a(u, v)$ to $V \times V$. Thus we have obtained a variational, or a weak, form:

$$a(u, v) = \langle f, v\rangle, \ \forall v \in V. \tag{1.2.28}$$

It is equivalent to an operator equation:

$$Tu = Rf. \tag{1.2.28$'$}$$

A relationship between them is revealed by the Riesz representation theorem:

$$a(u,v) = [Tu, v], \; u, v \in V,$$
$$\langle f, v\rangle = [Rf, v], \; f \in V', v \in V.$$

Apparently $T : V \to V$ is a bounded linear operator with $\|T\| \le M$.

Let $V$ be a separable Hilbert space. Then one can choose a sequence of finite dimensional subspaces $\{V_n\}$ of $V$ such that

$$V_n \subset V_{n+1} \; (n = 1, 2, \cdots), \; \overline{\bigcup_{n=1}^{\infty} V_n} = V.$$

Denote by $\Pi_n$ the orthogonal projection operator from $V$ to $V_n$. Then $\Pi_n$ is a self-adjoint linear operator satisfying $\|\Pi_n\| = 1$ and

$$\lim_{n\to\infty} |\Pi_n v - v| = 0, \; \forall v \in V.$$

The Galerkin method for the variational problem (1.2.28) is: Find $u_n \in V_n$ such that

$$a(u_n, v_n) = \langle f, v_n\rangle, \; \forall v_n \in V_n, \tag{1.2.29}$$

or equivalently

$$T_n u_n = \Pi_n R f \;\; (T_n = \Pi_n T|_{V_n}). \tag{1.2.29$'$}$$

Let $\{\phi_1, \cdots, \phi_N\}$ be a basis of $V_n$. Write $u_n$ as

$$u_n = \sum_{i=1}^{N} c_i \phi_i,$$

insert it into (1.2.29), and take $v_n = \phi_j$ $(1 \le j \le N)$, then we have

$$\sum_{i=1}^{N} a(\phi_i, \phi_j) c_i = \langle f, \phi_j\rangle, \; j = 1, 2, \cdots, N. \tag{1.2.29$''$}$$

Solving it for $c_i$ $(1 \le i \le N)$ yields an approximate solution $u_n$. (1.2.29) (or an equivalent form) is called a Galerkin approximate equation, and $u_n$ a Galerkin approximate solution.

**Theorem 1.2.6** *Let $V$ be a separable real Hilbert space, $V_n$ a finite dimensional subspace of $V$, and $a(\cdot,\cdot)$ a bilinear form defined on $V\times V$ possessing the following properties:*

(i) *Boundedness: There is a constant $M > 0$ such that:*

$$|a(u,v)| \leq M|u||v|, \ \forall u, v \in V;$$

(ii) *Positive definiteness: There exists a constant $\alpha > 0$ such that*

$$a(v,v) \geq \alpha|v|^2, \ \forall v \in V.$$

*Then, both the variational problem (1.2.28) and the approximation problem (1.2.29) have unique solutions $u$ and $u_n$, respectively. Moreover, the following error estimate holds:*

$$|u - u_n| \leq \frac{M}{\alpha} \inf_{v_n \in V_n} |u - v_n|. \tag{1.2.30}$$

**Proof** By Theorem 1.2.3, (1.2.28) has a unique solution. It follows from (ii) that the homogeneous equation

$$a(u_n, v_n) = 0, \ \forall v_n \in V_n$$

corresponding to (1.2.29) admits only the trivial solution. Thus (1.2.29) has a unique solution for any given $f$. (1.2.28) and (1.2.29) lead to the error equation

$$a(u - u_n, v_n) = 0, \ \forall v_n \in V_n. \tag{1.2.31}$$

So (1.2.30) follows from

$$|u - u_n|^2 \leq \frac{1}{\alpha} a(u - u_n, u - u_n)$$
$$= \frac{1}{\alpha} a(u - u_n, u - v_n) \leq \frac{M}{\alpha} |u - u_n||u - v_n|, \ \forall v_n \in V_n.$$

This completes the proof. □

Theorem 1.2.6 can also be deduced from Theorem 1.2.5 by noting that the positive definiteness implies (1.2.25).

(1.2.30) gives an error estimate of the approximate solution $u_n$ in $|\cdot|$ norm, i.e., a $V$-estimate. Next we turn to discuss the error estimate of $u_n$ in $\|\cdot\|$ norm, namely the $H$-estimate. To this end we use the Aubin-Nitsche dual argument.

**Theorem 1.2.7** *Let the assumptions of Theorem 1.2.6 hold. Then*

$$\|u - u_n\| \le M|u - u_n|\Big(\sup_{g\in H}\frac{1}{\|g\|}\inf_{\tilde{\phi}\in V_n}|\phi - \tilde{\phi}|\Big). \tag{1.2.32}$$

*Here for any $g \in H$, $\phi \in V$ is the unique solution to the dual variational problem:*

$$a(v, \phi) = (g, v), \ \forall v \in V. \tag{1.2.33}$$

**Proof** The unique solvability of the dual problem (1.2.33) can be proved similarly to Theorem 1.2.6. To show (1.2.32) we use

$$\|u - u_n\| = \sup_{g\in H}\frac{|(g, u - u_n)|}{\|g\|}. \tag{1.2.34}$$

So setting $v = u - u_n$ in (1.2.33) yields

$$a(u - u_n, \phi) = (g, u - u_n).$$

By virtue of (1.2.31) we have

$$(g, u - u_n) = a(u - u_n, \phi - \tilde{\phi}), \ \forall\tilde{\phi} \in V_n.$$

Now, employ the boundedness and take the infimum for $\tilde{\phi} \in V_n$ to get

$$|(g, u - u_n)| \le M|u - u_n| \inf_{\tilde{\phi}\in V_n}|\phi - \tilde{\phi}|. \tag{1.2.35}$$

Finally, a combination of (1.2.34) and (1.2.35) leads to the desired result and completes the proof. □

### 1.2.6 Generalized Galerkin methods

Let $H$ be a separable real Hilbert space equipped with an inner product $(\cdot, \cdot)$ and the corresponding norm $\|\cdot\|$. $U$ and $V$ are dense linear subsets supplied with new inner products $[\cdot, \cdot]_U$, $[\cdot, \cdot]_V$ and related norms $|\cdot|_U$, $|\cdot|_V$, respectively. $U$ and $V$ respectively become Hilbert spaces under these new inner products. We also assume that the imbeddings of $U$ and $V$ in $H$ are continuous.

Suppose $A$ is a linear operator from a linear dense subset $D(A)$ of $U$ to $V'$. For $f \in V'$ let us consider the operator equation:

$$Au = f. \tag{1.2.36}$$

Assume the bilinear form $a(u, v) = \langle Au, v\rangle$ on $D(A) \times V$ satisfies

$$|a(u, v)| \leq M|u|_U|v|_V, \ \forall u \in D(A), v \in V. \tag{1.2.37}$$

Then as in Proposition 1.2.1 we may continuously extend $a(u, v)$ onto $U \times V$ such that the above estimate is still valid for any $(u, v) \in U \times V$. This results in a variational form of (1.2.36): Find $u \in U$ such that

$$a(u, v) = \langle f, v\rangle, \ \forall v \in V. \tag{1.2.38}$$

It is equivalent to a bounded linear operator equation

$$Tu = Rf, \tag{1.2.38$'$}$$

where $T : U \to V$ and $R : V' \to V$ are determined respectively by the Riesz representation theorem:

$$a(u, v) = [Tu, v]_V, \ u \in U, v \in V,$$

$$\langle f, v\rangle = [Rf, v], \ f \in V', v \in V.$$

It is obvious that $\|T\| \leq M$.

The separability of the spaces enables us to choose two families of finite dimensional subspaces $\{U_n\}$ and $\{V_n\}$ of $U$ and $V$ respectively such that

$$U_n \subset U_{n+1}, \ V_n \subset V_{n+1}, \ n = 1, 2, \cdots; \ \overline{\bigcup_{n=1}^{\infty} U_n} = U, \ \overline{\bigcup_{n=1}^{\infty} V_n} = V.$$

$U_n$ is referred to as the trial function space, and $V_n$ the test function space. The generalized Galerkin method for (1.2.38) is: Find $u_n \in U_n$ such that

$$a(u_n, v_n) = \langle f, v_n\rangle, \ \forall v_n \in V_n, \tag{1.2.39}$$

or equivalently

$$T_n u_n = Q_n g, \tag{1.2.39$'$}$$

where $T_n = Q_n T|_{U_n}$, $g = Rf$, and $Q_n$ is the orthogonal projection operator from $V$ to $V_n$.

**Theorem 1.2.8** *Let $U$ and $V$ be separable real Hilbert spaces, $U_n$ and $V_n$ their subspaces respectively, and $a(\cdot,\cdot)$ the bilinear form defined on $U \times V$ satisfying*

$$|a(u,v)| \leq M|u|_U|v|_V, \quad \forall u \in U, v \in V, \tag{1.2.40}$$

$$\inf_{\substack{u_n \in U_n \\ |u_n|_U=1}} \sup_{\substack{v_n \in V_n \\ |v_n|_V=1}} |a(u_n,v_n)| \geq \alpha > 0, \tag{1.2.41}$$

*where $M$ and $\alpha$ are constants. Then, (1.2.38) and (1.2.39) possess unique solutions $u$ and $u_n$ respectively, and the following error estimate holds:*

$$|u-u_n|_U \leq \left(1+\frac{M}{\alpha}\right) \inf_{w \in U_n} |u-w|_U. \tag{1.2.42}$$

**Proof** Note that

$$|T_n u_n|_V = \sup_{\substack{v \in V \\ |v|_V=1}} |[T_n u_n, v]| \geq \sup_{\substack{v_n \in V_n \\ |v_n|_V=1}} |[Tu_n, v_n]|.$$

So it follows from (1.2.41) that

$$|T_n u_n|_V \geq \alpha |u_n|_U, \ \forall u_n \in U_n.$$

This together with Theorem 1.2.5 gives the desired result and completes the proof. □

The estimate (1.2.42) in the above theorem indicates that the convergence order of $u - u_n$ is determined by the trial function space $U_n$, while the test function space $V_n$ only influences the constant in the right-hand side of the estimate. This motivates us to speed up the convergence by choosing the trial function spaces with better approximate properties, and to simplify the approximation scheme by choosing simple and flexible test function spaces. It is this idea that the generalized difference method discussed in this book is based on.

Theorem 1.2.8 is difficult to apply in practice. So we shall modify the above framework to suit the need for the numerical analysis for the generalized difference method.

Let $H$ be a separable real Hilbert space supplied with an inner product $(\cdot,\cdot)$ and the related norm $\|\cdot\|$. $U$ is a dense linear subset of $H$, and is a Hilbert space with an inner product $[\cdot,\cdot]$ and the related norm $|\cdot|$. $a(\cdot,\cdot)$ is a bounded and positive definite bilinear form on $U \times U$. For $f \in H$, consider the equation

$$a(u,v) = (f,v), \quad \forall v \in U. \tag{1.2.43}$$

Set $D = \{u \in U$ : the linear functional $a(u,\cdot)$ is continuous on $U$ with respect to the topology induced by $H\}$. In other words, $D$ is such a subset of $U$ that for each element $u$ of $D$ there exists a constant $M(u) > 0$ such that

$$|a(u,v)| \leq M(u)\|v\|, \; \forall v \in U.$$

The density of $U$ in $H$ enables us to continuously extend $a(u,\cdot)$ to a bounded linear functional on $H$, and by the Riesz representation theorem there is a unique $Au \in H$ such that

$$a(u,v) = (Au,v), \; u \in D, v \in H. \tag{1.2.44}$$

Obviously $A$ is a linear operator from $D$ to $H$.

Choose a trial function space $U_h \in U$ and a test function space $V_h \in H$ with dimensions $\dim U_h = \dim V_h = N$, where $h$ is a parameter. Construct a discrete bilinear form $a_h(\cdot,\cdot)$ defined on $U_h \times V_h$ satisfying

$$a_h(u,v_h) = (Au,v_h), \; \forall u \in D, v_h \in V_h. \tag{1.2.45}$$

The approximation scheme is: Find $u_h \in U_h$ such that

$$a_h(u_h,v_h) = (f,v_h), \; \forall v_h \in V_h. \tag{1.2.46}$$

Let $\Gamma_h$ be a linear operator from $U$ to $V_n$ satisfying $\Gamma_h U_h = V_h$. Then (1.2.46) is equivalent to

$$a_h(u_h,\Gamma_h w_h) = (f,\Gamma_h w_h), \; \forall w_h \in U_h. \tag{1.2.46$'$}$$

**Theorem 1.2.9** *Suppose $a_h(\cdot, \Gamma_h \cdot)$ is uniformly positive definite in the following sense: There exists a constant $\alpha > 0$ independent of the subspace $U_h$ such that*

$$a_h(w_h, \Gamma_h w_h) \geq \alpha |w_h|^2, \ \forall w_h \in U_h. \tag{1.2.47}$$

*Then (1.2.46) has a unique solution $u_h \in U_h$. If in addition the solution to (1.2.43) belongs to D, then we have the following error estimate:*

$$|u - u_h| \leq \inf_{\tilde{u} \in U_h} \left\{ |u - \tilde{u}| + \frac{1}{\alpha} \sup_{w_h \in U_h} \frac{|a_h(u - \tilde{u}, \Gamma_h w_h)|}{|w_h|} \right\}. \tag{1.2.48}$$

**Proof** Consider the homogeneous equation related to (1.2.46)

$$a_h(u_h, v_h) = 0, \ \forall v_h \in V_h.$$

Then we have $u_h = 0$ by setting $v_h = \Gamma_h u_h$ and using the uniformly positive definite condition (1.2.47). Thus the homogeneous equation admits only the trivial solution and consequently (1.2.46) has a unique solution.

Now let $u \in D$ and $u_h \in U_h$ be the solutions to (1.2.43) and (1.2.46) respectively. By (1.2.43)-(1.2.45) we know that

$$a_h(u, v_h) = (f, v_h), \ \forall v_h \in V_h. \tag{1.2.49}$$

Subtracting (1.2.46) from (1.2.49) yields an error equation

$$a_h(u - u_h, v_h) = 0, \ \forall v_h \in V_h. \tag{1.2.50}$$

For any $\tilde{u} \in U_h$, by (1.2.47) and (1.2.50) we have

$$\alpha |\tilde{u} - u_h|^2 \leq a_h(\tilde{u} - u_h, \Gamma_h(\tilde{u} - u_h)) = a_h(\tilde{u} - u, \Gamma_h(\tilde{u} - u_h)).$$

Hence

$$|\tilde{u} - u_h| \leq \frac{1}{\alpha} \sup_{w_h \in U_h} \frac{|a_h(\tilde{u} - u, \Gamma_h w_h)|}{|w_h|}.$$

Combine this equality with the triangular inequality

$$|u - u_h| \leq |u - \tilde{u}| + |\tilde{u} - u_h|,$$

and take the infimum with respect to $\tilde{u} \in U_h$, then we obtain (1.2.48). This completes the proof. □

## Bibliography and Comments

For the convenience of later use and the reader's reference, we provide in this chapter an outline of the Sobolev spaces and some basic results on their interpolation theories, variational problems and approximation methods. Most of the materials can be found in [A-27,26,19,2]. For a more systematical understanding of related topics, we refer to [B-1] for Sobolev space theory, [B-17] for the finite element method and the interpolation theory of Sobolev spaces, [A-26,3] and [B-47] for the generalized Galerkin method and the projection method. An original form of the general framework (Theorem 1.2.9) for the theory of generalized difference methods has been given in [A-30,53].

# Chapter 2

# TWO POINT BOUNDARY VALUE PROBLEMS

In this chapter we first illustrate the basic ideas of the generalized difference method by applying it to a second order ODE , i.e., a two point boundary value problem. It is shown how the generalized difference schemes are derived from the generalized Galerkin variational form. Then we present several examples of generalized difference schemes, and discuss their existence, uniqueness and convergence. Finally a fourth order problem is considered.

## 2.1 Basic Ideas of the Generalized Difference Method

### 2.1.1 A variational form

Consider the boundary value problem of the second order ODE on an interval $I = [a, b]$

$$\text{(P1)}\quad \begin{cases} Lu \equiv -\dfrac{\mathrm{d}}{\mathrm{d}x}\Big(p\dfrac{\mathrm{d}u}{\mathrm{d}x}\Big)+r\dfrac{\mathrm{d}u}{\mathrm{d}x}+qu = f,\ x\in(a,b), & \text{(2.1.1a)}\\ u(a)=0,\ u'(b)=0, & \text{(2.1.1b)} \end{cases}$$

where $p \in C^1(I)$, $p(x) \geq p_{\min} > 0$, and $r, q, f \in C(I)$.

Let $u \in C^1[a,b] \cap C^2(a,b)$ be the solution of (P1), and $H_E^1(I) = \{v \in H^1(I) : \ v(a) = 0\}$. Use any function $v \in H_E^1(I)$ (called a test function) to multiply (2.1.1a) and integrate it on [a,b], then we have

$$\int_a^b Luv\mathrm{d}x = \int_a^b fv\mathrm{d}x.$$

By integrating by parts and the boundary condition (2.1.1b) we have

$$\begin{aligned}\int_a^b Luv\mathrm{d}x &= \int_a^b (pu'v' + ru'v + quv)\mathrm{d}x - pu'v|_a^b \\ &= \int_a^b (pu'v' + ru'v + quv)\mathrm{d}x.\end{aligned}$$

Write

$$a(u,v) = \int_a^b (pu'v' + ru'v + quv)\mathrm{d}x \tag{2.1.2}$$

and denote by $(\cdot,\cdot)$ the inner product of $L^2$, then we find that $u$ is the solution of the following problem:

$$(\mathrm{P2}) \begin{cases} \text{Find } u \in H_E^1(I) \text{ such that} \\ a(u,v) - (f,v) = 0, \ \forall v \in H_E^1(I). \end{cases} \tag{2.1.3}$$

On the other hand, if $u$ is a solution of (P2) and $u \in C^1[a,b] \cap C^2(a,b)$, then integrating (2.1.3) by parts leads to

$$\int_a^b [-(pu')' + ru' + qu - f]v\mathrm{d}x + p(b)u'(b)v(b) = 0, \ \forall v \in H_E^1(I). \tag{2.1.4}$$

In particular

$$\int_a^b (Lu - f)v\mathrm{d}x = 0, \ \forall v \in C_0^\infty(I).$$

So $u$ satisfies (2.1.1a) by the fundamental lemma of variational methods (Theorem 1.1.3). Now using (2.1.4) again we find that

$$p(b)u'(b)v(b) = 0, \ \forall v \in H_E^1(I).$$

Hence, setting $v(b) \neq 0$ shows $u'(b) = 0$. Therefore $u$ satisfies (2.1.1b) as well and is the solution of (P1).

To sum up we have the following theorem.

**Theorem 2.1.1** (Variational principle) *Suppose that $u \in C^1[a,b] \cap C^2(a,b)$ is the solution of* (P1)*, then it is the solution of* (P2)*. Conversely, if $u$ is the solution of* (P2) *and $u \in C^1[a,b] \cap C^2(a,b)$, then it is the solution of* (P1)*.*

(P1) is a differential form and (P2) is its Galerkin variational form. Solutions of (P1) are called classical solutions, while those of (P2) generalized solutions or weak solutions. In mechanics, the left hand side of (2.1.3) represents virtual work, and hence Theorem 2.1.1 is also referred to as a virtual work principle. There are significant differences between the two boundary conditions in (2.1.1b). The right boundary condition $u'(b) = 0$ needs not to be satisfied by the functions in $H_E^1(I)$. But it will be satisfied *naturally* by the solution of the variational problem. Therefore, it is called a natural boundary condition. In mechanics, this boundary condition corresponds to the force. On the contrary, the left boundary condition $u(a) = 0$ must be imposed on $H_E^1(I)$. Hence it is called an essential boundary condition. It is a geometrical condition. Variational principles are commonly used to describe physical phenomena and also lay the foundation of the numerical methods for differential equations. Compared with the differential forms, the merit of the variational problems is that, for instance, the second derivative of $u$ is not involved in the problem (P2), and the natural boundary conditions are much easier to deal with.

### 2.1.2 Galerkin methods

It is usually difficult to directly solve the variational forms to get the precise solutions. The main trouble lies in the fact that $H_E^1(I)$ is an infinite dimensional space. The idea to overcome this is to approximate infinite dimensional spaces by finite ones.

Let us choose a finite dimensional subspace $U_h$ in $U = H_E^1(I)$, and use it to replace $U$ in (P2) to obtain an approximate problem:

$$(\text{P2})_h \begin{cases} \text{Find } u_h \in U_h \text{ such that} \\ a(u_h, v_h) = (f, v_h), \ \forall v_h \in U_h. \end{cases} \tag{2.1.5}$$

This is precisely the so-called Galerkin method, or the variational method, and $u_h$ is the Galerkin approximation of $u$.

Galerkin methods in the early stages use smooth functions (usually algebraic or triangular polynomials, or special functions related to certain specific problems) to construct the finite dimensional space $U_h$. There are some disadvantages to follow this approach in practice. Mainly they are: the difficulties in constructing globally defined polynomials to satisfy the boundary conditions for multidimensional irregular regions; the big computing work for calculating the integrals to form the Galerkin equation; and the non-sparseness and the large condition number of the coefficient matrix of the Galerkin equation. Therefore the classical Galerkin methods cannot match finite difference methods which possess, on the contrary, advantages such as sparse coefficient matrices, less computing work and simple programming. The finite element method, initiated by R. Courant (1943) and developed in the fifties, provides a new approach for Galerkin methods to construct the subspace $U_h$. It decomposes the solution region into a network like the difference method, and utilizes spline functions to construct the subspace $U_h$, which contains low order piecewise polynomials satisfying the essential boundary conditions as well as certain global smoothness. Such kinds of Galerkin methods are called finite element methods or Galerkin finite element methods, and $U_h$'s the finite element spaces. The coefficient matrices of the finite element method are sparse, and their computation is simple and flexible. It is particularly powerful in dealing with irregular regions and natural boundary conditions. Therefore, the finite element method is an effective numerical method for elliptic and parabolic equations.

### 2.1.3 Generalized Galerkin variational principles

Our generalized difference methods are based on the traditional difference methods and have absorbed certain ideas of finite element methods (mainly the variational forms and the finite element spaces). We will see that the generalized difference methods enjoy the advantages of both the difference methods and the finite element methods. To set up the generalized difference method and its theoretical foundation, we need to design a new variational principle of generalized Galerkin type. In this chapter we only take a one-dimensional problem as an example to illustrate the idea.

Let us discretize the interval $I = [a, b]$ into a set of points (or nodes)

$$a = x_0 < x_1 < x_2 < \cdots < x_n = b.$$

The subintervals $I_i = [x_{i-1}, x_i]$ are called elements. All these elements compose a discretization of $I$, denoted by $\sigma = \{I_i : 1 \le i \le n\}$. Denote by $I_i^0$ the interior of $I_i$, and $\mathcal{P}_k$ the set of all the polynomials of degrees less than or equal to $k$. Write

$$S_\sigma^{(k)}(I) = \{v \in L^2(I) : v|_{I_i^0} \in \mathcal{P}_k,\ i = 1, 2, \cdots, n\}$$

and call it the set of piecewise polynomials of degree $k$ with respect to $\sigma$. Similarly call

$$S^{(k)}(I) = \bigcup_\sigma S_\sigma^{(k)}(I)$$

the piecewise polynomials of degree $k$ on $I$. In particular, it is called the set of piecewise constant (or step) and piecewise linear functions when $k = 0$ and $k = 1$ respectively. We also write

$$S_{\sigma,E}^{(k)}(I) = \{v \in S_\sigma^{(k)}(I) : v(a^+) = 0\},$$

$$S_E^{(k)}(I) = \bigcup_\sigma S_{\sigma,E}^{(k)}(I).$$

Use any $v \in V = S_E^{(k)}(I)$ to multiply equation (2.1.1a), and integrate it on $I$. Then by integrating by parts we have

$$(Lu, v) = \sum_{i=1}^n \int_{x_{i-1}}^{x_i} (pu'v' + ru'v + quv)\mathrm{d}x - \sum_{i=1}^n pu'v\Big|_{x_{i-1}^+}^{x_i^-}. \quad (2.1.6)$$

Noticing $v(a^+) = 0$ and $u'(b) = 0$, we know that $u$ satisfies

$$a_\sigma(u, v) = (f, v),\ \forall v \in V, \quad (2.1.7)$$

where

$$a_\sigma(u,v) = \sum_{i=1}^{n}\int_{x_{i-1}}^{x_i}(pu'v' + ru'v + quv)\mathrm{d}x + \sum_{i=1}^{n-1} p(x_i)u'(x_i)[v(x_i^+) - v(x_i^-)]. \tag{2.1.8}$$

Let us introduce a generalized function $\delta(x)$ defined as the derivative of the following jump function:

$$\sigma(x) = \begin{cases} 0, & \text{for } x < 0, \\ 1, & \text{for } x > 0. \end{cases}$$

So formally we have

$$\delta(x) = \sigma'(x) = \begin{cases} 0, & \text{for } x \neq 0, \\ \infty, & \text{for } x = 0. \end{cases}$$

For any smooth function $g(x)$ we have

$$\int_\alpha^\beta g(x)\delta(x)\mathrm{d}x = g(0), \ \alpha < 0 < \beta.$$

The piecewise polynomial function $v \in V$ mentioned above can be expressed as the sum of a continuous function $v_1$ and a step function $v_2$:

$$v = v_1 + v_2,$$

$$v_2 = \sum_{i=1}^{n-1}[v(x_i^+) - v(x_i^-)]\sigma(x - x_i).$$

So if $u'$ is continuous or $u \in H^2(I)$, then in the sense of generalized functions we have

$$\begin{aligned} a(u,v) &\equiv \int_a^b (pu'v' + ru'v + quv)\mathrm{d}x \\ &= \sum_{i=1}^{n}\int_{x_{i-1}}^{x_i}(pu'v' + ru'v + quv)\mathrm{d}x \\ &\quad + \sum_{i=1}^{n-1} p(x_i)u'(x_i)[v(x_i^+) - v(x_i^-)] \\ &= a_\sigma(u,v). \end{aligned} \tag{2.1.9}$$

Now, it follows from (2.1.7) and (2.1.9) that the solution $u$ of (P1) also solves the following problem

$$\text{(P3)} \quad \begin{cases} \text{Find } u \in H_E^1(I) \cap H^2(I) \text{ such that} \\ a(u,v) - (f,v) = 0, \ \forall v \in V. \end{cases} \tag{2.1.10}$$

On the other hand, if $u$ is a solution of (P3) and $u \in C^1[a,b] \cap C^2(a,b)$, then it follows from (2.1.6) and (2.1.7) that

$$(Lu - f, v) + p(b)u'(b)v(b^-) = 0, \ \forall v \in V. \tag{2.1.11}$$

In particular

$$(Lu - f, v) = 0, \ \forall v \in \{v \in V : v(b^-) = 0\}.$$

Since the set $\{v \in V : v(b^-) = 0\}$ is dense in $L^2(I)$, the above equation is valid for any $v \in L^2(I)$. This verifies (2.1.1a). Now (2.1.11) becomes

$$p(b)u'(b)v(b^-) = 0, \ \forall v \in V.$$

Taking $v \in V$ with $v(b^-) \neq 0$ implies $u'(b) = 0$. Finally, we note $u \in H_E^1(I)$, so (2.1.1b) is valid and $u$ is the solution of (P1).

The above discussion leads to the following theorem.

**Theorem 2.1.2** *Suppose that $u \in C^1[a,b] \cap C^2(a,b)$ is the solution of (P1); then it is the solution of (P3). Conversely, if $u$ is the solution of (P3) and $u \in C^1[a,b] \cap C^2(a,b)$, then it is the solution of (P1).*

We shall call (P3) the generalized Galerkin variational problem with respect to (P1), and Theorem 2.1.2 the generalized Galerkin variational principle.

Now we make a convention that in the sequel the bilinear form $a(u,v)$ is understood according to (2.1.9), which coincides with the original definition when $v \in H_E^1(I)$, and that (2.1.2) should be interpreted in terms of generalized functions when $v \in S_E^{(k)}(I)$.

### 2.1.4 Generalized difference methods

Let us decompose the interval $I = [a, b]$ into a grid $T_h$ with nodes

$$a = x_0 < x_1 < x_2 < \cdots < x_n = b.$$

Accordingly we place a dual grid

$$a = x_0 < x_{1/2} < x_{3/2} < \cdots < x_{n-1/2} < x_n = b,$$

where $x_{i-1/2} = (x_{i-1} + x_i)/2$, $i = 1, 2, \cdots, n$. Write $I_0^* = [x_0, x_{1/2}]$, $I_i^* = [x_{i-1/2}, x_{i+1/2}]$ $(i = 1, 2, \cdots, n-1)$ and $I_n^* = [x_{n-1/2}, x_n]$. Then, these dual elements $I_i^*$ $(i = 0, 1, 2, \cdots, n)$ lead to a dual grid, written as $T_h^*$.

Choose the trial function space $U_h \subset U = H_E^1(I)$ as a finite element space with respect to $T_h$, and the test function space $V_h \subset V = S_E^{(k)}(I)$ as a piecewise polynomial (of low order) space $S_{\sigma^*,E}^{(k)}$ with respect to the dual discretization $\sigma^*$ induced by $T_h^*$. Now we propose an approximation problem of (P3)

$$(\mathrm{P3})_h \begin{cases} \text{Find } u_h \in U_h \text{ such that} \\ a(u_h, v_h) = (f, v_h), \ \forall v_h \in V_h. \end{cases} \tag{2.1.12}$$

Different choices of $U_h$ and $V_h$ lead to different schemes. In particular, selecting $U_h$ and $V_h$ as piecewise linear and constant functions respectively yields the usual difference scheme, as we shall see in the next section. This explains why we call $(\mathrm{P3})_h$ the generalized difference method.

Let us elaborate on the general considerations in constructing the test function space $V_h$. The choice of $V_h$ certainly should be somehow related to $U_h$. For instance, $V_h$ should have the same degree of freedom as $U_h$. But it is not required for the functions in $V_h$ to possess global continuities, so we can choose $V_h$ as low order piecewise polynomial spaces to reduce the computation effort. Usually when the values $\mathrm{d}^j u(x_i)/\mathrm{d}x^j$ $(j \geq 0)$ make sense at the node $x_i$ for the functions in $U_h$, one can choose the basis function $\psi_i^{(j)}$ of $V_h$ with respect to the point $x_i$ to satisfy the following conditions:

(i) The support of $\psi_i^{(j)}$ belongs to the dual element $I_i^*$ containing $x_i$.

(ii) On $I_i^*$

$$\frac{\mathrm{d}^l}{\mathrm{d}x^l}\psi_i^{(j)}(x)\Big|_{x=x_i} = \begin{cases} 0, & \text{for } l \neq j, \\ 1, & \text{for } l = j. \end{cases}$$

These conditions imply that

$$\psi_i^{(j)}(x) = \begin{cases} (x - x_i)^j/j!, & x \in I_i^*, \\ 0, & x \notin I_i^*. \end{cases}$$

Finally, we point out that the generalized difference methods are different from the usual generalized Galerkin finite element methods or the nonstandard finite element methods, in that we only have $V_h \in L^2$ rather than $V_h \in H^1(I)$ and the corresponding bilinear form has to be understood in the sense of generalized functions. So the variational forms of the generalized difference methods differ essentially from the usual ones, causing difficulties in theoretical analysis.

## 2.2 Linear Element Difference Schemes

Consider the two point boundary value problem

$$\begin{cases} Lu \equiv -\dfrac{\mathrm{d}}{\mathrm{d}x}\Big(p\dfrac{\mathrm{d}u}{\mathrm{d}x}\Big) = f, \ x \in (a, b), & (2.2.1a) \\ u(a) = 0, \ u'(b) = 0, & (2.2.1b) \end{cases}$$

where $p \in C^1(I)$, $p(x) \geq p_{\min} > 0$, and $f \in L^2(I)$.

In this section we deduce a linear element difference scheme by choosing the trial and the test function spaces as linear finite element and piecewise constant function spaces respectively. This will result in the usual difference scheme as we have promised.

### 2.2.1 Trial and test function spaces

Discretize the interval $I = [a, b]$ into a grid $T_h$ with nodes

$$a = x_0 < x_1 < \cdots < x_n = b.$$

Denote the length of the element $I_i$ by $h_i = x_i - x_{i-1}$ and write $h = \max_{1\le i\le n} h_i$. We assume the grid satisfies the quasi-uniform condition $h_i \geq \mu h$ $(i = 1, 2, \cdots, n)$ for some positive constant $\mu$.

The trial space $U_h$ is taken as the linear element space with respect to $T_h$, which consists of all the functions $u_h$ satisfying

(i) $u_h \in C(I)$, $u_h(a) = 0$ and

(ii) $u_h$ is linear on each $I_i$ and is determined uniquely by its values at the endpoints of the element.

Obviously $U_h$ is an $n$-dimensional subspace of $H_E^1(I)$.

To construct the nodal basis functions we consider an interpolation problem on the reference element [0,1]: Find a linear function $N_0(\xi)$ such that

$$N_0(0) = 1,\ N_0(1) = 0.$$

Then, $N_0(\xi) = 1 - \xi$. Notice that the affine mapping

$$\xi = \frac{x_i - x}{h_i}$$

maps the left interval $I_i = [x_{i-1}, x_i]$ of the node $x_i$ onto the reference element [0,1] with $x_i \to 0$, $x_{i-1} \to 1$. Thus, on the interval $[x_{i-1}, x_i]$ the basis function $\phi_i$ of the node $x_i$ is of the form

$$\phi_i(x) = 1 - \frac{x_i - x}{h_i}.$$

Similarly the affine mapping

$$\xi = \frac{x - x_i}{h_{i+1}}$$

maps the right interval $I_i = [x_i, x_{i+1}]$ of the node $x_i$ onto the reference element [0,1] with $x_i \to 0$, $x_{i+1} \to 1$. Thus, on the interval $[x_i, x_{i+1}]$ $\phi_i$ can be expressed as

$$\phi_i(x) = 1 - \frac{x - x_i}{h_{i+1}}.$$

Therefore, the basis function with respect to $x_i$ is

$$\phi_i(x) = \begin{cases} 1 - h_i^{-1}|x - x_i|, & x_{i-1} \le x \le x_i, \\ 1 - h_{i+1}^{-1}|x - x_i|, & x_i \le x \le x_{i+1}, \\ 0, & \text{elsewhere}. \end{cases} \tag{2.2.2}$$

The functions $\{\phi_i(x) : i = 1, 2, \cdots, n\}$ form a basis of $U_h$ and any $u_h \in U_h$ has the following expression

$$u_h = \sum_{i=1}^{n} u_i \phi_i(x),$$

where $u_i = u_h(x_i)$. On the element $I_i$ we have

$$u_h = u_{i-1}(1-\xi) + u_i \xi, \ \left(\xi = \frac{x - x_{i-1}}{h_i}\right) \tag{2.2.3}$$

$$u_h' = (u_i - u_{i-1})/h_i, \ x \in I_i, \ i = 1, 2, \cdots, n. \tag{2.2.4}$$

Next we place a dual grid $T_h^*$ with nodes

$$a = x_0 < x_{1/2} < x_{3/2} < \cdots < x_{n-1/2} < x_n = b,$$

where $x_{i-1/2} = (x_{i-1} + x_i)/2$, $i = 1, 2, \cdots, n$. The dual elements are $I_0^* = [x_0, x_{1/2}]$, $I_i^* = [x_{i-1/2}, x_{i+1/2}]$ $(i = 1, 2, \cdots, n-1)$ and $I_n^* = [x_{n-1/2}, x_n]$. Accordingly we choose the test function space $V_h$ as the piecewise constant function (step function) space, which contains all the functions $v_h \in L^2(I)$ satisfying

(i) $v_h(x) = 0$, for $x \in I_0^*$ and

(ii) $v_h$ is a constant on each $I_i^*$ $(i = 1, 2, \cdots, n)$.

The basis functions of $V_h$ are

$$\psi_j(x) = \begin{cases} 1, & x \in I_j^*, \\ 0, & x \notin I_j^*, \end{cases} \quad j = 1, 2, \cdots, n. \tag{2.2.5}$$

Any $v_h \in V_h$ has the form

$$v_h = \sum_{i=1}^{n} v_i \psi_i(x),$$

where $v_i = v_h(x_i)$.

Typical basis functions of $U_h$ and $V_h$ are depicted in Fig. 2.2.1.

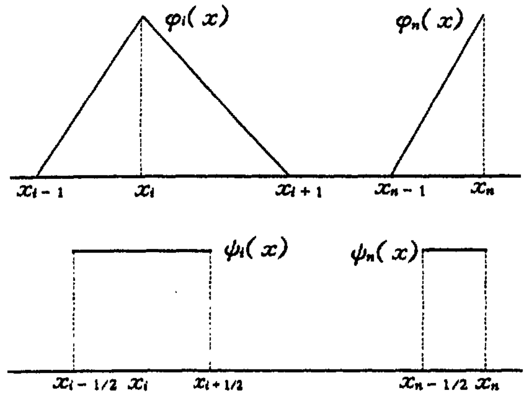



**Fig. 2.2.1**

### 2.2.2 Difference equations

Following $(\mathrm{P3})_h$ in Subsection 2.1.4, the linear element difference scheme is: Find $u_h = \sum_{i=1}^{n} u_i \phi_i(x)$ such that

$$a(u_h, \psi_j) = (f, \psi_j), \ j = 1, 2, \cdots, n, \tag{2.2.6}$$

where (cf. (2.1.9))

$$\begin{aligned}
a(u_h, \psi_j) &= \int_a^b p u_h' [\delta(x - x_{j-1/2}) - \delta(x - x_{j+1/2})] \mathrm{d}x \\
&= p_{j-1/2} u_h'(x_{j-1/2}) - p_{j+1/2} u_h'(x_{j+1/2}) \\
&= p_{j-1/2}(u_j - u_{j-1})/h_j - p_{j+1/2}(u_{j+1} - u_j)/h_{j+1}, \\
&\quad j = 1, 2, \cdots, n-1, \\
&\quad u_0 = 0, \\
a(u_h, \psi_n) &= p_{n-1/2}(u_n - u_{n-1})/h_n.
\end{aligned}$$

This results in a difference equation

$$\begin{cases} p_{j-1/2}(u_j - u_{j-1})/h_j - p_{j+1/2}(u_{j+1} - u_j)/h_{j+1} = \int_{x_{j-1/2}}^{x_{j+1/2}} f\mathrm{d}x, \\ \quad j = 1, 2, \cdots, n-1, \quad u_0 = 0, \\ p_{n-1/2}(u_n - u_{n-1})/h_n = \int_{x_{n-1/2}}^{x_n} f\mathrm{d}x. \end{cases} \tag{2.2.7}$$

The left-hand side of this equation coincides with the usual finite difference method. Furthermore, if we use $f_j = f(x_j)$ and $f_n = f(x_n)$ respectively to approximate the integrals in (2.2.7), then we end up with precisely the usual finite difference scheme. Therefore, we see that a finite difference scheme has been derived from the generalized difference method.

(2.2.7) is a linear system for the unknowns $u_1, u_2, \cdots, u_n$ with a symmetric tridiagonal coefficient matrix:

$$\begin{bmatrix} \frac{p_{\frac{1}{2}}}{h_1} + \frac{p_{\frac{3}{2}}}{h_2} & -\frac{p_{\frac{3}{2}}}{h_2} & & & \\ -\frac{p_{\frac{3}{2}}}{h_2} & \frac{p_{\frac{3}{2}}}{h_2} + \frac{p_{\frac{5}{2}}}{h_3} & -\frac{p_{\frac{5}{2}}}{h_3} & & \\ & -\frac{p_{\frac{5}{2}}}{h_3} & \frac{p_{\frac{5}{2}}}{h_3} + \frac{p_{\frac{7}{2}}}{h_4} & -\frac{p_{\frac{7}{2}}}{h_4} & \\ & & \ddots & \ddots & \ddots \\ & & -\frac{p_{n-\frac{3}{2}}}{h_{n-1}} & \frac{p_{n-\frac{3}{2}}}{h_{n-1}} + \frac{p_{n-\frac{1}{2}}}{h_n} & -\frac{p_{n-\frac{1}{2}}}{h_n} \\ & & & -\frac{p_{n-\frac{1}{2}}}{h_n} & \frac{p_{n-\frac{1}{2}}}{h_n} \end{bmatrix}$$

### 2.2.3 Convergence estimates

There have been thorough discussions for the convergence of the finite difference scheme corresponding to (2.2.6) (or (2.2.7)). Here we consider the error estimates of (2.2.6) in a Sobolev norm.

First, for $u_h \in U_h$ we have by (2.2.4) that

$$|u_h|_1 = \Big[\int_a^b (u_h')^2 \mathrm{d}x\Big]^{1/2} = \Big[\sum_{i=1}^n (u_i - u_{i-1})^2/h_i\Big]^{1/2}. \qquad (2.2.8)$$

Next, we define an interpolation operator $\Pi_h^* : U \to V_h$ by

$$\Pi_h^* u = \sum_{j=1}^n u_j \psi_j, \ \forall u \in U.$$

Finally we examine the positive definiteness of $a(u_h, \Pi_h^* u_h)$.

$$\begin{aligned} a(u_h, \Pi_h^* u_h) &= \sum_{j=1}^n u_j a(u_h, \psi_j) \\ &= \sum_{j=1}^{n-1} u_j [p_{j-1/2}(u_j - u_{j-1})/h_j - p_{j+1/2}(u_{j+1} - u_j)/h_{j+1}] \\ &\quad + u_n p_{n-1/2}(u_n - u_{n-1})/h_n \\ &= \sum_{j=1}^n p_{j-1/2}(u_j - u_{j-1})^2/h_j \geq p_{\min}|u_h|_1^2. \end{aligned}$$

This gives the following theorem.

**Theorem 2.2.1** *The discrete bilinear form $a(u_h, \Pi_h^* u_h)$ is positive definite, that is, there exists a constant $\alpha > 0$ such that*

$$a(u_h, \Pi_h^* u_h) \geq \alpha |u_h|_1^2, \ \forall u_h \in U_h. \qquad (2.2.9)$$

We notice that the seminorm $|\cdot|_1$ and the norm $\|\cdot\|_1$ are equivalent in the space $H_E^1(I)$. Thus the existence and uniqueness of the difference scheme (2.2.6) follows from Theorem 2.2.1. Now we turn to the convergence estimate.

**Theorem 2.2.2** *Let $u \in H^2(I)$ be the solution of the differential equation (2.2.1) and $u_h$ the solution of the difference scheme (2.2.6), then the following estimate holds:*

$$|u - u_h|_1 \leq Ch|u|_2. \qquad (2.2.10)$$

**Proof** It is obvious that

$$a(u,\psi_j)=(f,\psi_j),\ j=1,2,\cdots,n,$$

$$a(u_h,\psi_j)=(f,\psi_j),\ j=1,2,\cdots,n.$$

So

$$a(u-u_h,\psi_j)=0,\ j=1,2,\cdots,n. \tag{2.2.11}$$

Let $\Pi_h u$ be the interpolation projection of $u$ onto the trial function space $U_h$. Then, by Theorem 2.2.1 and (2.2.11) we get

$$\begin{aligned}|u_h-\Pi_h u|_1^2 &\le \frac{1}{\alpha}a(u_h-\Pi_h u,\Pi_h^*(u_h-\Pi_h u))\\ &= \frac{1}{\alpha}a(u-\Pi_h u,\Pi_h^*(u_h-\Pi_h u)).\end{aligned}$$

Thus

$$|u_h-\Pi_h u|_1 \le \frac{1}{\alpha}\sup_{w_h\in U_h}\frac{|a(u-\Pi_h u,\Pi_h^* w_h)|}{|w_h|_1}. \tag{2.2.12}$$

Write $w_j=w_h(x_j)$, then we have $\Pi_h^* w_h=\sum\limits_{j=1}^{n} w_j\psi_j$ and

$$\begin{aligned}a(u-\Pi_h u,\Pi_h^* w_h) &= \sum_{j=1}^{n} w_j a(u-\Pi_h u,\psi_j)\\ =& \sum_{j=1}^{n-1} w_j[p_{j-1/2}(u-\Pi_h u)'_{j-1/2}-p_{j+1/2}(u-\Pi_h u)'_{j+1/2}\\ &+w_n p_{n-1/2}(u-\Pi_h u)'_{n-1/2}\\ =& \sum_{j=1}^{n-1} p_{j-1/2}(u-\Pi_h u)'_{j-1/2}(w_j-w_{j-1}).\end{aligned}$$

By the Cauchy inequality we have

$$\begin{aligned}&|a(u-\Pi_h u,\Pi_h^* w_h)|\\ \le\ & C\Big\{\sum_{j=1}^{n}[(u-\Pi_h u)'_{j-1/2}]^2\Big\}^{1/2}\Big\{\sum_{j=1}^{n}(w_j-w_{j-1})^2\Big\}^{1/2}.\end{aligned} \tag{2.2.13}$$

By (2.2.4)

$$(u-\Pi_h u)'_{j-1/2} = u'_{j-1/2} - (u_j - u_{j-1})/h_j.$$

By the mean value theorem there exists $\xi_0 \in I_j$ such that

$$u'(\xi_0) = (u_j - u_{j-1})/h_i, \text{ i.e. } (u-\Pi_h u)'(\xi_0) = 0.$$

Hence,

$$(u-\Pi_h u)'_{j-1/2} = \int_{\xi_0}^{x_{j-1/2}} (u-\Pi_h u)'' \mathrm{d}x = \int_{\xi_0}^{x_{j-1/2}} u'' \mathrm{d}x,$$

which yields

$$|(u-\Pi_h u)'_{j-1/2}|^2 \le h\Big[\int_{x_{j-1}}^{x_j} (u'')^2 \mathrm{d}x\Big],$$

$$\Big\{\sum_{j=1}^{n} [(u-\Pi_h u)'_{j-1/2}]^2\Big\}^{1/2} \le h^{1/2}|u|_2. \tag{2.2.14}$$

Combining (2.2.13), (2.2.14) and (2.2.8) we get

$$|a(u-\Pi_h u, \Pi_h^* w_h)| \le Ch|u|_2|w_h|_1.$$

This together with (2.2.12) yields

$$|u_h - \Pi_h u|_1 \le Ch|u|_2.$$

By the interpolation theory in Sobolev spaces

$$|u-\Pi_h u|_1 \le Ch|u|_2.$$

These two estimates imply (2.2.10) and complete the proof. □

## 2.3 Quadratic Element Difference Schemes

Consider the two point boundary value problem:

$$\begin{cases} Lu \equiv -\dfrac{\mathrm{d}}{\mathrm{d}x}\Big(p\dfrac{\mathrm{d}u}{\mathrm{d}x}\Big) + qu = f,\ x \in (a,b), & (2.3.1a) \\ u(a) = 0,\ u'(b) = 0, & (2.3.1b) \end{cases}$$

where $p, q \in C^1(I)$, $p(x) \geq p_{\min} > 0$, $q \geq 0$ and $f \in L^2(I)$.

In this section we derive a quadratic element difference scheme by choosing the trial and the test spaces as the quadratic element space of Lagrangian type and the piecewise constant function space respectively.

### 2.3.1 Trial and test spaces

Perform the same discretization as in §2.2 to get the grid $T_h$. But now the interpolation points include, besides the nodal points $x_j$, the midpoints of the elements $x_{j-1/2} = (x_j + x_{j-1})/2$ $(j = 1, 2, \cdots, n)$ as well.

Select the trial function space $U_h$ as the quadratic element space of Lagrangian type with respect to $T_h$. So any function $u_h$ in $U_h$ satisfies the following conditions:

(i)$u_h \in C(I)$, $u_h(a) = 0$;

(ii)$u_h$ is a quadratic polynomial at each element $I_i$ and it is determined uniquely by its values at the two endpoints and the midpoint of the element.

Obviously $U_h$ is a $2n$-dimensional subspace of $U = H_E^1(I)$.

To obtain the basis functions, let us first construct the quadratic functions $N_0(\xi)$ and $N_{1/2}(\xi)$ such that

$$\begin{aligned} &N_0(0) = 1, && N_0(0.5) = N_0(1) = 0, \\ &N_{1/2}(0.5) = 1, && N_{1/2}(0) = N_{1/2}(1) = 0. \end{aligned}$$

It is an easy matter to get

$$N_0(\xi) = (2\xi - 1)(\xi - 1),$$
$$N_{1/2}(\xi) = 4\xi(1 - \xi).$$

Setting $\xi = (x - x_i)/h_{i+1}$ and $\xi = (x_i - x)/h_i$ respectively in $N_0(\xi)$, we end up with the right and left halves of the basis function with respect to the node $x_i$. So we have

$$\phi_i(x) = \begin{cases} (2|x - x_i|/h_i - 1)(|x - x_i|h_i - 1), \\ \qquad x_{i-1} \le x \le x_i, \\ (2|x - x_i|/h_{i+1} - 1)(|x - x_i|h_{i+1} - 1), \\ \qquad x_i \le x \le x_{i+1}, \\ 0, \quad \text{elsewhere.} \end{cases}$$

Similarly set $\xi = (x - x_{j-1})/h_i$ in $N_{1/2}(\xi)$ to get the basis function with respect to the node $x_{i-1/2}$

$$\phi_{i-1/2}(x) = \begin{cases} 4(1 - (x - x_{i-1})/h_i)(x - x_{i-1})/h_i, \\ \qquad x_{i-1} \le x \le x_i. \\ 0, \quad \text{elsewhere.} \end{cases}$$

The set $\{\phi_i(x), \phi_{i-1/2}(x); 1 \le i \le n\}$ is a basis of $U_h$ and any $u_h \in U_h$ can be uniquely written as

$$u_h = \sum_{i=1}^{n} [u_i\phi_i(x) + u_{i-1/2}\phi_{i-1/2}(x)],$$

where $u_i = u_h(x_i)$ and $u_{i-1/2} = u_h(x_{i-1/2})$. In the element $I_i = [x_{i-1}, x_i]$

$$\begin{aligned} u_h &= u_{i-1}(2\xi - 1)(\xi - 1) + 4u_{i-1/2}\xi(1 - \xi) + u_i(2\xi - 1)\xi \\ &= (\xi^2, \xi, 1) \begin{bmatrix} 2 & -4 & 2 \\ -3 & 4 & -1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} u_{i-1} \\ u_{i-1/2} \\ u_i \end{bmatrix}, \end{aligned} \tag{2.3.2}$$

$$\begin{aligned} u'_h &= u_{i-1}(4\xi - 3)/h_i + u_{i-1/2}(-8\xi + 4)/h_i + u_i(4\xi - 1)/h_i \\ &= (\xi, 1) \begin{bmatrix} -4 & 4 \\ 3 & -1 \end{bmatrix} \begin{bmatrix} (u_{i-1/2} - u_{i-1})/h_i \\ (u_i - u_{i-1/2})/h_i \end{bmatrix}, \end{aligned} \tag{2.3.3}$$

where $\xi = (x - x_{i-1})/h_i$.

Next, we place a dual grid $T_h^*$ with nodal points

$$a = x_0 < x_{1/4} < x_{3/4} < \cdots < x_{n-3/4} < x_{n-1/4} < x_n = b,$$

where $x_{i-k/4} = x_i - \frac{k}{4}h_i$ $(k = 1, 3,\ i = 1, 2, \cdots, n)$. The test function space corresponding to $T_h^*$ is taken as the piecewise constant function space, which is a $2n$-dimensional subspace spanned by the basis functions of the nodes $x_i$

$$\psi_i(x) = \begin{cases} 1, & x_{i-1/4} \le x \le x_{i+1/4}, \\ 0, & \text{elsewhere,} \end{cases}$$

and the ones of $x_{i-1/2}$

$$\psi_{i-1/2}(x) = \begin{cases} 1, & x_{i-3/4} \le x \le x_{i-1/4}, \\ 0, & \text{elsewhere.} \end{cases}$$

Any $v_h \in V_h$ can be uniquely expressed as

$$v_h = \sum_{j=1}^{n} [v_j \psi_j(x) + v_{j-1/2}\psi_{j-1/2}(x)].$$

Typical basis functions $\phi_i$ of $U_h$ and $\psi_i$ of $V_h$ are depicted in Fig. 2.3.1.

### 2.3.2 Difference equations

The quadratic difference scheme corresponding to the subspaces $U_h$ and $V_h$ given in the last subsection is: Find $u_h \in U_h$ such that

$$\begin{cases} a(u_h, \psi_j) = (f, \psi_j), & j = 1, 2, \cdots, n, \\ a(u_h, \psi_{j-1/2}) = (f, \psi_{j-1/2}), & j = 1, 2, \cdots, n, \end{cases} \tag{2.3.4}$$

where

$$\begin{aligned} a(u_h, \psi_j) &= p_{j-1/4}u_h'(x_{j-1/4}) - p_{j+1/4}u_h'(x_{j+1/4}) + \int_{x_{j-1/4}}^{x_{j+1/4}} q u_h \mathrm{d}x \\ &= 2p_{j-1/4}(u_j - u_{j-1/2})/h_j - 2p_{j+1/4}(u_{j+1/2} - u_j)/h_{j+1} \\ &+ \int_{x_{j-1/4}}^{x_{j+1/4}} q u_h \mathrm{d}x, \end{aligned} \tag{2.3.5}$$

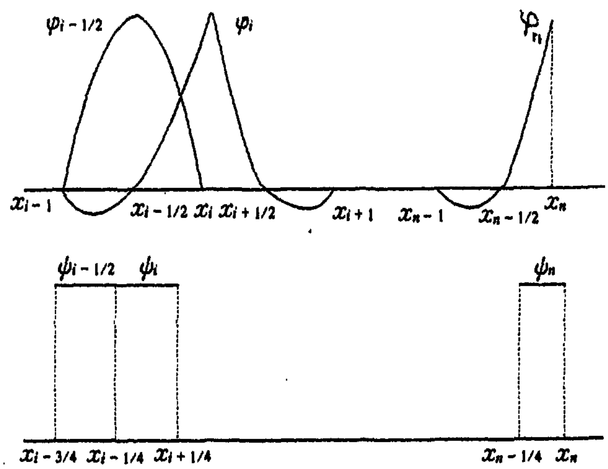



Fig. 2.3.1

$$\begin{aligned}
&a(u_h, \psi_{j-1/2}) \\
=\;& p_{j-3/4}u_h'(x_{j-3/4}) - p_{j-1/4}u_h'(x_{j-1/4}) + \int_{x_{j-3/4}}^{x_{j-1/4}} qu_h \mathrm{d}x \\
=\;& 2p_{j-3/4}(u_{j-1/2} - u_{j-1})/h_j - 2p_{j-1/4}(u_j - u_{j-1/2})/h_j \\
&+ \int_{x_{j-3/4}}^{x_{j-1/4}} qu_h \mathrm{d}x.
\end{aligned} \tag{2.3.6}$$

In the above expressions $u_0 = 0$ and, when $j = n$, the quantities on the right-hand side of $x_n = b$ should be dropped. For this reason we make the convention that $p_{n+1/4} = 0$ and $x_{n+1/4} = x_n$.

Exploiting the numerical quadrature formula

$$\int_{x_{j-1/4}}^{x_{j+1/4}} qu_h \mathrm{d}x = \frac{1}{4}(h_j + h_{j+1})q_j u_j$$

$$\int_{x_{j-3/4}}^{x_{j-1/4}} qu_h \mathrm{d}x = \frac{1}{2}h_j q_{j-1/2} u_{j-1/2}$$

leads to the following difference equation corresponding to $x_j$

$$\begin{aligned}
&a_h(u_h, \psi_j)\\
\equiv\ & p_{j-1/4}\frac{2(u_j - u_{j-1/2})}{h_j} - p_{j+1/4}\frac{2(u_{j+1/2} - u_j)}{h_{j+1}} + \frac{h_j + h_{j+1}}{4}q_j u_j\\
&= \int_{x_{j-1/4}}^{x_{j+1/4}} f\mathrm{d}x, \quad j = 1, 2, \cdots, n, \qquad (2.3.7)
\end{aligned}$$

and the one to $x_{j-1/2}$

$$\begin{aligned}
&a_h(u_h, \psi_{j-1/2})\\
\equiv\ & p_{j-3/4}\frac{2(u_{j-1/2} - u_{j-1})}{h_j} - p_{j-1/4}\frac{2(u_j - u_{j-1/2})}{h_j}\\
&+\frac{h_j}{2}q_{j-1/2}u_{j-1/2} \qquad (2.3.8)\\
=\ & \int_{x_{j-3/4}}^{x_{j-1/4}} f\mathrm{d}x, \quad j = 1, 2, \cdots, n.
\end{aligned}$$

This gives a finite difference scheme on the given grid. If the unknowns are arranged in the order $u_{1/2}, u_1, u_{3/2}, u_2, \cdots, u_{n-1/2}, u_n$, then the coefficient matrix of the resulting linear system

$$\begin{bmatrix} a_{00} & a_{01} & & & \\ a_{10} & a_{11} & a_{12} & & \\ & a_{21} & a_{22} & a_{23} & \\ & & \ddots & \ddots & \ddots \end{bmatrix}$$

is a symmetric tridiagonal matrix, where

$$\begin{aligned}
a_{00} &= \tfrac{2p_{1/4}}{h_1} + \tfrac{2p_{3/4}}{h_1} + \tfrac{h_1}{2}q_{1/2}, & a_{01} = a_{10} &= -\frac{2p_{3/4}}{h_1},\\
a_{11} &= \frac{2p_{3/4}}{h_1} + \frac{2p_{5/4}}{h_2} + \frac{h_1 + h_2}{4}q_1, & a_{12} = a_{21} &= -\frac{2p_{5/4}}{h_2},\\
a_{22} &= \frac{2p_{5/4}}{h_2} + \frac{2p_{7/4}}{h_2} + \frac{h_2}{2}q_{3/2}, & a_{23} = a_{32} &= -\frac{2p_{7/4}}{h_2},\\
a_{33} &= \frac{2p_{7/4}}{h_2} + \frac{2p_{9/4}}{h_3} + \frac{h_2 + h_3}{4}q_2, & a_{34} = a_{43} &= -\frac{2p_{9/4}}{h_3}, \cdots.
\end{aligned}$$

### 2.3.3 Convergence order estimates

Inspired by (2.3.2) and (2.3.3), we introduce the following discrete norms

$$|u_h|_{0,h} = \Big\{\sum_{i=1}^{n} h_i(u_{i-1}^2 + u_{i-1/2}^2 + u_i^2)\Big\}^{1/2}, \tag{2.3.9}$$

$$|u_h|_{1,h} = \Big\{\sum_{i=1}^{n}[(u_{i-1/2} - u_{i-1})^2 + (u_i - u_{i-1/2})^2]/h_i\Big\}^{1/2}. \tag{2.3.10}$$

**Theorem 2.3.1** *Within $U_h$, the norms $|\cdot|_{0,h}$ and $|\cdot|_{1,h}$ are equivalent to $|\cdot|_0$ and $|\cdot|_1$ respectively, namely, there exist positive constants $C_1, C_2, C_3$ and $C_4$ independent of $U_h$ such that*

$$C_1|u_h|_{0,h} \le |u_h|_0 \le C_2|u_h|_{0,h}, \ \forall u_h \in U_h, \tag{2.3.11}$$

$$C_3|u_h|_{1,h} \le |u_h|_1 \le C_4|u_h|_{1,h}, \ \forall u_h \in U_h, \tag{2.3.12}$$

**Proof** By (2.3.3)

$$|u_h|_1^2 = \sum_{i=1}^{n}\int_{x_{i-1}}^{x_i}(u_h')^2\mathrm{d}x = \sum_{i=1}^{n} h_i\delta_i^T A\delta_i,$$

where

$$\delta_i = \begin{bmatrix} (u_{i-1/2} - u_{i-1})/h_i \\ (u_i - u_{i-1/2})/h_i \end{bmatrix}, \quad A = G^T A_0 G,$$

$$G = \begin{bmatrix} -4 & 4 \\ 3 & -1 \end{bmatrix}, \quad A_0 = \int_0^1 \begin{bmatrix} \xi^2 & \xi \\ \xi & 1 \end{bmatrix}\mathrm{d}\xi = \begin{bmatrix} \frac{1}{3} & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}.$$

$A$ is positive definite since $A_0$ is positive definite and $G$ is nonsingular. Thus there exist positive constants $C_3$ and $C_4$ which are independent of $U_h$ and satisfy

$$C_3^2\delta_i^T\delta_i \le \delta_i^T A\delta_i \le C_4^2\delta_i^T\delta_i.$$

This implies (2.3.12). Similarly one can prove (2.3.11). This completes the proof. □

Define an interpolation projector $\Pi_h^* : U \to V_h$ as

$$\Pi_h^* u = \sum_{j=1}^{n} (u_{j-1/2}\psi_{j-1/2} + u_j\psi_j), \ \forall u \in U. \tag{2.3.13}$$

**Theorem 2.3.2** *For sufficiently small $h$, $a(u_h, \Pi_h^* u_h)$ is positive definite, that is, there exists a positive constant $\alpha$ such that*

$$a(u_h, \Pi_h^* u_h) \geq \alpha |u_h|_1^2, \ \forall u_h \in U_h. \tag{2.3.14}$$

**Proof** First we show that $a_h(u_h, \Pi_h^* u_h)$ is positive definite:

$$\begin{aligned}
& a_h(u_h, \Pi_h^* u_h) \\
= & \sum_{j=1}^{n} [u_{j-1/2} a_h(u_h, \psi_{j-1/2}) + u_j a_h(u_h, \psi_j)] \\
= & \sum_{j=1}^{n} [2p_{j-3/4}(u_{j-1/2} - u_{j-1})^2/h_j \\
& +2p_{j-1/4}(u_j - u_{j-1/2})^2/h_j] \\
& + \sum_{j=1}^{n} \left( \tfrac{h_j}{2} q_{j-1/2} u_{j-1/2}^2 + \tfrac{h_j + h_{j+1}}{4} q_j u_j^2 \right) \\
\geq & \ 2p_{\min} |u_h|_{1,h}^2 \geq \alpha_0 |u_h|_1^2 \ \ (\alpha_0 > 0).
\end{aligned} \tag{2.3.15}$$

Next, we deal with $a(u_h, \Pi_h^* u_h)$. Notice

$$\begin{aligned}
& |a(u_h, \Pi_h^* u_h) - a_h(u_h, \Pi_h^* u_h)| \\
= & \Big| \sum_{j=1}^{n} u_{j-1/2} \Big( \int_{x_{j-3/4}}^{x_{j-1/4}} q u_h \mathrm{d}x - \frac{h_j}{2} q_{j-1/2} u_{j-1/2} \Big) \\
& + \sum_{j=1}^{n} u_j \Big( \int_{x_{j-1/4}}^{x_{j+1/4}} q u_h \mathrm{d}x - \frac{h_j + h_{j+1}}{4} q_j u_j \Big) \Big| \\
\leq & \Big\{ \sum_{j=1}^{n} \Big[ \Big( \int_{x_{j-3/4}}^{x_{j-1/4}} (q u_h - q_{j-1/2} u_{j-1/2}) \mathrm{d}x \Big)^2 \\
& + \Big( \int_{x_{j-1/4}}^{x_{j+1/4}} (q u_h - q_j u_j) \mathrm{d}x \Big)^2 \Big] \Big\}^{1/2} \Big\{ \sum_{j=1}^{n} (u_{j-1/2}^2 + u_j^2) \Big\}^{1/2}.
\end{aligned} \tag{2.3.16}$$

By the Cauchy inequality we have

$$|qu_h - q_{j-1/2}u_{j-1/2}|^2 = \Big|\int_{x_{j-1/2}}^{x} (qu_h)'\mathrm{d}x\Big|^2$$
$$\leq \frac{h_j}{2}\int_{x_{j-3/4}}^{x_{j-1/4}} [(qu_h)']^2\mathrm{d}x,\ x_{j-3/4} \leq x \leq x_{i-1/4},$$

$$\Big[\int_{x_{j-3/4}}^{x_{j-1/4}} (qu_h - q_{j-1/2}u_{j-1/2})\mathrm{d}x\Big]^2$$
$$\leq \frac{h_j}{2}\int_{x_{j-3/4}}^{x_{j-1/4}} (qu_h - q_{j-1/2}u_{j-1/2})^2\mathrm{d}x$$
$$\leq \frac{h_j^3}{8}\int_{x_{j-3/4}}^{x_{j-1/4}} [(qu_h)']^2\mathrm{d}x.$$

Similarly,

$$\Big[\int_{x_{j-1/4}}^{x_{j+1/4}} (qu_h - q_j u_j)\mathrm{d}x\Big]^2 \leq \frac{h_j^3}{8}\int_{x_{j-1/4}}^{x_{j+1/4}} [(qu_h)']^2\mathrm{d}x.$$

So we have

$$\Big\{\sum_{j=1}^{n}\Big[\Big(\int_{x_{j-3/4}}^{x_{j-1/4}} (qu_h - q_{j-1/2}u_{j-1/2})\mathrm{d}x\Big)^2$$
$$+\Big(\int_{x_{j-1/4}}^{x_{j+1/4}} (qu_h - q_j u_j)\mathrm{d}x\Big)^2\Big]\Big\}^{1/2}$$
$$\leq \Big\{\sum_{j=1}^{n} \frac{h_j^3}{8}\Big(\int_{x_{j-3/4}}^{x_{j-1/4}} [(qu_h)']^2\mathrm{d}x + \int_{x_{j-1/4}}^{x_{j+1/4}} [(qu_h)']^2\mathrm{d}x\Big)\Big\}^{1/2}$$
$$\leq Ch^{3/2}|qu_h|_1. \tag{2.3.17}$$

Notice $q \in C^1(I)$ and the equivalence of the seminorm $|\cdot|_1$ and the full norm $\|\cdot\|_1$ in $H_E^1(I)$, then we have

$$|qu_h|_1 \leq |q'u_h|_0 + |qu_h'|_0 \leq C|u_h|_1. \tag{2.3.18}$$

It follows from the quasi-uniformity of the grid and (2.3.11) that

$$\begin{aligned}\Big\{\sum_{j=1}^{n}(u_{j-1/2}^2+u_j^2)\Big\}^{1/2} &\le Ch^{-1/2}|u_h|_{0,h} \\ \le\ & Ch^{-1/2}|u_h|_0 \le Ch^{-1/2}|u_h|_1.\end{aligned} \tag{2.3.19}$$

Combining (2.3.16)-(2.3.19) leads to

$$|a(u_h,\Pi_h^* u_h)-a_h(u_h,\Pi_h^* u_h)| \le Ch|u_h|_1^2.$$

This together with (2.3.15) implies the desired result. □

**Theorem 2.3.3** *Suppose $u \in H^3(I)$ and $u_h$ are the solutions of the problem (2.3.1) and the quadratic element difference scheme (2.3.4) respectively, then the following estimate holds:*

$$|u-u_h|_1 \le Ch^2|u|_3. \tag{2.3.20}$$

**Proof** Noticing

$$a(u,v_h)=(f,v_h),\ \forall v_h \in V_h,$$

$$a(u_h,v_h)=(f,v_h),\ \forall v_h \in V_h,$$

we have

$$a(u-u_h,v_h)=0,\ \forall v_h \in V_h. \tag{2.3.21}$$

Let $\Pi_h u$ be the interpolation projection of $u$ onto the trial space $U_h$. Then using (2.3.14) and (2.3.21) we find that

$$\begin{aligned}|u_h-\Pi_h u|_1^2 &\le \frac{1}{\alpha}a(u_h-\Pi_h u,\Pi_h^*(u_h-\Pi_h u)) \\ &= \frac{1}{\alpha}a(u-\Pi_h u,\Pi_h^*(u_h-\Pi_h u)).\end{aligned}$$

This gives

$$|u_h-\Pi_h u|_1 \le \frac{1}{\alpha}\sup_{w_h\in U_h}\frac{|a(u-\Pi_h u,\Pi_h^* w_h)|}{|w_h|_1}. \tag{2.3.22}$$

Notice

$$
\begin{aligned}
& a(u-\Pi_h u, \Pi_h^* w_h) \\
= & \sum_{j=1}^{n}[w_{j-1/2} a(u-\Pi_h u, \psi_{j-1/2}) + w_j a(u-\Pi_h u, \psi_j)] \\
= & \sum_{j=1}^{n}\Big\{w_{j-1/2}[p_{j-3/4}(u-\Pi_h u)'_{j-3/4} - p_{j-1/4}(u-\Pi_h u)'_{j-1/4}] \\
& +w_{j-1/2}\int_{x_{j-3/4}}^{x_{j-1/4}} q(u-\Pi_h u)\mathrm{d}x\Big\} \\
& +\sum_{j=1}^{n}\Big\{w_j[p_{j-1/4}(u-\Pi_h u)'_{j-1/4} - p_{j+1/4}(u-\Pi_h u)'_{j+1/4}] \\
& +w_j\int_{x_{j-1/4}}^{x_{j+1/4}} q(u-\Pi_h u)\mathrm{d}x\Big\} \\
= & \sum_{j=1}^{n}[p_{j-3/4}(u-\Pi_h u)'_{j-3/4}(w_{j-1/2}-w_{j-1}) \\
& +p_{j-1/4}(u-\Pi_h u)'_{j-1/4}(w_j - w_{j-1/2})] \\
& +\sum_{j=1}^{n}\Big[w_{j-1/2}\int_{x_{j-3/4}}^{x_{j-1/4}} q(u-\Pi_h u)\mathrm{d}x \\
& +w_j\int_{x_{j-1/4}}^{x_{1+1/4}} q(u-\Pi_h u)\mathrm{d}x\Big].
\end{aligned}
$$

Then, we use the Cauchy inequality and (2.3.12), (2.3.19) to get

$$
\begin{aligned}
& a(u-\Pi_h u, \Pi_h^* w_h) \\
\le & |p|_{\max}\Big\{\sum_{j=1}^{n}[((u-\Pi_h u)'_{j-3/4})^2 + ((u-\Pi_h u)'_{j-1/4})^2]\Big\}^{1/2} \\
& \cdot\Big\{\sum_{j=1}^{n}[(w_{j-1/2}-w_{j-1})^2 + (w_j - w_{j-1/2})^2]\Big\}^{1/2}
\end{aligned}
$$

$$
\begin{aligned}
&+|q|_{\max}\Big\{\sum_{j=1}^{n}\Big[\Big(\int_{x_{j-3/4}}^{x_{j-1/4}}|u-\Pi_h u|\mathrm{d}x\Big)^2\\
&+\Big(\int_{x_{j-1/4}}^{x_{j+1/4}}|u-\Pi_h u|\mathrm{d}x\Big)^2\Big]\Big\}^{1/2}\Big\{\sum_{j=1}^{n}(w_{j-1/2}^2+w_j^2)\Big\}^{1/2}\\
\le\ & Ch^{1/2}\Big\{\sum_{j=1}^{n}[(u-\Pi_h u)_{j-3/4}')^2+((u-\Pi_h u)_{j-1/4}')^2]\Big\}^{1/2}|w_h|_1\\
&+Ch^{-1/2}\Big\{\sum_{j=1}^{n}\Big[\Big(\int_{x_{j-3/4}}^{x_{j-1/4}}|u-\Pi_h u|\mathrm{d}x\Big)^2\\
&+\Big(\int_{x_{j-1/4}}^{x_{j+1/4}}|u-\Pi_h u|\mathrm{d}x\Big)^2\Big]\Big\}^{1/2}|w_h|_1.
\end{aligned}
\tag{2.3.23}
$$

The interpolation property gives

$$u-\Pi_h u=0,\ \text{when } x=x_{j-1},x_{j-1/2},x_j.$$

So by Rolle theorem there exist $\eta_1\in(x_{j-1},x_{j-1/2}),\eta_2\in(x_{j-1/2},x_j)$ and $\eta_3\in(\eta_1,\eta_2)$ satisfying

$$(u-\Pi_h u)'(\eta_i)=0,\ i=1,2,$$

$$(u-\Pi_h u)''(\eta_3)=0.$$

Thus,

$$
\begin{aligned}
(u-\Pi_h u)_{j-3/4}' &= \int_{\eta_1}^{x_{j-3/4}}(u-\Pi_h u)''\mathrm{d}x\\
&= \int_{\eta_1}^{x_{j-3/4}}\Big(\int_{\eta_3}^{x}(u-\Pi_h u)'''\mathrm{d}x\Big)\mathrm{d}x,
\end{aligned}
$$

$$|(u-\Pi_h u)_{j-3/4}'|\le h\int_{x_{j-1}}^{x_j}|u'''|\mathrm{d}x\le h^{3/2}\Big(\int_{x_{j-1}}^{x_j}|u'''|^2\mathrm{d}x\Big)^{1/2}.$$

Similarly we have

$$|(u-\Pi_h u)_{j-1/4}'|\le h^{3/2}\Big(\int_{x_{j-1}}^{x_j}|u'''|^2\mathrm{d}x\Big)^{1/2}.$$

Hence

$$\Big\{\sum_{j=1}^{n}[((u-\Pi_h u)'_{j-3/4})^2+((u-\Pi_h u)'_{j-1/4})^2]\Big\}^{1/2} \le Ch^{3/2}|u|_3. \quad (2.3.24)$$

On the other hand,

$$\begin{aligned} &\Big\{\sum_{j=1}^{n}\Big[\Big(\int_{x_{j-3/4}}^{x_{j-1/4}}|u-\Pi_h u|\mathrm{d}x\Big)^2+\Big(\int_{x_{j-1/4}}^{x_{j+1/4}}|u-\Pi_h u|\mathrm{d}x\Big)^2\Big]\Big\}^{1/2} \\ \le\ &\Big\{\sum_{j=1}^{n}h\Big[\int_{x_{j-3/4}}^{x_{j-1/4}}|u-\Pi_h u|^2\mathrm{d}x+\int_{x_{j-1/4}}^{x_{j+1/4}}|u-\Pi_h u|^2\mathrm{d}x\Big]\Big\}^{1/2} \\ \le\ & h^{1/2}|u-\Pi_h u|_0 \le h^{7/2}|u|_3. \end{aligned} \quad (2.3.25)$$

Substituting (2.3.24) and (2.3.25) into (2.3.23) yields

$$a(u-\Pi_h u,\Pi_h^* w_h) \le Ch^2|u|_3|w_h|_1,$$

which together with (2.3.22) implies

$$|u_h-\Pi_h u|_1 \le Ch^2|u|_3.$$

This together with the interpolation estimate

$$|u-\Pi_h u|_1 \le Ch^2|u|_3$$

gives (2.3.20) and completes the proof. □

## 2.4 Cubic Element Difference Schemes

Consider a more general two point boundary value problem

$$\begin{cases} Lu \equiv -\dfrac{\mathrm{d}}{\mathrm{d}x}\Big(p\dfrac{\mathrm{d}u}{\mathrm{d}x}\Big)+r\dfrac{\mathrm{d}u}{\mathrm{d}x}+qu=f,\ a<x<b, & (2.4.1a)\\ u(a)=0,\ u'(b)=0, & (2.4.1b)\end{cases}$$

where $p \in C^1(I),\ p \ge p_{\min} > 0,\ q, r \in C(I),\ f \in L^2(I)$.

In this section we shall deduce a generalized difference scheme with higher accuracy by choosing trial and test spaces as the cubic finite element space of Hermite type and the piecewise linear function space respectively. As in the last two sections, a numerical analysis indicates that this method enjoys the same convergence order as the usual cubic finite element method while its computation is more economical.

### 2.4.1 Trial and test spaces

As in §2.2 we place a quasi-uniform grid $T_h$ with nodes

$$a = x_0 < x_1 < x_2 < \cdots < x_n = b.$$

The trial space $U_h$ is chosen as the cubic finite element space of Hermite type, so each function $u_h$ in $U_h$ satisfies the following condition:

(i) $u_h \in C^1(I),\ u_h(a) = 0,$

(ii) $u_h$ is a cubic polynomial on each element $I_i = [x_{i-1}, x_i]$ and is determined uniquely by its values and derivatives at the two endpoints of $I_i$.

$U_h$ is a $(2n+1)$-dimensional subspace of $U = H^2(I) \cap H_E^1(I)$. To construct the basis functions, we first seek for the cubic polynomials to satisfy

$$N_0(0) = 1,\ N_0'(0) = N_0(1) = N_0'(1) = 0,$$

$$N_1'(0) = 1,\ N_1(0) = N_1(1) = N_1'(1) = 0.$$

It easily follows that

$$N_0(\xi) = (1-\xi)^2(2\xi+1),$$

$$N_1(\xi) = c\xi(1-\xi)^2. \quad \Big(c = (\frac{\mathrm{d}\xi}{\mathrm{d}x})^{-1}\Big).$$

Setting $\xi = (x - x_i)/h_{i+1}$ in $N_0(\xi)$ and $N_1(\xi)$, we obtain the right halves of the two basis functions $\phi_i^{(0)}(x)$ and $\phi_i^{(1)}(x)$ respectively.

Similarly, setting $\xi = (x_i - x)/h_i$ we get the left halves. So we have

$$\phi_i^{(0)}(x) = \begin{cases} (1 - h_i^{-1}|x - x_i|)^2(2h_i|x - x_i| + 1), \\ \qquad x_{i-1} \le x \le x_i, \\ (1 - h_{i+1}^{-1}|x - x_i|)^2(2h_{i+1}|x - x_i| + 1), \\ \qquad x_i \le x \le x_{i+1}, \\ 0, \qquad \text{elsewhere,} \end{cases}$$

$$\phi_i^{(1)}(x) = \begin{cases} (x - x_i)(h_i^{-1}|x - x_i| - 1)^2, \\ \qquad x_{i-1} \le x \le x_i, \\ (x - x_i)(h_{i+1}^{-1}|x - x_i| - 1)^2, \\ \qquad x_i \le x \le x_{i+1}, \\ 0, \qquad \text{elsewhere.} \end{cases}$$

Any $u_h \in U_h$ can be written uniquely as

$$u_h = \sum_{i=0}^{n}[u_i\phi_i^{(0)}(x) + u_i'\phi_i^{(1)}(x)],$$

where $u_0 = 0,\ u_i = u_h(x_i),\ u_i' = u_h'(x_i)$.

On the element $I_i = [x_{i-1}, x_i]$

$$\begin{aligned} u_h &= u_{i-1}(1-\xi)^2(2\xi+1) + u_i\xi^2(3-2\xi) \\ &\quad + u_{i-1}'h_i\xi(1-\xi)^2 + u_i'h_i(\xi-1)\xi^2 \\ &= (\xi^3, \xi^2, \xi, 1)\begin{bmatrix} 2 & -2 & 1 & 1 \\ -3 & 3 & -2 & -1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} u_{i-1} \\ u_i \\ h_iu_{i-1}' \\ h_iu_i' \end{bmatrix}, \end{aligned} \tag{2.4.2}$$

$$\begin{aligned} u_h' &= u_{i-1}h_i^{-1}(6\xi^2 - 6\xi) + u_ih_i^{-1}(-6\xi^2 + 6\xi) \\ &\quad + u_{i-1}'(3\xi^2 - 4\xi + 1) + u_i'(3\xi^2 - 2\xi) \\ &= (\xi^2, \xi, 1)\begin{bmatrix} -6 & 3 & 3 \\ 6 & -4 & -2 \\ 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} (u_i - u_{i-1})/h_i \\ u_{i-1}' \\ u_i' \end{bmatrix}, \end{aligned} \tag{2.4.3}$$

where $\xi = (x - x_{i-1})/h_i$.

Now we place a dual grid $T_h^*$ with nodes

$$a = x_0 < x_{1/2} < x_{3/2} < \cdots < x_{n-1/2} < x_n = b,$$

where $x_{j-1/2} = (x_j + x_{j-1})/2$. The test function space $V_h$ is chosen as the piecewise linear function space with basis functions at the point $x_j$:

$$\psi_j^{(0)}(x) = \begin{cases} 1, & x_{j-1/2} \le x \le x_{j+1/2}, \\ 0, & \text{elsewhere,} \end{cases}$$

$$\psi_j^{(1)}(x) = \begin{cases} x - x_j, & x_{j-1/2} \le x \le x_{j+1/2}, \\ 0, & \text{elsewhere.} \end{cases}$$

Any $v_h \in V_h$ has a unique expression

$$v_h = \sum_{j=0}^{n} [v_j \psi_j^{(0)}(x) + v_j' \psi_j^{(1)}(x)],$$

where $v_0 = 0$, $v_i = v_h(x_i)$, $v_i' = v_h'(x_i)$. $V_h$ is clearly a $(2n+1)$-dimensional subspace of $L^2(I)$. Typical basis functions of $U_h$ and $V_h$ are shown in Fig. 2.4.1.

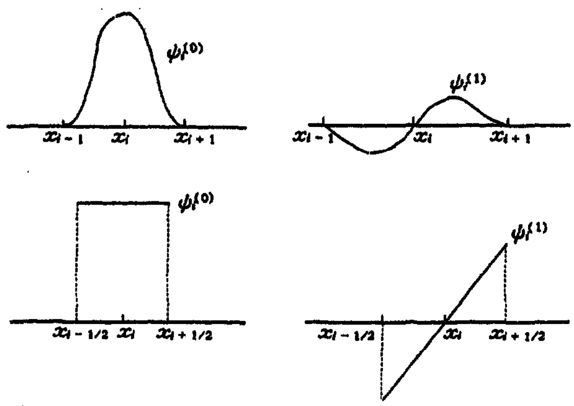



Fig. 2.4.1

### 2.4.2 Generalized difference methods

The cubic element difference scheme approximating (2.4.1) is: Find $u_h \in U_h$ such that

$$a(u_h, v_h) = (f, v_h), \ \forall v_h \in V_h, \tag{2.4.4}$$

or

$$\begin{cases} a\left(u_h, \psi_j^{(0)}\right) = \left(f, \psi_j^{(0)}\right), & j = 1, 2, \cdots, n, \\ a\left(u_h, \psi_j^{(1)}\right) = \left(f, \psi_j^{(1)}\right), & j = 0, 1, \cdots, n. \end{cases} \tag{2.4.4$'$}$$

Now we deal with the high order part of $a(u_h, v_h)$

$$b(u_h, v_h) \equiv \int_a^b p \frac{\mathrm{d}u_h}{\mathrm{d}x} \frac{\mathrm{d}v_h}{\mathrm{d}x} \mathrm{d}x.$$

Use (2.4.2), (2.4.3) and note $\xi = 1/2$ when $x = x_{j-1/2}$, then we have

$$u_{j-1/2} = u_h(x_{j-1/2}) = \frac{1}{2}u_{j-1} + \frac{1}{2}u_j + \frac{1}{8}h_j u'_{j-1} - \frac{1}{8}h_j u'_j, \quad j = 1, 2, \cdots, n, \tag{2.4.5}$$

$$u'_{j-1/2} = u'_h(x_{j-1/2}) = -\frac{3}{2}h_j^{-1}u_{j-1} + \frac{3}{2}h_j^{-1}u_j - \frac{1}{4}u'_{j-1} - \frac{1}{4}u'_j, \quad j = 1, 2, \cdots, n. \tag{2.4.6}$$

So we obtain for $j = 1, 2, \cdots, n-1$,

$$\begin{aligned} b(u_h, \psi_j^{(0)}) &= \int_a^b p \frac{\mathrm{d}u_h}{\mathrm{d}x} \frac{\mathrm{d}\psi_j^{(0)}}{\mathrm{d}x} \mathrm{d}x \\ &= p_{j-1/2}u'_{j-1/2} - p_{j+1/2}u'_{j+1/2} \\ &= \frac{3}{2}p_{j-1/2}(u_j - u_{j-1})/h_j - \frac{3}{2}p_{j+1/2}(u_{j+1} - u_j)/h_{j+1} \\ &\quad -\frac{1}{4}p_{j-1/2}u'_{j-1} + \frac{1}{4}(p_{j+1/2} - p_{j-1/2})u'_j + \frac{1}{4}p_{j+1/2}u'_{j+1}, \end{aligned} \tag{2.4.7}$$

$$\begin{aligned} b(u_h, \psi_n^{(0)}) &= \int_a^b p \frac{\mathrm{d}u_h}{\mathrm{d}x} \frac{\mathrm{d}\psi_n^{(0)}}{\mathrm{d}x} \mathrm{d}x = p_{n-1/2}u'_{j-1/2} \\ &= \frac{3}{2}p_{n-1/2}(u_n - u_{n-1})/h_n - \frac{1}{4}p_{n-1/2}(u'_{n-1} + u'_n), \end{aligned} \tag{2.4.8}$$

$$
\begin{aligned}
b(u_h, \psi_j^{(1)}) &= \int_a^b p \frac{\mathrm{d}u_h}{\mathrm{d}x} \frac{\mathrm{d}\psi_j^{(1)}}{\mathrm{d}x} \mathrm{d}x \\
&= -\frac{h_j}{2} p_{j-1/2} u'_{j-1/2} - \frac{h_{j+1}}{2} p_{j+1/2} u'_{j+1/2} + \int_{x_{j-1/2}}^{x_{j+1/2}} p u'_h \mathrm{d}x \\
&= -\frac{3}{4} p_{j-1/2}(u_j - u_{j-1}) - \frac{3}{4} p_{j+1/2}(u_{j+1} - u_j) \\
&\quad + \frac{1}{8} p_{j-1/2} h_j u'_{j-1} + \frac{1}{8}(p_{j-1/2} h_j + p_{j+1/2} h_{j+1}) u'_j \\
&\quad + \frac{1}{8} p_{j+1/2} h_{j+1} u'_{j+1} + \int_{x_{j-1/2}}^{x_{j+1/2}} p u'_h \mathrm{d}x.
\end{aligned}
\tag{2.4.9}
$$

Let us approximate the last integral in (2.4.9) in a symmetric fashion:

$$
\begin{aligned}
& \int_{x_{j-1/2}}^{x_{j+1/2}} p u'_h \mathrm{d}x = p u_h |_{x_{j-1/2}}^{x_{j+1/2}} - \int_{x_{j-1/2}}^{x_{j+1/2}} p' u_h \mathrm{d}x \\
\approx \; & p_{j+1/2} u_{j+1/2} - p_{j-1/2} u_{j-1/2} - u_j (p_{j+1/2} - p_{j-1/2}) \\
= \; & \frac{1}{2} p_{j-1/2}(u_j - u_{j-1}) + \frac{1}{2} p_{j+1/2}(u_{j+1} - u_j) - \frac{1}{8} p_{j-1/2} h_j u'_{j-1} \\
& + \frac{1}{8}(p_{j-1/2} h_j + p_{j+1/2} h_{j+1}) u'_j - \frac{1}{8} p_{j+1/2} h_{j+1} u'_{j+1}.
\end{aligned}
$$

Insert this into (2.4.9) to get an approximation of $b(u_h, \psi_j^{(1)})$:

$$
\begin{aligned}
b_h(u_h, \psi_j^{(1)}) = \; & -\frac{1}{4} p_{j-1/2}(u_j - u_{j-1}) - \frac{1}{4} p_{j+1/2}(u_{j+1} - u_j) \\
& + \frac{1}{4}(p_{j-1/2} h_j + p_{j+1/2} h_{j+1}) u'_j, \quad j = 1, 2, \cdots, n-1.
\end{aligned}
\tag{2.4.10}
$$

Similarly for the two endpoints:

$$
\begin{aligned}
b_h(u_h, \psi_0^{(1)}) &= -\frac{h_1}{2} p_{1/2} u'_{1/2} + p_{1/2} u_{1/2} - p_0 u_0 - u_0 (p_{1/2} - p_0) \\
&= \frac{1}{4} p_{1/2} h_1 u'_0 - \frac{1}{4} p_{1/2} u_1,
\end{aligned}
\tag{2.4.11}
$$

$$\begin{aligned} b_h(u_h, \psi_n^{(1)}) &= -\frac{h_n}{2} p_{n-1/2} u'_{n-1/2} \\ &\quad + p_n u_n - p_{n-1/2} u_{n-1/2} - u_n (p_n - p_{n-1/2}) \\ &= -\frac{1}{4} p_{n-1/2} (u_n - u_{n-1}) + \frac{1}{4} p_{n-1/2} h_n u'_n. \end{aligned} \tag{2.4.12}$$

If we define $p_{-1/2}, p_{n+1/2}, h_{-1}, h_{n+1}, u_{-1}, u_{n+1}$ and $u'_{n+1}$ as zero, then (2.4.8), (2.4.11) and (2.4.12) can be expressed in terms of (2.4.7) and (2.4.10) respectively, that is, (2.4.7) is valid for $j = 1, 2, \cdots, n$, and (2.4.10) for $j = 0, 1, \cdots, n$. Write

$$b_h(u_h, \psi_j^{(0)}) \equiv b(u_h, \psi_j^{(0)}), \; j = 1, 2, \cdots, n,$$

then we have the following bilinear form approximating $b(u_h, v_h)$

$$b_h(u_h, v_h) = (v'_0, v_1, v'_1, \cdots, v_n, v'_n) A (u'_0, u_1, u'_1, \cdots, u_n, u'_n)^T, \tag{2.4.13}$$

where $A$ is a symmetric matrix of the form

$$\begin{bmatrix} a_{00} & a_{01} & 0 & & & & & \\ a_{10} & a_{11} & a_{12} & a_{13} & a_{14} & & & \\ 0 & a_{21} & a_{22} & a_{23} & 0 & & & \\ & a_{31} & a_{32} & a_{33} & a_{34} & a_{35} & a_{36} & \\ & a_{41} & 0 & a_{43} & a_{44} & a_{45} & 0 & \\ & & & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix},$$

where

$$\begin{aligned} a_{00} &= \tfrac{1}{4} p_{1/2} h_1, & a_{01} &= a_{10} = -\frac{1}{4} p_{1/2}, \\ a_{11} &= \frac{3}{2}\Big(\frac{p_{1/2}}{h_1} + \frac{p_{3/2}}{h_2}\Big), & a_{12} &= a_{21} = \frac{1}{4}(p_{3/2} - p_{1/2}), \\ a_{13} &= a_{31} = -\frac{3}{2}\frac{p_{3/2}}{h_2}, & a_{14} &= a_{41} = -a_{23} = -a_{32} = \frac{1}{4} p_{3/2}, \\ a_{22} &= \frac{1}{4}(p_{1/2} h_1 + p_{3/2} h_2), & a_{33} &= \frac{3}{2}\Big(\frac{p_{3/2}}{h_2} + \frac{p_{5/2}}{h_3}\Big), \\ a_{34} &= a_{43} = \frac{1}{4}(p_{5/2} - p_{3/2}), & a_{35} &= -\frac{3}{2}\frac{p_{5/2}}{h_3}, \\ a_{36} &= -a_{45} = \frac{1}{4} p_{5/2}, & a_{44} &= \frac{1}{4}(p_{3/2} h_2 + p_{5/2} h_3). \end{aligned}$$

These give the leading terms (corresponding to the highest order derivative) of the matrix of the generalized difference equation. We see that the coefficient matrix of the generalized difference equation is sparse, and of the same bandwidth as the finite element method, but it is much more economical to obtain the discrete equation for the generalized difference method than the finite element method.

### 2.4.3 Some lemmas

First let us introduce a discrete norm for $u_h \in U_h$:

$$|u_h|_{1,h} = \Big\{\sum_{j=1}^{n} h_j\Big[\Big(\frac{u_j - u_{j-1}}{h_j}\Big)^2 + u_{j-1}'^2 + u_j'^2\Big]\Big\}^{1/2}. \tag{2.4.14}$$

The proof of the following lemma is similar to that of Theorem 2.3.1 by noting (2.4.3).

**Lemma 2.4.1** *The norms $|\cdot|_{1,h}$ and $|\cdot|_1$ are equivalent, i.e. there exist constants $C_1$ and $C_2$ independent of the subspace $U_h$ such that*

$$C_1|u_h|_{1,h} \le |u_h|_1 \le C_2|u_h|_{1,h}, \ \forall u_h \in U_h. \tag{2.4.15}$$

Next we show the positive definiteness of the leading term of the difference equation.

**Lemma 2.4.2** *The bilinear form $b(u_h, \Pi_h^* u_h)$ is positive definite, i.e., there exists a constant $\beta > 0$ independent of the subspace $U_h$ such that*

$$b(u_h, \Pi_h^* u_h) \ge \beta |u_h|_1^2, \ \forall u_h \in U_h. \tag{2.4.16}$$

**Proof** First we show the positive definiteness of $b_h(u_h, \Pi_h^* u_h)$. By (2.4.7), (2.4.10) and the definition for the quantities outside the endpoints, we have

$$
\begin{aligned}
& b_h(u_h, \Pi_h^* u_h) \\
&= \sum_{j=0}^{n} [u_j b_h(u_h, \psi_j^{(0)}) + u_j' b_h(u_h, \psi_j^{(1)})] \\
&= \sum_{j=1}^{n} \Big[ \frac{3}{2} p_{j-1/2} \frac{u_j - u_{j-1}}{h_j} u_j - \frac{3}{2} p_{j+1/2} \frac{u_{j+1} - u_j}{h_{j+1}} u_j \\
&\quad - \frac{1}{4} p_{j-1/2} (u_{j-1}' + u_j') u_j + \frac{1}{4} p_{j+1/2} (u_{j+1}' + u_j') u_j \Big] \\
&\quad + \sum_{j=0}^{n} \Big[ -\frac{1}{4} p_{j-1/2} (u_j - u_{j-1}) u_j' - \frac{1}{4} p_{j+1/2} (u_{j+1} - u_j) u_j' \\
&\quad + \frac{1}{4} p_{j-1/2} h_j u_j'^2 + \frac{1}{4} p_{j+1/2} h_{j+1} u_j'^2 \Big] \\
&= \sum_{j=1}^{n} \Big[ \frac{3}{2} p_{j-1/2} \frac{u_j - u_{j-1}}{h_j} u_j - \frac{3}{2} p_{j-1/2} \frac{u_j - u_{j-1}}{h_j} u_{j-1} \\
&\quad - \frac{1}{4} p_{j-1/2} (u_{j-1}' + u_j') u_j + \frac{1}{4} p_{j-1/2} (u_{j-1}' + u_j') u_{j-1} \Big] \\
&\quad + \sum_{j=1}^{n} \Big[ -\frac{1}{4} p_{j-1/2} (u_j - u_{j-1}) u_j' - \frac{1}{4} p_{j-1/2} (u_j - u_{j-1}) u_{j-1}' \\
&\quad + \frac{1}{4} p_{j-1/2} h_j u_j'^2 + \frac{1}{4} p_{j-1/2} h_j u_{j-1}'^2 \Big] \\
&= \sum_{j=1}^{n} h_j p_{j-1/2} \Big[ \frac{3}{2} \Big( \frac{u_j - u_{j-1}}{h_j} \Big)^2 - \frac{1}{2} \frac{u_j - u_{j-1}}{h_j} (u_j' + u_{j-1}') \\
&\quad + \frac{1}{4} (u_j'^2 + u_{j-1}'^2) \Big] \\
&\geq \frac{1}{8} P_{\min} \sum_{j=1}^{n} h_j \Big[ \Big( \frac{u_j - u_{j-1}}{h_j} \Big)^2 + u_j'^2 + u_{j-1}'^2 \Big] \\
&\geq \beta' |u_h|_1^2 \quad (\beta' > 0).
\end{aligned}
\tag{2.4.17}
$$

Next we estimate

$$\begin{aligned}
&E(u_h, \Pi_h^* u_h) \equiv b(u_h, \Pi_h^* u_h) - b_h(u_h, \Pi_h^* u_h) \\
&= \sum_{j=0}^{n} u_j' \int_{x_{j-1/2}}^{x_{j+1/2}} p'(u_j - u_h)\mathrm{d}x \\
&= \sum_{j=0}^{n} u_j' \int_{x_{j-1/2}}^{x_{j+1/2}} p' u_h'(\xi_j)(x - x_j)\mathrm{d}x.
\end{aligned}$$

It follows from (2.4.3) that on the interval $I_i = [x_{j-1}, x_j]$

$$|u_h'(x)| \le C\Big(\Big|\frac{u_j - u_{j-1}}{h_j}\Big| + |u_{j-1}'| + |u_j'|\Big).$$

This together with the quasi-uniformity of the grid leads to

$$\Big\{\sum_{j=0}^{n}[u_h'(\xi_j)]^2\Big\}^{1/2} \le Ch^{-1/2}|u_h|_{1,h},$$

$$\Big\{\sum_{j=0}^{n}(u_j')^2\Big\}^{1/2} \le Ch^{-1/2}|u_h|_{1,h}.$$

So we have

$$|E(u_h, \Pi_h^* u_h)| \le Ch|u_h|_1^2. \tag{2.4.18}$$

Combining (2.4.17) and (2.4.18) yields (2.4.16). This completes the proof. □

Denote by $P_h$ the orthogonal projector from $L^2$ to $V_h$. Then the cubic element difference scheme (2.4.4) is equivalent to

$$(Lu_h, P_h v) = (f, P_h v), \quad \forall v \in L^2(I), \tag{2.4.19}$$

or

$$L_h u_h = f_h,$$

where $L_h = P_h L$, $f_h = P_h f$.

**Lemma 2.4.3** *Suppose the homogeneous equation*

$$a(u,v) = 0, \ \forall v \in H^1_E(I) \tag{2.4.20}$$

*admits only the trivial solution, then there exists a constant $\alpha > 0$ independent of the subspace $U_h$ such that for sufficiently small $h$*

$$|||L_h u_h||| \equiv \sup_{\substack{w_h \in U_h \\ |w_h|_1 = 1}} |(Lu_h, \Pi^*_h w_h)| \geq \alpha |u_h|_1, \ \forall u_h \in U_h. \tag{2.4.21}$$

**Proof** First write $L$ as

$$L = L_1 + L_2, \tag{2.4.22}$$

where

$$\begin{aligned} L_1 u &= -\frac{\mathrm{d}}{\mathrm{d}x}\Big(p\frac{\mathrm{d}u}{\mathrm{d}x}\Big) + r\frac{\mathrm{d}u}{\mathrm{d}x} + (q+\lambda)u, \\ L_2 u &= -\lambda u, \end{aligned}$$

where $\lambda$ is a positive constant to be chosen. We find

$$\begin{aligned} & (L_1 u_h, \Pi^*_h u_h) \\ = \ & \Big(-\frac{\mathrm{d}}{\mathrm{d}x}\Big(p\frac{\mathrm{d}u_h}{\mathrm{d}x}\Big), \Pi^*_h u_h\Big) + \Big(r\frac{\mathrm{d}u_h}{\mathrm{d}x}, \Pi^*_h u_h\Big) + ((q+\lambda)u_h, \Pi^*_h u_h) \\ \geq \ & \beta |u_h|_1^2 - \max_{x\in I}|r| \cdot |u_h|_1 \cdot |\Pi^*_h u_h|_0 + (\lambda - \max_{x\in I}|q|)|u_h|_0^2 \\ & -(\lambda + \max_{x\in I}|q|)|u_h|_0 |\Pi^*_h u_h - u_h|_0. \end{aligned} \tag{2.4.23}$$

By the interpolation theory

$$|\Pi^*_h u_h - u|_0 \leq Ch^2 |u|_2, \ \forall u \in U. \tag{2.4.24}$$

The inverse property of the finite elements implies

$$|u_h|_2 \leq Ch^{-1}|u_h|_1 \leq Ch^{-2}|u_h|_0, \ \forall u_h \in U_h. \tag{2.4.25}$$

Hence

$$|\Pi^*_h u_h|_0 \leq |u_h|_0 + |\Pi^*_h u_h - u_h|_0 \leq C|u_h|_0. \tag{2.4.26}$$

Then we combine (2.4.23)-(2.4.26) to obtain

$$(L_1u_h, \Pi_h^*u_h) \geq \beta|u_h|_1^2 + C_3|u_h|_0^2 - (C_4 + C_5h)|u_h|_1|u_h|_0,$$

where the constants $C_3$ and $C_5$ increase with $\lambda$. Exploiting the $\epsilon$-inequality $ab \leq \frac{\epsilon}{2}a^2 + \frac{1}{2\epsilon}b^2$ $(\epsilon > 0)$, we see that there exists a positive constant $\gamma$ such that for sufficiently small $h$

$$(L_1u_h, \Pi_h^*u_h) \geq \gamma|u_h|_1^2, \ \forall u_h \in U_h, \tag{2.4.27}$$

where $\gamma$ is independent of $U_h$.

Next we turn to show (2.4.21). Suppose by contradiction that there exists a sequence $\{\tilde{u}_h\}$, $\tilde{u}_h \in U_h$, satisfying

$$|\tilde{u}_h|_1 = 1, \ |||L_h\tilde{u}_h||| \to 0 \text{ as } h \to 0. \tag{2.4.28}$$

Since $H_E^1(I)$ is weakly sequentially compact, $\{\tilde{u}_h\}$ has a subsequence (again written as $\{\tilde{u}_h\}$) which converges weakly to some $\tilde{u} \in H_E^1(I)$. Take any $w \in U$ and write $\Pi_h w$ as the interpolation projection of $w$ onto $U_h$. It is clear that $\Pi_h^*(w - \Pi_h w) = 0$. It follows from the interpolation theory that when $h$ is sufficiently small

$$|\Pi_h w|_1 \leq |w|_1 + |\Pi_h w - w|_1 \leq |w|_1 + Ch|w|_2 \leq \|w\|_2. \tag{2.4.29}$$

Now by (2.4.28)

$$\begin{aligned} &|(L\tilde{u}_h, \Pi_h^* w)| = |(L\tilde{u}_h, \Pi_h^*\Pi_h w)| \\ \leq \ & C|||L_h\tilde{u}_h||| \cdot |\Pi_h w|_1 \leq C|||L_h\tilde{u}_h||| \cdot \|w\|_2 \to 0 \ \ (h \to 0). \end{aligned} \tag{2.4.30}$$

On the other hand, it follows from (2.4.24) and (2.4.25) that

$$\begin{aligned} &|(L\tilde{u}_h, \Pi_h^* w - w)| \leq C\|\tilde{u}_h\|_2|\Pi_h^* w - w|_0 \\ \leq \ & Ch\|\tilde{u}_h\|_1\|w\|_2 \leq Ch\|w\|_2 \to 0 \ \ (h \to 0). \end{aligned} \tag{2.4.31}$$

Combining (2.4.30) and (2.4.31) leads to

$$a(\tilde{u}_h, w) = (L\tilde{u}_h, w) \to 0 \ \ (h \to 0). \tag{2.4.32}$$

For fixed $w \in H_E^1(I)$, $a(u, w)$ is a bounded linear functional on $H_E^1(I)$, which implies

$$a(\tilde{u}_h, w) \to a(\tilde{u}, w) \ \ (h \to 0). \tag{2.4.33}$$

By (2.4.32) and (2.4.33) we have

$$a(\tilde{u}, w) = 0, \ \forall w \in U. \tag{2.4.34}$$

In fact, the above equality is valid for any $w \in H_E^1(I)$ since $U$ is dense in $H_E^1(I)$. The assumption of the lemma then implies $\tilde{u} = 0$. So the sequence $\{\tilde{u}_h\}$ converges weakly to zero. By the compactness of the imbedding of $H_E^1(I)$ in $L^2$ we know that $\{\tilde{u}_h\}$ converges strongly to zero in $L^2$, which gives

$$|L_2\tilde{u}_h|_0 \to 0 \ \ (h \to 0).$$

Furthermore, it follows from (2.4.26) that

$$|(L_2\tilde{u}_h, \Pi_h^*\tilde{u}_h)| \le C|L_2\tilde{u}_h|_0|\tilde{u}_h|_0 \to 0 \ \ (h \to 0). \tag{2.4.35}$$

Finally by (2.4.28) and (2.4.35) we conclude

$$\begin{aligned} & |(L_1\tilde{u}_h, \Pi_h^*\tilde{u}_h)| \\ \le \ & |(L\tilde{u}_h, \Pi_h^*\tilde{u}_h)| + |(L_2\tilde{u}_h, \Pi_h^*\tilde{u}_h)| \\ \le \ & |||L_h\tilde{u}_h||| + |(L_2\tilde{u}_h, \Pi_h^*\tilde{u}_h)| \to 0 \ \ (h \to 0). \end{aligned} \tag{2.4.36}$$

This contradicts (2.4.27) and completes the proof. □

### 2.4.4 Existence, uniqueness and stability

**Theorem 2.4.1** *Assume that the homogeneous equation (2.4.20) admits only the trivial solution. Then for sufficiently small h, the cubic element difference scheme (2.4.4) has a unique solution for any given* $f \in L^2(I)$.

**Proof** By virtue of the well-known results in linear algebra theories, one only needs to show that the homogeneous equation

$$L_h u_h = 0$$

admits solely the trivial solution, which follows from (2.4.21). □

The stability problem of the scheme considers the difference between the solutions of equation (2.4.19) and its perturbed equation

$$(L_h + E_h)\hat{u}_h = f_h + g_h, \tag{2.4.37}$$

where $E_h = P_h E$, $E$ being a linear perturbation of $L$, and $g_h = P_h g$, $g$ a perturbation of $f$.

**Definition 2.4.1** *Let equation (2.4.19) be uniquely solvable, for any $f \in L^2(I)$ and $0 < h \le h_0$. (2.4.19) is said to be stable, if there exist positive constants $\alpha_0$, $\beta_0$ and $\delta_0$ independent of the subspace $U_h$ and the function $f$ such that the perturbed equation (2.4.37) always has a unique solution $\hat{u}_h \in U_h$ for any $g_h = P_h g \in V_h$, $E_h = P_h E : U_h \to V_h$, and $|||E_h||| \le \delta_0$, provided $0 < h \le h_0$; and that this solution satisfies*

$$|\hat{u}_h - u_h|_1 \le \alpha_0 |||E_h||| |u_h|_1 + \beta_0 |||g_h|||, \tag{2.4.38}$$

*where*

$$|||E_h||| = \sup_{\substack{u_h \in U_h \\ |u_h|_1 = 1}} |||E_h u_h|||,$$

$$|||E_h u_h||| = \sup_{\substack{w_h \in U_h \\ |w_h|_1 = 1}} |(E u_h, \Pi_h^* w_h)|,$$

$$|||g_h||| = \sup_{\substack{w_h \in U_h \\ |w_h|_1 = 1}} |(g, \Pi_h^* w_h)|.$$

**Theorem 2.4.2** *Let the conditions of Theorem 2.4.1 hold. Then the cubic element difference scheme (2.4.4) is stable.*

**Proof** By Theorem 2.4.1 we know that (2.4.19), and hence (2.4.4), is uniquely solvable. Choose $\delta_0$ such that $0 < \delta_0 < \alpha$. Then it follows from Lemma 2.4.3 that there exists $h_0$ such that for $0 < h \le h_0$ and $|||E_h||| < \delta_0$ we have

$$\begin{aligned} |||(L_h + E_h)u_h||| &= \sup_{\substack{w_h \in U_h \\ |w_h|_1 = 1}} |((L_h + E_h)u_h, \Pi_h^* w_h)| \\ &\ge \alpha |u_h|_1 - |||E_h u_h||| \ge (\alpha - \delta_0)|u_h|_1, \quad \forall u_h \in U_h. \end{aligned} \tag{2.4.39}$$

Thus (2.4.37) has a unique solution $\hat{u}_h$. By (2.4.19) and (2.4.37) we have

$$(L_h + E_h)(\hat{u}_h - u_h) = g_h - E_h u_h.$$

Hence,

$$\begin{aligned}(\alpha - \delta_0)|\hat{u}_h - u_h|_1 &\le |||(L_h + E_h)(\hat{u}_h - u_h)||| \\ &= |||g_h - E_h u_h||| \le |||g_h||| + |||E_h||| \cdot |u_h|_1.\end{aligned}$$

Therefore, (2.4.38) holds for $\alpha_0 = \beta_0 = (\alpha - \delta_0)^{-1}$. □

### 2.4.5 Convergence order estimates

**Theorem 2.4.3** *Let the conditions of Theorem 2.4.1 be satisfied, and let $u$ be the solution of (2.4.1) satisfying $u \in H^4(I)$ and $u_h$ the solution of the cubic element difference scheme (2.4.4). Then the following error estimate holds for sufficiently small $h$:*

$$|u - u_h|_1 \le Ch^3|u|_4. \tag{2.4.40}$$

**Proof** Clearly we have

$$(Lu, \Pi_h^* w_h) = (Lu_h, \Pi_h^* w_h), \ \forall w_h \in U_h. \tag{2.4.41}$$

By Lemma 2.4.3, (2.4.41), (2.4.24) and (2.4.25) we deduce that

$$\begin{aligned} & |\Pi_h u - u_h|_1 \\ \le\ & C|||L_h(\Pi_h u - u)||| \\ \le\ & C \sup_{\substack{w_h \in U_h \\ |w_h|_1 = 1}} |(L(\Pi_h u - u), \Pi_h^* w_h)| \\ \le\ & C \sup_{\substack{w_h \in U_h \\ |w_h|_1 = 1}} \{|(L(\Pi_h u - u), w_h)| + |(L(\Pi_h u - u), w_h - \Pi_h^* w_h)|\} \\ \le\ & C(|\Pi_h u - u|_1 + \sup_{\substack{w_h \in U_h \\ |w_h|_1 = 1}} h^2 \|\Pi_h u - u\|_2 |w_h|_2) \\ \le\ & C(|\Pi_h u - u|_1 + h\|\Pi_h u - u\|_2) \\ \le\ & Ch^3|u|_4. \end{aligned} \tag{2.4.42}$$

This leads to (2.4.40) and completes the proof. □

We remark that the error estimate (2.4.40) is of the optimal order precisely as the cubic finite element method of Hermite type.

**Definition 2.4.2** *In the deduction of the error estimates of generalized difference methods in §2.2, §2.3 and this section, the key point is to show that the bilinear forms $a(u_h, \Pi_h^* u_h)$ $(u_h \in U_h)$ satisfy the inequalities (2.2.9), (2.3.14) or (2.4.16) respectively. Henceforth in such a case we say that the bilinear form $a(u_h, \Pi_h^* u_h)$ is uniformly positive definite or $U_h$−elliptic.*

### 2.4.6 Numerical examples

The usual second order central difference method (FD), the cubic finite element method (FE) and the above three generalized difference methods (GD1, GD2, GD3) are used to solve the following boundary value problem

$$\begin{cases} -u''(x) = x^2, \ x \in (0,1), & (2.4.43a) \\ u(0) = 0, \ u'(\pi) = 0. & (2.4.43b) \end{cases}$$

The true solution (TS) of (2.4.43) is

$$u = \frac{1}{3}\pi^3 x - \frac{1}{12}x^4.$$

Take the step length $h = \pi/n$ and write $x_i = i\pi/16$, $i = 1, 2, \cdots, 16$. The numerical results of the five methods are given in Table 2.4.1. We recall that (GD) is more economical than (FE). Table 2.4.1 shows that these (GD)'s are much more accurate than (FD), and (GD3) is nearly accurate as (FE).

**Table 2.4.1 Numerical results**

| $n$ | meth. | $x_2$ | $x_4$ | $x_6$ | $x_8$ | $x_{10}$ | $x_{12}$ | $x_{14}$ | $x_{16}$ |
|---|---|---|---|---|---|---|---|---|---|
| 2 | FD | | | | 18.26438 | | | | 30.44063 |
| | GD1 | | | | 15.98118 | | | | 25.36695 |
| | GD2 | | 8.10157 | | 15.79093 | | 21.92656 | | 24.60594 |
| | GD3 | | | | 15.70229 | | | | 24.35250 |
| | FE | | | | 15.71638 | | | | 24.35250 |
| 4 | FD | | 8.37117 | | 16.36184 | | 22.83047 | | 25.87453 |
| | GD1 | | 8.10157 | | 15.79093 | | 21.92656 | | 24.60594 |
| | GD2 | 4.05772 | 8.08968 | 12.02453 | 15.74336 | 19.07971 | 21.81954 | 23.70125 | 24.41569 |
| | GD3 | | 8.08401 | | 15.72572 | | 21.78229 | | 24.35250 |
| | FE | | 8.08492 | | 15.72666 | | 21.78320 | | 24.35250 |
| 8 | FD | 4.09046 | 8.15714 | 12.12869 | 15.88020 | 19.26321 | 22.04567 | 23.97199 | 24.73301 |
| | GD1 | 4.05772 | 8.08968 | 12.02453 | 15.74336 | 19.07971 | 21.81954 | 23.70125 | 24.41569 |
| | GD2 | 4.05698 | 8.08671 | 12.01784 | 15.73147 | 19.06113 | 21.79279 | 23.66484 | 24.36813 |
| | GD3 | 4.05666 | 8.08507 | 12.01560 | 15.72753 | 19.05499 | 21.78395 | 23.65281 | 24.35250 |
| | FE | 4.05671 | 8.08513 | 12.01565 | 15.72759 | 19.05505 | 21.78401 | 23.65287 | 24.35250 |
| 16 | FD | 4.06519 | 8.10363 | 12.04336 | 15.76729 | 19.10714 | 21.84047 | 23.73269 | 24.44763 |
| | GD1 | 4.05698 | 8.08671 | 12.01784 | 15.73147 | 19.06113 | 21.79279 | 23.66484 | 24.36813 |
| | GD2 | 4.05697 | 8.08596 | 12.01617 | 15.72850 | 19.05649 | 21.78610 | 23.65573 | 24.35624 |
| | GD3 | 4.05676 | 8.08578 | 12.01572 | 15.72765 | 19.05511 | 21.78407 | 23.65291 | 24.35250 |
| | FE | 4.05676 | 8.08579 | 12.01572 | 15.72765 | 19.05511 | 21.78407 | 23.65292 | 24.35250 |
| TS | | 4.05677 | 8.08579 | 12.01572 | 15.72766 | 19.05512 | 21.78407 | 23.65292 | 24.35250 |

## 2.5 Estimates in $L^2$ and Maximum Norms

### 2.5.1 $L^2$-estimates

First let us consider the linear element difference scheme introduced in §2.2 for the two point boundary value problem (2.2.1).

**Theorem 2.5.1** *Let $u_h$ be the solution to (2.2.6), and $u$ to (2.2.1) with $u \in H_E^1(I) \cap W^{3,1}(I)$. Then the following estimate holds:*

$$\|u - u_h\|_0 \le Ch^2 \|u\|_{3,1}. \tag{2.5.1}$$

**Proof** Let us introduce an auxiliary problem: For given $g = u - u_h$, find $w \in H_E^1(I)$ such that

$$a(v, w) = (g, v), \ \forall v \in H_E^1(I). \tag{2.5.2}$$

By the differential equation theory we know that problem (2.5.2) possesses a unique solution satisfying

$$\|w\|_2 \le C\|g\|_0. \tag{2.5.3}$$

It follows from (2.5.2) and (2.2.11) that

$$\begin{aligned} &\|u-u_h\|_0^2 = a(u-u_h, w) \\ = \ & a(u-u_h, w-\Pi_h w) + a(u-u_h, \Pi_h w) - a(u-u_h, \Pi_h^* w). \end{aligned} \tag{2.5.4}$$

By (2.2.10) and (2.5.3) we have

$$\begin{aligned} |a(u-u_h, w-\Pi_h w)| & \le C|u-u_h|_1 |w-\Pi_h w|_1 \\ & \le Ch^2 |u|_2 \|u-u_h\|_0. \end{aligned} \tag{2.5.5}$$

Next we compute

$$\begin{aligned} & a(u-u_h, \Pi_h w) \\ = \ & \int_a^b p(u-u_h)'(\Pi_h w)' \mathrm{d}x \\ = \ & \sum_{j=1}^n \Big[\int_{x_{j-1}}^{x_j} (p-p_{j-1/2})(u-u_h)' \mathrm{d}x \\ & + \int_{x_{j-1}}^{x_j} p_{j-1/2}(u-u_h)' \mathrm{d}x\Big] \frac{w_j - w_{j-1}}{h_j}, \end{aligned} \tag{2.5.6}$$

$$\begin{aligned} & a(u-u_h, \Pi_h^* w) \\ = \ & \sum_{j=1}^n w_j a(u-u_h, \psi_j) \\ = \ & \sum_{j=1}^n p_{j-1/2}(u-u_h)'_{j-1/2}(w_j - w_{j-1}). \end{aligned} \tag{2.5.7}$$

Thus

$$\begin{aligned} & a(u-u_h, \Pi_h w) - a(u-u_h, \Pi_h^* w) \\ = \ & \sum_{j=1}^n \int_{x_{j-1}}^{x_j} (p-p_{j-1/2})(u-u_h)' \mathrm{d}x \frac{w_j - w_{j-1}}{h_j} \\ & + \sum_{j=1}^n p_{j-1/2}(u_j - u_{j-1} - h_j u'_{j-1/2}) \frac{w_j - w_{j-1}}{h_j}. \end{aligned} \tag{2.5.8}$$

It follows from (2.2.10) and (2.5.3) that

$$\begin{aligned}&\Big|\sum_{j=1}^{n}\int_{x_{j-1}}^{x_j}(p-p_{j-1/2})(u-u_h)'\mathrm{d}x\frac{w_j-w_{j-1}}{h_j}\Big| \\ \le\;& Ch|u-u_h|_1|w|_1 \le Ch^2|u|_2\|u-u_h\|_0.\end{aligned} \tag{2.5.9}$$

Using the Taylor expansion with an integral remainder we have

$$\begin{aligned}&\Big|\sum_{j=1}^{n}p_{j-1/2}(u_j-u_{j-1}-h_ju'_{j-1/2})\frac{w_j-w_{j-1}}{h_j}\Big| \\ \le\;& Ch^2|u|_{3,1}|w|_{1,\infty} \le Ch^2|u|_{3,1}\|u-u_h\|_0.\end{aligned} \tag{2.5.10}$$

Combining (2.5.4),(2.5.5),(2.5.8)-(2.5.10) and noting the imbedding relation $W^{3,1}(I) \to H^2(I)$, we have

$$\|u-u_h\|_0^2 \le Ch^2\|u\|_{3,1}\|u-u_h\|_0.$$

This validates the estimate (2.5.1) and completes the proof. □

Theorem 2.5.1 indicates that the solution of the linear element difference scheme possesses an optimal order estimate in $L^2$ norm, but it requires higher smoothness of the true solution $u$ than the corresponding finite element method. Maybe it is a reasonable punishment for taking only piecewise constant functions as the test space. The deduction of the $L^2$-estimate for the quadratic element difference scheme is rather tedious and is left as an exercise for interested readers.

In the next theorem we give an $L^2$ estimate for the cubic element difference scheme, which is more like the counterpart for the finite element method since here the test space has a better approximation property.

**Theorem 2.5.2** *Let $u_h$ be the solution to the cubic element difference scheme (2.4.4) and $u$ to (2.4.1) with $u \in H_E^1(I)\cap H^4(I)$. Then the following estimate holds:*

$$\|u-u_h\|_0 \le Ch^4|u|_4. \tag{2.5.11}$$

**Proof** As in the proof of Theorem 2.5.1 we introduce an auxiliary problem: For given $g = u - u_h$, find $w \in H_E^1(I)$ such that

$$a(v, w) = (g, v), \ \forall v \in H_E^1(I).$$

This together with (2.4.1) and (2.4.4) implies

$$\begin{aligned} & \|u - u_h\|_0^2 = a(u - u_h, w) \\ = \ & (L(u - u_h), w - \Pi_h^* w) \\ \leq \ & C\|u - u_h\|_2 \|w - \Pi_h^* w\|_0. \end{aligned} \tag{2.5.12}$$

By the approximation theory, the inverse property of the finite elements and (2.4.42) we have

$$\begin{aligned} \|u - u_h\|_2 & \leq \|u - \Pi_h u\|_2 + \|\Pi_h u - u_h\|_2 \\ & \leq Ch^2|u|_4 + Ch^{-1}|\Pi_h - u_h|_1 \\ & \leq Ch^2|u|_4. \end{aligned} \tag{2.5.13}$$

Also note

$$\|w - \Pi_h^* w\|_0 \leq Ch^2|w|_2 \leq Ch^2\|u - u_h\|_0. \tag{2.5.14}$$

(2.5.11) now follows from (2.5.12)-(2.5.14). This completes the proof. □

### 2.5.2 Maximum norm estimates

The $L^2$-estimates easily results in the $L^\infty$-estimates which indicate the uniform convergence of the approximate solutions to the true solutions.

**Theorem 2.5.3** *Under the assumptions of Theorem 2.5.1, the solution of the linear element scheme (2.2.6) satisfies the following error estimate:*

$$\|u - u_h\|_{0,\infty} \leq Ch^{3/2}\|u\|_{3,1}. \tag{2.5.15}$$

**Proof** Clearly we have

$$\|u - u_h\|_{0,\infty} \leq \|u - \Pi_h u\|_{0,\infty} + \|\Pi_h u - u_h\|_{0,\infty}. \tag{5.16}$$

For the first term in the right side we have, by the Sobolev interpolation theorem, for some $I_i \in T_h$ that

$$\|u - \Pi_h u\|_{0,\infty} = \|u - \Pi_h u\|_{0,\infty,I_i} \leq Ch^{3/2}|u|_{2,I_i} \leq Ch^{3/2}|u|_2. \tag{5.17}$$

For the second term, we use the inverse property of the finite element method and the $L^2$-estimate (2.5.1) to obtain

$$\begin{aligned} &\|\Pi_h u - u_h\|_{0,\infty} \\ \leq\ & Ch^{-1/2}\|\Pi_h u - u_h\|_0 \\ \leq\ & Ch^{-1/2}(\|u - u_h\|_0 + \|u - \Pi_h u\|_0) \\ \leq\ & Ch^{3/2}\|u\|_{3,1}. \end{aligned} \tag{2.5.18}$$

Combining (2.5.16)-(2.5.18) yields (2.5.15) and this completes the proof. □

The next theorem can be similarly proved.

**Theorem 2.5.4** *Under the assumption of Theorem 2.5.2, the following estimate holds for the cubic element difference scheme (2.4.4)*

$$\|u - u_h\|_{0,\infty} \leq Ch^{7/2}\|u\|_4. \tag{2.5.19}$$

## 2.6 Superconvergence

In this section we first give an outline of the concept of optimal stress points and then, in particular, we show some superconvergence results for the generalized difference methods for two point boundary value problems.

### 2.6.1 Optimal stress points

In the error analysis, the determination of the upper bound of $\|u - u_h\|_m$ is usually reduced to the estimation of $\|u - \Pi_h u\|_m$ and $a(u - \Pi_h u, \Pi_h^* w_h)$. By the approximation theory, in general we can only obtain, limited by the degree $k$ of the approximate polynomials, that

$$\|u - \Pi_h u\|_m \leq Ch^{k+1-m}\|u\|_{k+1}.$$

In general this estimate can not be improved even if the solution $u$ possesses an higher smoothness. Therefore,

$$\|u - u_h\|_m = O(h^{k+1-m})$$

is the optimal order error estimate. But this fact does not exclude the possibility that the approximation of the derivatives may be of higher order accuracy at some special points, called optimal stress points. The following definition describes an example of such points.

**Definition 2.6.1** *Point $x_0$ is called a optimal stress point if there exists a $q \in [1, \infty]$ such that*

$$|\bar{\nabla}(u - \Pi_h u)(x_0)| \leq Ch^{k+1-\frac{N}{q}}\|u\|_{k+2,q,E}, \ \forall u \in W^{k+2,q}(E), \quad (2.6.1)$$

*where $E$ denotes the union of all the elements containing $x_0$, $\bar{\nabla}v(x_0)$ the arithmetic mean of the values $\nabla v(x_0)$ at every element in $E$, $N$ the dimension of the region, and $C$ a constant independent of the grid $T_h$ and the solution $u$.*

The superconvergence theory of finite elements has clarified the distribution of the interpolation optimal stress points for some most in use finite elements. For instance, the set of the interpolation optimal stress points for the one-dimensional $\mathcal{P}_k$ type Lagrange element is

$$N_k = F\hat{N}_k,$$

where $F$ is the invertible affine mapping from the reference element $\hat{K} = [-1, 1]$ to the finite element $K$, and $\hat{N}_k$ is the set of the interpolation optimal stress points on [-1,1]:

$$\hat{N}_1 = \{0\},\ \hat{N}_2 = \left\{-\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right\},\ \hat{N}_3 = \left\{-\sqrt[3]{\frac{3}{5}}, 0, \sqrt[3]{\frac{3}{5}}\right\}, \cdots.$$

For any $x_0 \in N_k$,

$$|(u - \Pi_h u)'(x_0)| \le Ch^k \|u\|_{k+2,1,K}.$$

For the two-dimensional linear triangular elements, the interpolation optimal stress points are the midpoints of the sides. For a uniform mesh one has

$$|\bar{\nabla}(u - \Pi_h u)(x_0)| \le C|u|_{3,1,E}. \tag{2.6.3}$$

In the case of nearly uniform mesh

$$|\bar{\nabla}(u - \Pi_h u)(x_0)| \le Ch^{2-2/q}\|u\|_{3,q,E} \quad (q > 2). \tag{2.6.4}$$

Further details can be found in the references on the superconvergence of finite element methods.

Just as in the case of finite element methods, we can also obtain superconvergence results for generalized difference methods, provided we manage to get the super interpolation weak estimates. In detail we have the following theorem.

**Theorem 2.6.1** *Let $u$ and $u_h$ be solutions of the boundary value problem and its generalized difference scheme, respectively. Assume the bilinear form of the generalized difference scheme satisfies the following interpolation weak estimate: There exists $p \in [1, \infty]$ such that*

$$\begin{aligned} &|a(u - \Pi_h u, \Pi_h^* w_h)| \\ \le\ & Ch^{k+1}\|u\|_{k+2,p}\|w_h\|_1, \ \forall w_h \in U_h. \end{aligned} \tag{2.6.5}$$

*Then one has*

$$\|\Pi_h u - u_h\|_1 \le Ch^{k+1}\|u\|_{k+2,p}. \tag{2.6.6}$$

*Moreover, let $N_k$ be the set of interpolation optimal stress points: For any $x_0 \in N_k$ there exists $q \in [1, \infty]$ such that*

$$|\bar{\nabla}(u - \Pi_h u)(x_0)| \le Ch^{k+1-\frac{N}{q}}\|u\|_{k+2,q,E}. \tag{2.6.7}$$

*Then one has*

$$\Big[\frac{1}{r}\sum_{x_0 \in N_k} |\bar{\nabla}(u - u_h)(x_0)|^2\Big]^{1/2} \le Ch^{k+1}\|u\|_{k+2,\infty}, \tag{2.6.8}$$

*where $r$ is the number of points in $N_k$. (Usually the number of super points in each element is fixed, so $r = O(h^{-N})$ for an $N-$dimensional problem.)*

**Proof** It follows from the uniform $U_h-$ellipticity of the bilinear form and the weak estimate (2.6.5) that

$$\begin{aligned} & \|\Pi_h u - u_h\|_1^2 \\ \le \; & Ca(\Pi_h u - u_h, \Pi_h^*(\Pi_h u - u_h)) \\ = \; & Ca(\Pi_h u - u, \Pi_h^*(\Pi_h u - u_h)) \\ \le \; & Ch^{k+1}\|u\|_{k+2,p}\|\Pi_h u - u_h\|_1, \end{aligned}$$

which implies (2.6.6).

The inverse property of finite element methods (Theorem 1.1.13) leads to

$$|\bar{\nabla}(\Pi_h u - u_h)(x_0)| \le Ch^{-N/2}\|\Pi_h u - u_h\|_{1,E}.$$

Noticing $r = O(h^{-N})$ and using (2.6.6), we have

$$\begin{aligned} & \Big[\frac{1}{r}\sum_{x_0 \in N_k} |\bar{\nabla}(\Pi_h u - u_h)(x_0)|^2\Big]^{1/2} \\ \le \; & C\|\Pi_h u - u_h\|_1 \le Ch^{k+1}\|u\|_{k+2,p}. \end{aligned} \tag{2.6.9}$$

(2.6.7) gives

$$|\bar{\nabla}(u - \Pi_h u)(x_0)| \le Ch^{k+1}\|u\|_{k+2,\infty}.$$

Thus

$$\Big[\frac{1}{r}\sum_{x_0 \in N_k} |\bar{\nabla}(u - \Pi_h u)(x_0)|^2\Big]^{1/2} \le Ch^{k+1}\|u\|_{k+2,\infty}. \tag{2.6.10}$$

Now (2.6.8) follows from (2.6.9) and (2.6.10). This completes the proof. □

**Remark** If $p = q = 2$ in (2.6.5) and (2.6.7), then the right-hand side of (2.6.8) can be replaced by $Ch^{k+1}\|u\|_{k+2}$.

### 2.6.2 Superconvergence for linear element difference schemes

Let us consider the linear element difference scheme (2.2.6) for the two point boundary value problem (2.2.1). We first deduce a relevant interpolation weak estimate, then give a superconvergence result.

**Theorem 2.6.2** *For the linear element difference scheme (2.2.6) approximating the two point boundary value problem (2.2.1), the following interpolation weak estimate holds:*

$$\begin{aligned} &|a\,(u-\Pi_h u, \Pi_h^* w_h)\,| \le Ch^2\|u\|_{3,p}\|w_h\|_{1,p'},\\ &\forall u \in W^{3,p}(I), w_h \in U_h;\ 1 \le p, p' \le +\infty; \frac{1}{p}+\frac{1}{p'}=1. \end{aligned} \tag{2.6.11}$$

**Proof** In §2.2 we have found that

$$a\,(u-\Pi_h u, \Pi_h^* w_h) = \sum_{j=1}^{n} p_{j-1/2}(u-\Pi_h u)'(x_{j-1/2})[w_h(x_j)-w_h(x_{j-1})].$$

On $I_j = [x_{j-1}, x_j]$

$$(\Pi_h u)' = [u(x_j) - u(x_{j-1})]/h_j.$$

By the Taylor expansion with integral remainder we have

$$\begin{aligned} & u(x_j) - u(x_{j-1}) \\ = \ & u'(x_{j-1/2})h_j + \frac{1}{2!}\int_{x_{j-1/2}}^{x_j} u'''(x)(x_j-x)^2 \mathrm{d}x \\ & -\frac{1}{2!}\int_{x_{j-1/2}}^{x_{j-1}} u'''(x)(x_{j-1}-x)^2 \mathrm{d}x. \end{aligned}$$

Thus

$$\begin{aligned} & (u-\Pi_h u)'(x_{j-1/2}) \\ = \ & -\frac{1}{2h_j}\Big[\int_{x_{j-1/2}}^{x_j} u'''(x)(x_j-x)^2 \mathrm{d}x + \int_{x_{j-1}}^{x_{j-1/2}} u'''(x)(x_{j-1}-x)^2 \mathrm{d}x\Big]. \end{aligned}$$

Noting

$$w_h(x_j) - w_h(x_{j-1}) = w_h'(x)h_j,\ x \in I_j,$$

we have

$$|a\,(u-\Pi_h u,\Pi_h^* w_h)\,| \le Ch^2\sum_{j=1}^{n}|u|_{3,p,I_j}|w_h|_{1,p',I_j} \le Ch^2|u|_{3,p}|w_h|_{1,p'}.$$

This completes the proof. □

Now, combining Theorems 2.6.1 and 2.6.2 leads to the following superconvergence result for the linear element difference scheme.

**Theorem 2.6.3** *Let $u \in H_E^1(I)$ be the solution of the two point boundary value problem (2.2.1) and $u_h \in U_h$ of the linear element difference scheme (2.2.6). Assume in addition $u \in H^3(I)$. Then*

$$\|\Pi_h u - u_h\|_1 \le Ch^2\|u\|_3, \tag{2.6.12}$$

$$\Big[\frac{1}{n}\sum_{j=1}^{n}|(u-u_h)'(x_{j-1/2})|^2\Big]^{1/2} \le Ch^2\|u\|_3. \tag{2.6.13}$$

### 2.6.3 Superconvergence for cubic element difference schemes

Consider the cubic element difference scheme (2.4.4) for the two point boundary value problem (2.4.1). First we give a lemma.

**Lemma 2.6.1** *If $u \in H^5(I)$, then for $j = 1,2,\cdots,n$*

$$\begin{aligned}
&(\Pi_h u)(x_{j-1/2})\\
=\;& u(x_{j-1/2}) - \frac{1}{4!}u^{(4)}(x_{j-1/2})\Big(\frac{h}{2}\Big)^4\\
&+\frac{1}{2\cdot 4!}\Big[\int_{x_{j-1/2}}^{x_j} u^{(5)}(x)(x_j-x)^4\mathrm{d}x\\
&-\int_{x_{j-1}}^{x_{j-1/2}} u^{(5)}(x)(x_{j-1}-x)^4\mathrm{d}x\Big]\\
&-\frac{h}{8\cdot 3!}\Big[\int_{x_{j-1/2}}^{x_j} u^{(5)}(x)(x_j-x)^3\mathrm{d}x\\
&+\int_{x_{j-1}}^{x_{j-1/2}} u^{(5)}(x)(x_{j-1}-x)^3\mathrm{d}x\Big],
\end{aligned} \tag{2.6.14}$$

$$
\begin{aligned}
&(\Pi_h u)'(x_{j-1/2}) \\
= \; & u'(x_{j-1/2}) + \frac{3}{2\cdot 4!h}\Big[\int_{x_{j-1/2}}^{x_j} u^{(5)}(x)(x_j - x)^4 \mathrm{d}x \\
& + \int_{x_{j-1}}^{x_{j-1/2}} u^{(5)}(x)(x_{j-1} - x)^4 \mathrm{d}x\Big] \\
& - \frac{1}{4\cdot 3!}\Big[\int_{x_{j-1/2}}^{x_j} u^{(5)}(x)(x_j - x)^3 \mathrm{d}x \\
& - \int_{x_{j-1}}^{x_{j-1/2}} u^{(5)}(x)(x_{j-1} - x)^3 \mathrm{d}x\Big].
\end{aligned}
\tag{2.6.15}
$$

**Proof** By (2.4.2) and (2.4.3) we have

$$
\begin{aligned}
&(\Pi_h u)(x_{j-1/2}) \\
= \; & \frac{1}{2}[u(x_j) + u(x_{j-1})] - \frac{h}{8}[u'(x_j) - u'(x_{j-1})], \\
&(\Pi_h u)'(x_{j-1/2}) \\
= \; & \frac{3}{2h}[u(x_j) - u(x_{j-1})] - \frac{1}{4}[u'(x_j) + u'(x_{j-1})].
\end{aligned}
$$

Then the Taylor expansion with integral remainders leads to (2.6.14) and (2.6.15). □

**Theorem 2.6.4** *Let $T_h$ be a uniform grid ($h_j = h,\ j = 1, 2, \cdots, n$). Then the cubic element scheme (2.4.4) for the two point boundary value problem (2.4.1) satisfies the following interpolation weak estimate:*

$$
|a\,(u - \Pi_h u, \Pi_h^* w_h)\,| \le Ch^4 \|u\|_5 \|w_h\|_1,\ \forall u \in H^5(I),\ w_h \in U_h. \tag{2.6.16}
$$

**Proof** First we have,

$$
\begin{aligned}
&a\,(u - \Pi_h u, \Pi_h^* w_h) \\
= \; & \sum_{j=1}^{n} (L(u - \Pi_h u), \psi_j^{(0)}) w_h(x_j) + \sum_{j=0}^{n} (L(u - \Pi_h u), \psi_j^{(1)}) w_h'(x_j).
\end{aligned}
\tag{2.6.17}
$$

Using (2.6.15) and the equivalent norm (2.4.14), we find that

$$\begin{aligned}
&\Big|\sum_{j=1}^{n}(-[p(u-\Pi_h u)']', \psi_j^{(0)})w_h(x_j)\Big| \\
= &\Big|\sum_{j=1}^{n}[p(x_{j-1/2})(u-\Pi_h u)'(x_{j-1/2}) \\
&-p(x_{j+1/2})(u-\Pi_h u)'(x_{j+1/2})]w_h(x_j)\Big| \\
= &\Big|\sum_{j=1}^{n}p(x_{j-1/2})(u-\Pi_h u)'(x_{j-1/2})[w_h(x_j)-w_h(x_{j-1})]\Big| \\
\le &\ Ch^4|u|_5|w_h|_1.
\end{aligned} \tag{2.6.18}$$

It follows from (2.6.14) that

$$|(u-\Pi_h u)(x_{j+1/2})-(u-\Pi_h u)(x_{j-1/2})| \le Ch^4\int_{x_{j-1}}^{x_{j+1}}|u^{(5)}(x)|\mathrm{d}x. \tag{2.6.19}$$

It is clear from (2.4.2) that the norm $\|w_h\|_0$ is equivalent to the discrete norm

$$\begin{aligned}
\|w\|_{0,h} = &\Big\{\sum_{j=1}^{n}h_j[(w_h(x_{j-1}))^2+(w_h(x_j))^2 \\
&+(h_j w_h'(x_{j-1}))^2+(h_j w_h'(x_j))^2]\Big\}^{1/2}.
\end{aligned} \tag{2.6.20}$$

So we find

$$\begin{aligned}
&\Big|\sum_{j=1}^{n}(r(u-\Pi_h u)', \psi_j^{(0)})w_h(x_j)\Big| \\
= &\Big|\sum_{j=1}^{n}r(x_j)[(u-\Pi_h u)(x_{j+1/2})-(u-\Pi_h u)(x_{j-1/2})]w_h(x_j) \\
&+\sum_{j=1}^{n}\int_{x_{j-1/2}}^{x_{j+1/2}}[r(x)-r(x_j)](u-\Pi_h u)'\mathrm{d}x w_h(x_j)\Big| \\
\le &\ Ch^4\|u\|_5\|w_h\|_0.
\end{aligned} \tag{2.6.21}$$

It is obvious that

$$\begin{aligned}
&\Big|\sum_{j=1}^{n}(q(u-\Pi_h u),\psi_j^{(0)})w_h(x_j)\Big| \\
\le\ & C\sum_{j=1}^{n}\int_{x_{j-1/2}}^{x_{j+1/2}}|u-\Pi_h u|\mathrm{d}x|w_h(x_j)| \\
\le\ & Ch^4|u|_4\|w_h\|_0.
\end{aligned} \tag{2.6.22}$$

Similarly as in (2.6.21) we use (2.6.15) to obtain

$$\begin{aligned}
&\Big|\sum_{j=0}^{n}(-[p(u-\Pi_h u)']',\psi_j^{(1)})w_h'(x_j)\Big| \\
=\ & \Big|\sum_{j=0}^{n}\Big[-\frac{h}{2}p(x_{j+1/2})(u-\Pi_h u)'(x_{j+1/2}) \\
& -\frac{h}{2}p(x_{j-1/2})(u-\Pi_h u)'(x_{j-1/2}) \\
& +\int_{x_{j-1/2}}^{x_{j+1/2}}p(u-\Pi_h u)'\mathrm{d}x\Big]w_h'(x_j)\Big| \\
\le\ & Ch^4\|u\|_5|w_h|_1.
\end{aligned} \tag{2.6.23}$$

It is an easy matter to deduce that

$$\begin{aligned}
&\Big|\sum_{j=1}^{n}(r(u-\Pi_h u)'+q(u-\Pi_h u),\psi_j^{(1)})w_h'(x_j)\Big| \\
=\ & \Big|\sum_{j=1}^{n}\int_{x_{j-1/2}}^{x_{j+1/2}}[r(u-\Pi_h u)'+q(u-\Pi_h u)](x-x_j)\mathrm{d}x w_h'(x_j)\Big| \\
\le\ & Ch^4|u|_4|w_h|_1.
\end{aligned} \tag{2.6.24}$$

Now (2.6.16) follows from (2.6.17), (2.6.18) and (2.6.21)-(2.6.24). This completes the proof. □

Theorems 2.6.4 and 2.6.1 imply the following superconvergence estimate.

**Theorem 2.6.5** *Let $u$ be the solution of the two boundary value problem (2.4.1) with $u \in H_E^1(I) \cap H^5(I)$, $u_h$ of the cubic element difference scheme (2.4.4), and $T_h$ a uniform grid. Then we have*

$$\|\Pi_h u - u_h\|_1 \le Ch^4 \|u\|_5,$$

$$\left[\frac{1}{r}\sum |(u - u_h)'(x_0)|^2\right]^{1/2} \le Ch^4 \|u\|_5.$$

## 2.7 Generalized Difference Methods for a Fourth Order Equation

As an example of high order equations, let us consider the beam balance equation:

$$\begin{cases} Lu \equiv \dfrac{\mathrm{d}^2}{\mathrm{d}x^2}\left(p\dfrac{\mathrm{d}u^2}{\mathrm{d}x^2}\right) = f(x), \ a \le x \le b, & (2.7.1a) \\ u(a) = u(b) = 0, & (2.7.1b) \\ u'(a) = u'(b) = 0, & (2.7.1c) \end{cases}$$

where $p \ge p_{\min} > 0$, $p \in C^1(I)$, $f \in L^2(I)$. In this section we shall first derive a generalized difference scheme in terms of the Hermite cubic element, then give its error estimates.

### 2.7.1 Generalized difference equations

The variational problem in accordance with (2.7.1) is: Find $u \in U = H_0^2(I)$ such that

$$a(u, v) = (f, v), \ \forall v \in U, \tag{2.7.2}$$

where

$$a(u, v) = \int_a^b pu''v''\mathrm{d}x. \tag{2.7.3}$$

Discretize $I$ as in §2.4, and take the trial and test spaces $U_h$ and $V_h$ as Hermite cubic element and piecewise linear function spaces, respectively. They and their derivatives are identically zero at the boundary nodes $a$ and $b$.

The generalized difference scheme approximating (2.7.1) is: Find $u_h = \sum_{i=1}^{n-1}(u_i\phi_i^{(0)} + u_i'\phi_i^{(1)})$ such that

$$a(u_h, v_h) = (f, v_h), \ \forall v_h \in V_h, \tag{2.7.4}$$

or equivalently

$$\begin{cases} a\left(u_h, \psi_j^{(0)}\right) = \left(f, \psi_j^{(0)}\right), & j = 1, 2, \cdots, n-1, \quad (2.7.4a) \\ a\left(u_h, \psi_j^{(1)}\right) = \left(f, \psi_j^{(1)}\right), & j = 1, 2, \cdots, n-1. \quad (2.7.4b) \end{cases}$$

As pointed out in §2.1.3, here we can explain $a(u,v)$ either as (2.7.3) in the sense of generalized functions, or as the following bilinear form by piecewise integrating $(Lu, v)$ by parts

$$\begin{aligned} a(u,v) &= \sum_{j=0}^{n}\int_{x_{j-1/2}}^{x_{j+1/2}}(pu'')''v\mathrm{d}x \\ &= \sum_{j=0}^{n}\Big[(pu'')'v\Big|_{x_{j-1/2}^+}^{x_{j+1/2}^-} - \int_{x_{j-1/2}}^{x_{j+1/2}}(pu'')'v'\mathrm{d}x\Big] \\ &= \sum_{j=0}^{n}\Big[(pu'')'v\Big|_{x_{j-1/2}^+}^{x_{j+1/2}^-} - pu''v'\Big|_{x_{j-1/2}^+}^{x_{j+1/2}^-}\Big], \\ & \quad v \in V_h, \end{aligned} \tag{2.7.5}$$

where we make the convention that $x_{-1/2} = x_0$ and $x_{n+1/2} = x_n$.

Next we calculate (2.7.4a) and (2.7.4b).

$$\begin{aligned} a(u_h, \psi_j^{(0)}) &= \int_a^b pu_h''\psi_j^{(0)''}\mathrm{d}x \\ &= \int_a^b pu_h''[\delta'(x - x_{j-1/2}) - \delta'(x - x_{j+1/2})]\mathrm{d}x \\ &= -(pu_h'')'_{j-1/2} + (pu_h'')'_{j+1/2} \\ &= -(p'u_h'')_{j-1/2} - (pu_h''')_{j-1/2} \\ &\quad +(p'u_h'')_{j+1/2} + (pu_h''')_{j+1/2}. \end{aligned} \tag{2.7.6}$$

$$
\begin{aligned}
& a(u_h, \psi_j^{(1)}) = \int_a^b p u_h'' \psi_j^{(1)\prime\prime} \mathrm{d}x \\
= \quad & \int_a^b p u_h'' \Big[\delta(x - x_{j-1/2}) - \delta(x - x_{j+1/2}) \\
& -\frac{h_j}{2}\delta'(x - x_{j-1/2}) - \frac{h_{j+1}}{2}\delta'(x - x_{j+1/2})\Big] \mathrm{d}x \\
= \quad & (pu_h'')_{j-1/2} - (pu_h'')_{j+1/2} + \frac{h_j}{2}(pu_h'')'_{j-1/2} + \frac{h_{j+1}}{2}(pu_h'')'_{j+1/2}.
\end{aligned}
\tag{2.7.7}
$$

For sake of brevity we shall write $u_h(x_j) = (u_h)_j = u_j$ etc. when there is no possible confusion. By (2.4.3) we have on $[x_{j-1}, x_j]$

$$
\begin{aligned}
u_h'' = \quad & (12\xi - 6)h_j^{-2}u_{j-1} - (12\xi - 6)h_j^{-2}u_j \\
& +(6\xi - 4)h_j^{-1}u'_{j-1} + (6\xi - 2)h_j^{-1}u'_j \\
= \quad & [\xi, 1]\begin{bmatrix} 6 & 0 \\ -3 & 1 \end{bmatrix}\begin{bmatrix} \{u'_{j-1} + u'_j - 2h_j^{-1}(u_j - u_{j-1})\}/h_j \\ (u'_j - u'_{j-1})/h_j \end{bmatrix},
\end{aligned}
\tag{2.7.8}
$$

$$
u_h''' = 6\frac{u'_{j-1} + u'_j - 2h_j^{-1}(u_j - u_{j-1})}{h_j^2}. \tag{2.7.9}
$$

Substituting (2.7.8) and (2.7.9) into (2.7.6) and (2.7.7) yields

$$
\begin{aligned}
& a(u_h, \psi_j^{(0)}) \\
= \quad & -12p_{j-1/2}h_j^{-3}u_{j-1} + 12(p_{j-1/2}h_j^{-3} + p_{j+1/2}h_{j+1}^{-3})u_j \\
& -12p_{j+1/2}h_{j+1}^{-3}u_{j+1} + (-6p_{j-1/2}h_j^{-2} + p'_{j-1/2}h_j^{-1})u'_{j-1} \\
& +(6p_{j+1/2}h_{j+1}^{-2} + p'_{j+1/2}h_{j+1}^{-1})u'_{j+1} \\
& +(-6p_{j-1/2}h_j^{-2} - p'_{j-1/2}h_j^{-1} + 6p_{j+1/2}h_{j+1}^{-2} - p'_{j+1/2}h_{j+1}^{-1})u'_j,
\end{aligned}
\tag{2.7.10}
$$

$$
\begin{aligned}
& a(u_h, \psi_j^{(1)}) \\
= \quad & 6p_{j-1/2}h_j^{-2}u_{j-1} + (-6p_{j-1/2}h_j^{-2} + 6p_{j+1/2}h_{j+1}^{-2})u_j \\
& -6p_{j+1/2}h_{j+1}^{-2}u_{j+1}
\end{aligned}
$$

$$+(2p_{j-1/2}h_j^{-1} - \frac{1}{2}p'_{j-1/2})u'_{j-1} + (2p_{j+1/2}h_{j+1}^{-1} + \frac{1}{2}p'_{j+1/2})u'_{j+1}$$
$$+(4p_{j-1/2}h_j^{-1} + \frac{1}{2}p'_{j-1/2} + 4p_{j+1/2}h_{j+1}^{-1} - \frac{1}{2}p'_{j+1/2})u'_j, \tag{2.7.11}$$

where $u_0 = u'_0 = u_n = u'_n = 0$.

### 2.7.2 Positive definiteness of $a(u_h, \Pi_h^* u_h)$

It is well-known that the seminorm $|\cdot|_2$ is equivalent to the full norm $\|\cdot\|_2$ on the space $H_0^2$. Now we introduce an equivalent discrete norm. Motivated by (2.7.8), we define

$$\begin{aligned} |u_h|_{2,h} = & \Big\{\sum_{j=1}^n h_j \Big[\Big(\frac{u'_{j-1} + u'_j - 2h_j^{-1}(u_j - u_{j-1})}{h_j}\Big)^2 \\ & + \Big(\frac{u'_j - u'_{j-1}}{h_j}\Big)^2\Big]\Big\}^{1/2}, \quad u_h \in U_h. \end{aligned} \tag{2.7.12}$$

The following lemma can be easily proved similarly as Theorem 2.3.1.

**Lemma 2.7.1** The norms $|\cdot|_{2,h}$ and $|\cdot|_2$ are equivalent on $U_h$, i.e., there exist constants $c_1$ and $c_2$ independent of the subspace $U_h$ such that

$$c_1|u_h|_{2,h} \le |u_h|_2 \le c_2|u_h|_{2,h}, \quad \forall u_h \in U_h. \tag{2.7.13}$$

Using Lemma 2.7.1 one can show the following uniform ellipticity theorem.

**Theorem 2.7.1** *$a(u_h, \Pi_h^* u_h)$ is positive definite for sufficiently small $h$, i.e., there exists a positive constant $\alpha$ independent of $U_h$ such that*

$$a(u_h, \Pi_h^* u_h) \ge \alpha |u_h|_2^2, \quad \forall u_h \in U_h. \tag{2.7.14}$$

**Proof** By (2.7.6) and (2.7.7) we find that

$$\begin{aligned} & a(u_h, \Pi_h^* u_h) \\ = & \sum_{j=1}^n [a(u_h, \psi_j^{(0)})u_j + a(u_h, \psi_j^{(1)})u'_j] \end{aligned}$$

$$
\begin{aligned}
= & \sum_{j=1}^{n}[(p'_{j+1/2}u''_{j+1/2}-p'_{j-1/2}u''_{j-1/2})u_j \\
& +(p_{j+1/2}u'''_{j+1/2}-p_{j-1/2}u'''_{j-1/2})u_j \\
& -(p_{j+1/2}u''_{j+1/2}-p_{j-1/2}u''_{j-1/2})u'_j \\
& +\frac{1}{2}(h_{j+1}p'_{j+1/2}u''_{j+1/2}+h_jp'_{j-1/2}u''_{j-1/2})u'_j \\
& +\frac{1}{2}(h_{j+1}p_{j+1/2}u'''_{j+1/2}+h_jp_{j-1/2}u'''_{j-1/2})u'_j] \\
= & \sum_{j=1}^{n}[p'_{j-1/2}u''_{j-1/2}(u_{j-1}-u_j)+p_{j-1/2}u'''_{j-1/2}(u_{j-1}-u_j) \\
& +p_{j-1/2}u''_{j-1/2}(u'_j-u'_{j-1})+\frac{1}{2}h_jp'_{j-1/2}u''_{j-1/2}(u'_j+u'_{j-1}) \\
& +\frac{1}{2}h_jp_{j-1/2}u'''_{j-1/2}(u'_j+u'_{j-1})] \\
= & \sum_{j=1}^{n}[\frac{1}{2}h_jp'_{j-1/2}u''_{j-1/2}(u'_j+u'_{j-1}-2h_j^{-1}(u_j-u_{j-1})) \\
& +\frac{1}{2}h_jp_{j-1/2}u'''_{j-1/2}(u'_j+u'_{j-1}-2h_j^{-1}(u_j-u_{j-1})) \\
& +p_{j-1/2}u''_{j-1/2}(u'_j-u'_{j-1})].
\end{aligned} \tag{2.7.15}
$$

By (2.7.8) one has

$$u''_{j-1/2}=\frac{u'_j-u'_{j-1}}{h_j}. \tag{2.7.16}$$

Finally we use (2.7.16), (2.7.9) and (2.7.15) to conclude that

$$
\begin{aligned}
& a(u_h,\Pi_h^*u_h) \\
= & \sum_{j=1}^{n}\Big[\frac{1}{2}h_j^2p'_{j-1/2}\frac{u'_j-u'_{j-1}}{h_j}\frac{u'_j+u'_{j-1}-2h_j^{-1}(u_j-u_{j-1})}{h_j} \\
& +3p_{j-1/2}h_j\Big(\frac{u'_j+u'_{j-1}-2h_j^{-1}(u_j-u_{j-1})}{h_j}\Big)^2
\end{aligned}
$$

$$+p_{j-1/2}h_j\Big(\frac{u_j'-u_{j-1}'}{h_j}\Big)^2\Big]$$
$$\geq\ p_{\min}|u_h|_{2,h}^2-\frac{1}{4}h|u_h|_{2,h}^2\max_{x\in I}|p'(x)|.$$

This leads to the desired result. □

We remark that Theorem 2.7.1 implies the existence, uniqueness, and stability of the solution of the generalized difference scheme (2.7.4).

### 2.7.3 Convergence order estimates

**Theorem 2.7.2** *Let $u$ be the solution to (2.7.1) satisfying $u\in H_0^2(I)\cap H^4(I)$ and $u_h\in U_h$ to the generalized difference scheme (2.7.4). Then the following error estimate holds for sufficiently small $h$*

$$|u-u_h|_2\leq Ch^2|u|_4. \tag{2.7.17}$$

**Proof** Clearly we have

$$a(u-u_h,\Pi_h^*w_h)=0,\quad \forall w_h\in U_h. \tag{2.7.18}$$

By Theorem 2.7.1

$$\alpha|u_h-\Pi_hu|_2^2\leq a(u_h-\Pi_hu,\Pi_h^*(u_h-\Pi_hu))=a(u-\Pi_hu,\Pi_h^*(u_h-\Pi_hu)).$$

Consequently

$$|u_h-\Pi_hu|_2\leq C\sup_{w_h\in U_h}\frac{|a(u-\Pi_hu,\Pi_h^*w_h)|}{|w_h|_2}. \tag{2.7.19}$$

Write $e_h=u-\Pi_hu$ and $e_h(x_{j-1/2})=e_{j-1/2}$. Then, similar to (2.7.15) we have

$$\begin{aligned}
&a(e_h,\Pi_h^*w_h)\\
=&\sum_{j=1}^{n}\Big[\frac{1}{2}p_{j-1/2}'h_je_{j-1/2}''(w_{j-1}'+w_j'-2h_j^{-1}(w_j-w_{j-1}))\\
&+\frac{1}{2}p_{j-1/2}h_je_{j-1/2}'''(w_{j-1}'+w_j'-2h_j^{-1}(w_j-w_{j-1}))\\
&+p_{j-1/2}e_{j-1/2}''(w_j'-w_{j-1}')\Big].
\end{aligned}\tag{2.7.20}$$

By the Cauchy inequality we have

$$
\begin{aligned}
&a(e_h, \Pi_h^* w_h)\\
\le\ & C\Big\{\sum_{j=1}^{n}[(e''_{j-1/2})^2 h_j^3 + (e'''_{j-1/2})^2 h_j^3 + (e''_{j-1/2})^2 h_j]\Big\}^{1/2}\\
&\cdot\Big\{\sum_{j=1}^{n} h_j\Big[\Big(\frac{w'_{j-1}+w'_j-2h_j^{-1}(w_j-w_{j-1})}{h_j}\Big)^2+\Big(\frac{w'_j-w'_{j-1}}{h_j}\Big)^2\Big]\Big\}^{1/2}.
\end{aligned}
\tag{2.7.21}
$$

By the interpolation condition and Rolle theorem we know that $e''_h = (u - \Pi_h u)''$ has two roots $\xi_1, \xi_2$, and $e'''_h$ has one root $\eta$ in $(x_{j-1}, x_j)$. Hence,

$$e'''_h(x) = \int_\eta^x e_h^{(4)}\mathrm{d}x = \int_\eta^x u^{(4)}\mathrm{d}x,$$

$$(e'''_h(x))^2 \le h\int_{x_{j-1}}^{x_j} |u^{(4)}|^2\mathrm{d}x,\ x \in I_j,$$

$$e''_h(x) = \int_{\xi_1}^x e'''_h(x)\mathrm{d}x,$$

$$(e''_h(x))^2 \le h^3\int_{x_{j-1}}^{x_j} |u^{(4)}|^2\mathrm{d}x,\ x \in I_j.$$

Substituting these estimates into (2.7.21) and using Lemma 2.7.1 we have

$$|a(e_h, \Pi_h^* w_h)| \le Ch^2|u|_4|w_h|_2. \tag{2.7.22}$$

This together with (2.7.19) results in

$$|u_h - \Pi_h u|_2 \le Ch^2|u|_4. \tag{2.7.23}$$

Finally, the desired result (2.7.17) follows from (2.7.23) and the interpolation property

$$|u - \Pi_h u|_2 \le Ch^2|u|_4. \qquad \square$$

### 2.7.4 Numerical examples

The five point difference method (FDM), the cubic finite element method (FEM) and the cubic generalized difference method (GDM) discussed in this section are used to solve the following model problem:

$$\begin{cases} u^{(4)}(x) = x - \dfrac{1}{2},\ 0 < x < 1, \\ u(0) = u(1) = 0,\ u'(0) = u'(1) = 0. \end{cases}$$

Set the step length $h = 0.1$. The results of the three methods and the true solution (TS) $u = \frac{1}{5!}x^2(x-1)^2(x-\frac{1}{2})$ are given in Table 2.7.1 below. We see that (GDM) is much more accurate than (FDM) and is nearly as accurate as (FEM), while we recall that (GDM) needs much less computational effort than (FEM).

The upper members of each pair in Table 2.7.1 stand for function values and the lower ones derivatives.

**Table 2.7.1 Numerical results**

| x | FDM | FEM | GDM | TS |
|---|---|---|---|---|
| 0.1 | -0.00003882 | -0.00002700 | -0.00002710 | -0.00002700 |
| | | -0.00041250 | -0.00041243 | -0.00041250 |
| 0.2 | -0.00007976 | -0.00006400 | -0.00006401 | -0.00006400 |
| | | -0.00026667 | -0.00026653 | -0.00026667 |
| 0.3 | -0.00008729 | -0.00007350 | -0.00007351 | -0.00007350 |
| | | 0.00087500 | 0.00087675 | 0.00087500 |
| 0.4 | -0.00005588 | -0.00004800 | -0.00004807 | -0.00004800 |
| | | 0.00040000 | 0.00040020 | 0.00040000 |
| 0.5 | 0.00000000 | 0.00000000 | 0.00000000 | 0.00000000 |
| | | 0.00052083 | 0.00052104 | 0.00052083 |

## Bibliography and Comments

[A-25] is the earliest paper on the generalized difference methods. (Before that, at the Conference of China Mathematical Society at Chengdu in 1978, Ronghua Li gave a talk on the way to construct the generalized difference schemes.) The original motive is to generalize the usual finite difference method (including the difference method on irregular networks) such as to possess the advantages of both finite element and finite difference methods, in particular to enjoy the same convergence order as finite element methods and less computational effort as finite difference methods, while including the usual difference methods as its special case. The key step is to use the general terms of the Taylor series as the basis functions of the test function space so as to gain the computational simplicity at the price of the loss of global smoothness. The error estimates for generalized Galerkin methods by Babuska (see (1.2.42) in Chapter 1) give an inspiration to the possible convergence orders to be reached, but it fails to provide a rigorous and practical approach for further development. The references [A-25,30] and [A-53] provide a framework for the error estimates of the generalized difference methods. It is indicated by theoretical analysis as well as numerical experiments that the generalized difference methods indeed have the same convergence order as the finite element methods. The results in §§2.1, 2.2 and 2.4 come out of [A-25] and [A-53]. For the generalization of the quadratic element difference scheme to two-dimensional problems, see [A-41] and §3.4 of this book. Another form of quadratic element difference scheme is constructed in [A-54]. There are some difficulties in using the Nitsche argument to estimate the $L^2$-error when piecewise constant functions are adopted as test functions: A higher smoothness, compared with the finite element methods, is required to obtain the optimal order estimates. It is an open question whether this result could be improved (cf. [A-10]). The estimates in $L^2$ and maximum norms for quadratic element difference schemes are rather tedious and are left for interested readers.

[A-36] generalizes some superconvergence results of finite element methods to generalized difference methods, resulting in certain superconvergence estimates for the cubic element difference scheme for two point boundary value problems. Superconvergence results for linear element difference schemes are given in §2.6. The superconvergence of quadratic element difference schemes remains to be tackled.

When proceeding from second order to higher order differential equations, the construction of the generalized difference method and its theoretical analysis will encounter new difficulties. The results on the generalized difference method for a beam balance problem in §2.7 of this chapter belong to [A-35]. A class of nonconforming generalized difference methods are presented in Chapter 4 for a two-dimensional high order differential equation.

# Chapter 3

# SECOND ORDER ELLIPTIC EQUATIONS

## 3.1 Introduction

The research on the difference methods on irregular meshes over a plane region can be traced back at least to MacNeal [B-65]. But it did not develop very much theoretically or practically at that time. In the last twenty odd years, there has appeared increasingly more research on the theories and applications of the difference schemes on irregular meshes. These methods are also called in the references the finite control volume methods, or the finite volume methods. (See the corresponding references at the end of the book.) The generalized difference methods can be viewed as a generalization of the difference methods on irregular networks by absorbing the idea of the finite element methods.

Let $\Omega$ be a bounded region with a piecewise smooth boundary $\partial\Omega$ on the $(x, y)$ plane. Consider the first boundary value problem of the second order elliptic partial differential equation:

$$
\begin{cases}
Au \equiv -\Big[\dfrac{\partial}{\partial x}\Big(a_{11}\dfrac{\partial u}{\partial x} + a_{12}\dfrac{\partial u}{\partial y}\Big) & (3.1.1a)\\
\qquad + \dfrac{\partial}{\partial y}\Big(a_{21}\dfrac{\partial u}{\partial x} + a_{22}\dfrac{\partial u}{\partial y}\Big)\Big] + qu = f, \ (x,y) \in \Omega & \\
u|_{\partial\Omega} = 0, & (3.1.1b)
\end{cases}
$$

where the coefficients $a_{ij}(x,y)$ $(i,j = 1,2)$ and $q(x,y)$ are sufficiently smooth functions satisfying the elliptic condition: There exists a constant $r > 0$ such that

$$\sum_{i,j=1}^{n} a_{ij}(x,y)\xi_i\xi_j \ge r\sum_{i=1}^{n}\xi_i^2, \ \ q(x,y) \ge 0$$

holds for any real vector $(\xi_1, \xi_2) \in R^2$ and $(x,y) \in \overline{\Omega}$. We also require $f \in L^2(\Omega)$.

The corresponding variational problem for (3.1.1) is: Find $u \in U = H_0^1(\Omega)$ satisfying

$$a(u,v) = (f,v), \ \ \forall v \in U, \tag{3.1.2}$$

where

$$a(u,v) = \int_\Omega \Big[\Big(a_{11}\frac{\partial u}{\partial x} + a_{12}\frac{\partial u}{\partial y}\Big)\frac{\partial v}{\partial x} + \Big(a_{21}\frac{\partial u}{\partial x} + a_{22}\frac{\partial u}{\partial y}\Big)\frac{\partial v}{\partial y} + quv\Big]\mathrm{d}x\mathrm{d}y, \tag{3.1.3a}$$

$$(f,v) = \int_\Omega fv\,\mathrm{d}x\mathrm{d}y. \tag{3.1.3b}$$

The solution to (3.1.2) is called the generalized solution or the weak solution of (3.1.1).

Let $U_h$ and $V_h$ respectively be the suitably chosen trial and test spaces with the same finite dimension. The generalized Galerkin method is: Find $u_h \in U_h$ such that

$$a(u_h, v_h) = (f, v_h), \ \ \forall v_h \in V_h. \tag{3.1.4}$$

If $U_h = V_h \subset U$, then (3.1.4) becomes the standard Galerkin method. Usually the (conforming) finite element methods use the interpolation of spline functions to construct the piecewise polynomial space $U_h = V_h \subset U$. Breach of the inclusion relationship, namely

$U_h = V_h \not\subset U$, leads to nonconforming methods. As mentioned in §2.1.3, generalized difference methods choose $U_h \subset U$ like finite element methods, but choose $V_h$ ($\neq U_h$) as lower order piecewise polynomial spaces.

It should be pointed out that for the generalized difference methods, we always have $V_h \not\subset U$. As in the case of nonconforming finite element methods, this is due to the loss of continuity of the functions in $V_h$ on the boundary of two neighbouring elements. So the bilinear form $a(u, v)$ must be revised accordingly. For nonconforming finite element methods, the idea is to write the integral on the whole region as a sum of the integrals on every element $K$, so (3.1.3a) is rewritten as

$$\begin{aligned} a(u,v) = & \sum_K \int_K \Big[\Big(a_{11}\frac{\partial u}{\partial x} + a_{12}\frac{\partial u}{\partial y}\Big)\frac{\partial v}{\partial x} + \Big(a_{21}\frac{\partial u}{\partial x} + a_{22}\frac{\partial u}{\partial y}\Big)\frac{\partial v}{\partial y} \\ & + quv\Big] \mathrm{d}x\mathrm{d}y. \end{aligned} \tag{3.1.5}$$

Now $a(u, v)$ is well-defined on $U_h \times V_h$. For the generalized difference methods, we place a dual grid and interpret (3.1.3) in the sense of generalized functions, i.e., $\delta$-functions on the boundary of neighbouring dual elements. Or equivalently, we take $a(u, v)$ as the bilinear form resulting from the piecewise integrations in parts on the dual elements $K^*$:

$$\int_\Omega Au \cdot v \mathrm{d}x\mathrm{d}y = \sum_{K^*} \int_{K^*} Au \cdot v \mathrm{d}x\mathrm{d}y.$$

So we have

$$\begin{aligned} & a(u,v) \\ = & \sum_{K^*} \int_{K^*} \Big[\Big(a_{11}\frac{\partial u}{\partial x} + a_{12}\frac{\partial u}{\partial y}\Big)\frac{\partial v}{\partial x} \\ & + \Big(a_{21}\frac{\partial u}{\partial x} + a_{22}\frac{\partial u}{\partial y}\Big)\frac{\partial v}{\partial y} + quv\Big] \mathrm{d}x\mathrm{d}y \\ & - \sum_{K^*} \int_{\partial K^*} \Big[\Big(a_{11}\frac{\partial u}{\partial x} + a_{12}\frac{\partial u}{\partial y}\Big)v\mathrm{d}y - \Big(a_{21}\frac{\partial u}{\partial x} + a_{22}\frac{\partial u}{\partial y}\Big)v\mathrm{d}x\Big], \end{aligned} \tag{3.1.6}$$

where $\int_{\partial K^*}$ denotes the line integrals, in the counterclockwise direction, on the boundary $\partial K^*$ of the dual element.

Now (3.1.4) is an algebraic system for the approximate solutions of $u$ and its derivatives. Different choices of $U_h$ and $V_h$ lead to different schemes. In particular, if we take $V_h$ as the piecewise constant function space with the characteristic functions of the dual elements $K^*$ as the basis functions, then the above method becomes the integral interpolation method based on the integral conservation law (the balance equation)

$$\begin{aligned} & \int_{K^*} Au\mathrm{d}x\mathrm{d}y \\ = & -\int_{\partial K^*}\Big[\Big(a_{11}\frac{\partial u}{\partial x}+a_{12}\frac{\partial u}{\partial y}\Big)\mathrm{d}y \\ & -\Big(a_{21}\frac{\partial u}{\partial x}+a_{22}\frac{\partial u}{\partial y}\Big)\mathrm{d}x\Big]+\int_{K^*} qu\mathrm{d}x\mathrm{d}y \\ = & \int_{K^*} f\mathrm{d}x\mathrm{d}y. \end{aligned}$$

So the generalized difference method is a significant generalization of the finite difference method.

In the following sections different generalized difference schemes will be deduced and discussed by introducing different $U_h$ and $V_h$.

## 3.2 Generalized Difference Methods on Triangular Meshes

### 3.2.1 Trial and test function spaces

The construction of the trial and test spaces is always related to a certain mesh decomposition. Suppose $\Omega$ is a polygonal region with boundary $\partial\Omega$. Divide $\overline{\Omega}$ into a sum of finite number of small triangles such that they have no overlapping internal region; that a vertex of any triangle does not belong to a side of any other triangle; and that each vertex of $\partial\Omega$ is a vertex of a small triangle. Each triangle is called an element and the vertexes of the triangles are called nodes. Two elements are adjacent, if they share a common side. Two nodes are adjacent if they are the endpoints of the same side. All the elements $K$ constitute a triangulation of $\overline{\Omega}$, denoted by $T_h$, where $h$ is the maximum length of all the sides.

The following definition will be used throughout this book.

**Definition 3.2.1** *We use $\overline{PQ}$ to denote the line segment with endpoints $P$ and $Q$ on the plane, which may bear a direction from $P$ to $Q$ when, e.g., it is a path of line integral. We also identify $\overline{PQ}$ with the corresponding vector of $R^2$ in the usual sense. Its length is denoted by $|\overline{PQ}|$.*

Now we construct a dual decomposition $T_h^*$ related to $T_h$. Let $P_0$ be a node of a triangle, $P_i$ $(i = 1, 2, \cdots, 6)$ the adjacent nodes of $P_0$, and $M_i$ the midpoint of $\overline{P_0P_i}$ (cf. Fig. 3.2.1). Choose a point $Q_i$ in an element $\triangle P_0P_iP_{i+1}$ $(P_7 = P_1)$ and connect successively $M_1, Q_1, \cdots, M_6, Q_6, M_1$ to form a polygonal region $K_{P_0}^*$, called a dual element. The modification of the definition is obvious when $P_0$ is on the boundary. All the dual elements constitute a new decomposition, called a dual decomposition (or a dual grid). $Q_i$ is called a node of the dual decomposition. The following two dual decompositions are most important for the triangulation $T_h$:

(1) Barycenter dual decomposition. Take the barycenter $Q_i$ of the triangle $\triangle P_0P_iP_{i+1}$ as the node of the dual decomposition, as shown in Fig. 3.2.1.

(2) Circumcenter dual decomposition. Assume that the interior angles of any element of $T_h$ are not greater than $90°$. Then, take the circumcenter $Q_i$ of the element $\triangle P_0P_iP_{i+1}$ as the node of the dual decomposition. Now $\overline{Q_iQ_{i+1}}$ is the perpendicular bisector of $\overline{P_0P_{i+1}}$, cf. Fig. 3.2.2.

In the sequel we denote by $\overline{\Omega}_h$ the set of the nodes of the decomposition $T_h$, $\Omega_h = \overline{\Omega}_h \backslash \partial\Omega$ the set of the interior nodes, and $\Omega_h^*$ the set of the nodes of the dual decomposition $T_h^*$. For $Q \in \Omega_h^*$, $K_Q$ denotes the triangular element containing $Q$. Let $S_{K_Q}$ (or $S_Q$) and $S_{P_0}^*$ be the areas of the triangular element $K_Q$ and the dual element $K_{P_0}^*$ respectively. It is easy to check that if $T_h$ and $T_h^*$ are quasi-uniform (cf. Definition 1.1.10), then there exist constant $c_1, c_2, c_3 > 0$ independent of $h$ such that

$$c_1h^2 \leq S_Q \leq h^2,\ Q \in \Omega_h^*, \tag{3.2.1a}$$

$$c_2h^2 \leq S_{P_0}^* \leq c_3h^2,\ P_0 \in \overline{\Omega}_h. \tag{3.2.1b}$$

It can be readily shown that (3.2.1a) is actually a necessary and sufficient condition for the triangulation $T_h$ to be quasi-uniform. Besides, for barycenter and circumcenter dual decompositions, (3.2.1b) can be deduced from (3.2.1a). In the sequel we always assume that the decomposition is quasi-uniform.

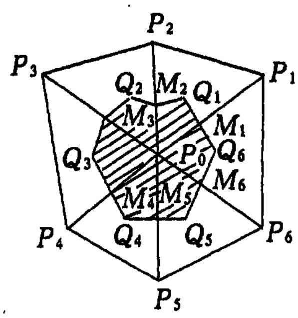



Fig. 3.2.1

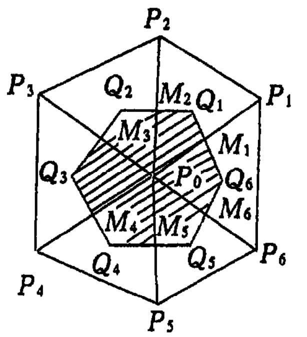



Fig. 3.2.2

The trial function space $U_h$ is chosen as the linear element space related to $T_h$. So $U_h$ is the set of all the functions $u_h$ satisfying the following conditions:

(i) $u_h \in C(\overline{\Omega})$, $u_h|_{\partial\Omega} = 0$;

(ii) $u_h|_K \in \mathcal{P}_1$, namely $u_h$ is a linear function of $x$ and $y$ on each triangular element $K \in T_h$, determined solely by its values on the three vertexes.

It is obvious that $U_h \subset U = H_0^1(\Omega)$.

Let $K = \Delta P_iP_jP_k$ be any triangular element and $P(x, y)$ a point in the element (cf. Fig. 3.2.3). Introduce the area coordinates

$(\lambda_i, \lambda_j, \lambda_k)$,

$$\lambda_i = \frac{S_i}{S} = \frac{1}{2S}\begin{vmatrix} 1 & x & y \\ 1 & x_j & y_j \\ 1 & x_k & y_k \end{vmatrix}, \tag{3.2.2a}$$

$$\lambda_j = \frac{S_j}{S} = \frac{1}{2S}\begin{vmatrix} 1 & x_i & y_i \\ 1 & x & y \\ 1 & x_k & y_k \end{vmatrix}, \tag{3.2.2b}$$

$$\lambda_k = \frac{S_k}{S} = \frac{1}{2S}\begin{vmatrix} 1 & x_i & y_i \\ 1 & x_j & y_j \\ 1 & x & y \end{vmatrix}, \tag{3.2.2c}$$

where $S_i, S_j, S_k$ and $S$ are the areas of $\Delta PP_jP_k, \Delta P_iPP_k, \Delta P_iP_jP$, and $\Delta P_iP_jP_k$, respectively. The mapping (3.2.2) maps $\Delta P_iP_jP_k$ onto a reference element $\hat{K}$ with vertexes $\hat{P}_i(0,0), \hat{P}_j(1,0)$ and $\hat{P}_k(0,1)$ on the $(\lambda_j, \lambda_k)$ plane (cf. Fig. 3.2.4).

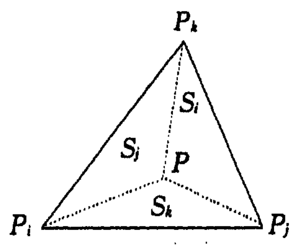


Fig. 3.2.3

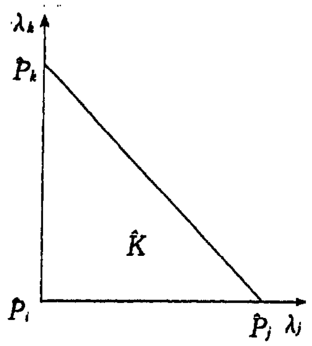


Fig. 3.2.4

The area coordinates and the orthogonal coordinates have the following relationship:

$$x = x_i\lambda_i + x_j\lambda_j + x_k\lambda_k, \tag{3.2.3a}$$

$$y = y_i\lambda_i + y_j\lambda_j + y_k\lambda_k, \tag{3.2.3b}$$

$$\lambda_i + \lambda_j + \lambda_k = 1. \tag{3.2.4}$$

It is easy to deduce that on the element $K$

$$u_h = u_i\lambda_i + u_j\lambda_j + u_k\lambda_k = u_i + (u_j - u_i)\lambda_j + (u_k - u_i)\lambda_k, \tag{3.2.5}$$

$$\begin{aligned} \frac{\partial u_h}{\partial x} &= \frac{1}{2S}\Big[\frac{\partial u_h}{\partial \lambda_j}(y_k - y_i) + \frac{\partial u_h}{\partial \lambda_k}(y_i - y_j)\Big] \\ &= \frac{1}{2S}[u_i(y_j - y_k) + u_j(y_k - y_i) + u_k(y_i - y_j)], \end{aligned} \tag{3.2.6}$$

$$\begin{aligned} \frac{\partial u_h}{\partial y} &= \frac{1}{2S}\Big[\frac{\partial u_h}{\partial \lambda_j}(x_i - x_k) + \frac{\partial u_h}{\partial \lambda_k}(x_j - x_i)\Big] \\ &= \frac{1}{2S}[u_i(x_k - x_j) + u_j(x_i - x_k) + u_k(x_j - x_i)], \end{aligned} \tag{3.2.7}$$

where and in the sequel, when there is no danger of confusion, we write in short $u_i = u_h(x_i, y_i)$ etc.

For $u \in U = H_0^1(\Omega)$, let $\Pi_h u$ be the interpolation projection of $u$ onto the trial function space $U_h$. By the interpolation theory of Sobolev spaces we have, if $u \in H^2(\Omega)$, that

$$|u - \Pi_h u|_m \le Ch^{2-m}|u|_2, \; m = 0, 1, 2. \tag{3.2.8}$$

The test space $V_h$ is chosen as the piecewise constant function space with respect to $T_h^*$, spanned by the following basis functions: For any point $P_0 \in \dot{\Omega}_k$

$$\phi_{P_0}(P) = \begin{cases} 1, & P \in K_{P_0}^*, \\ 0, & \text{elsewhere.} \end{cases} \tag{3.2.9}$$

For any $v_h \in V_h$

$$v_h = \sum_{P_0 \in \dot{\Omega}_h} v_h(P_0)\phi_{P_0}. \tag{3.2.10}$$

For $w \in U$, let $\Pi_h^* w$ be the interpolation projection of $w$ onto the test space $V_h$:

$$\Pi_h^* w = \sum_{P_0 \in \dot{\Omega}_h} w(P_0)\phi_{P_0}. \tag{3.2.11}$$

By the interpolation theory we have

$$|w - \Pi_h^* w|_0 \le Ch|w|_1. \tag{3.2.12}$$

### 3.2.2 Generalized difference equation

Choose the trial function space $U_h$ and the test function space $V_h$ as above, then the generalized difference scheme is: Find $u_h \in U_h$ such that

$$a(u_h, v_h) = (f, v_h), \ \forall v_h \in V_h, \tag{3.2.13}$$

or equivalently

$$a(u_h, \phi_{P_0}) = (f, \phi_{P_0}), \ \forall P_0 \in \dot{\Omega}_h, \tag*{(3.2.13)$'$}$$

where

$$a(u_h, v_h) = \sum_{P_0 \in \dot{\Omega}_h} v_h(P_0) a(u_h, \phi_{P_0}), \tag{3.2.14a}$$

$$\begin{aligned} &a(u_h, \phi_{P_0}) \\ &= -\int_{\partial K^*_{P_0}} [W_h^{(1)} \cos\langle n, x\rangle + W_h^{(2)} \cos\langle n, y\rangle] \mathrm{d}s + \int_{K^*_{P_0}} q u_h \mathrm{d}x\mathrm{d}y \\ &= -\int_{\partial K^*_{P_0}} W_h^{(1)} \mathrm{d}y + \int_{\partial K^*_{P_0}} W_h^{(2)} \mathrm{d}x + \int_{K^*_{P_0}} q u_h \mathrm{d}x\mathrm{d}y, \end{aligned} \tag{3.2.14b}$$

where $n$ is the unit outer normal vector and

$$W_h^{(1)} = a_{11}\frac{\partial u_h}{\partial x} + a_{12}\frac{\partial u_h}{\partial y}, \ W_h^{(2)} = a_{21}\frac{\partial u_h}{\partial x} + a_{22}\frac{\partial u_h}{\partial y}.$$

Since $\phi_{P_0}$ is taken as the characteristic function of $K^*_{P_0}$, (3.2.13)$'$ is in fact an integral conversation law (i.e. a balance equation) on $K^*_{P_0}$

$$\int_{K^*_{P_0}} Au \mathrm{d}x\mathrm{d}y = \int_{K^*_{P_0}} f \mathrm{d}x\mathrm{d}y.$$

We integrate in parts the left-hand side and then replace $u$ by $u_h$, that is, we use the piecewise linear interpolation of the solution $u$.

As in Figs. 3.2.1 or 3.2.2, we employ different numerical integration formulas to approximate (3.2.14). For instance,

$$\begin{aligned} & -\int_{\overline{Q_1M_2Q_2}} W_h^{(1)}\mathrm{d}y + \int_{\overline{Q_1M_2Q_2}} W_h^{(2)}\mathrm{d}x \\ \doteq & -[W_h^{(1)}]_{M_2}(y_{Q_2}-y_{Q_1}) + [W_h^{(2)}]_{M_2}(x_{Q_2}-x_{Q_1}) \\ = & -\Big[a_{11}(M_2)\frac{u_{P_2}-u_{P_0}}{x_{P_2}-x_{P_0}} + a_{12}(M_2)\frac{u_{P_2}-u_{P_0}}{y_{P_2}-y_{P_0}}\Big](y_{Q_2}-y_{Q_1}) \\ & +\Big[a_{21}(M_2)\frac{u_{P_2}-u_{P_0}}{x_{P_2}-x_{P_0}} + a_{22}(M_2)\frac{u_{P_2}-u_{P_0}}{y_{P_2}-y_{P_0}}\Big](x_{Q_2}-x_{Q_1}) \end{aligned} \tag{3.2.15}$$

or

$$\begin{aligned} & -\int_{\overline{M_1Q_1M_2}} W_h^{(1)}\mathrm{d}y + \int_{\overline{M_1Q_1M_2}} W_h^{(2)}\mathrm{d}x \\ \doteq & -[W_h^{(1)}]_{Q_1}(y_{M_2}-y_{M_1}) + [W_h^{(2)}]_{Q_1}(x_{M_2}-x_{M_1}) \end{aligned} \tag{3.2.16}$$

will lead to different conservative difference equations.

Next we take the Poisson equation

$$-\Delta u = f$$

as an example to study in detail the generalized difference scheme (3.2.13). Now

$$\begin{aligned} a(u_h, \phi_{P_0}) &= -\int_{\partial K^*_{P_0}} \frac{\partial u_h}{\partial n}\mathrm{d}s = -\int_{\partial K^*_{P_0}} \Big(\frac{\partial u_h}{\partial x}\mathrm{d}y - \frac{\partial u_h}{\partial y}\mathrm{d}x\Big) \\ &= -\sum_{i=1}^{6}\int_{\overline{M_iQ_iM_{i+1}}} \Big(\frac{\partial u_h}{\partial x}\mathrm{d}y - \frac{\partial u_h}{\partial y}\mathrm{d}x\Big). \end{aligned} \tag{3.2.17}$$

Since $\frac{\partial u_h}{\partial x}$ and $\frac{\partial u_h}{\partial y}$ are constants on each triangular element $K$, the above integral is independent of the position of $Q_i$. Hence

$$\begin{aligned}
&a(u_h, \phi_{P_0})\\
&= \sum_{i=1}^{6}\Big[-\frac{\partial u_h(Q_i)}{\partial x}(y_{M_{i+1}} - y_{M_i}) + \frac{\partial u_h(Q_i)}{\partial y}(x_{M_{i+1}} - x_{M_i})\Big]\\
&= \sum_{i=1}^{6} \frac{1}{4S_{Q_i}}\{[(u_{P_i} - u_{P_0})(y_{P_{i+1}} - y_{P_0})\\
&\quad +(u_{P_{i+1}} - u_{P_0})(y_{P_0} - y_{P_i})](y_{P_i} - y_{P_{i+1}})\\
&\quad +[(u_{P_i} - u_{P_0})(x_{P_0} - x_{P_{i+1}})\\
&\quad +(u_{P_{i+1}} - u_{P_0})(x_{P_i} - x_{P_0})](x_{P_{i+1}} - x_{P_i})\},
\end{aligned}$$

where $M_7 = M_1$ and $P_7 = P_1$. For a triangular element $K_{Q_i}$, write its side lengths $|\overline{P_{i+1}P_0}| = a_i$, $|\overline{P_iP_0}| = b_i$ and $|\overline{P_{i+1}P_i}| = c_i$. Then the difference equation corresponding to $P_0$ is

$$\begin{aligned}
a(u_h, \phi_{P_0}) = \ &\sum_{i=1}^{6} \frac{1}{4S_{Q_i}}\Big[(u_{P_i} - u_{P_0})\frac{b_i^2 - c_i^2 - a_i^2}{2}\\
&+(u_{P_{i+1}} - u_{P_0})\frac{a_i^2 - b_i^2 - c_i^2}{2}\Big] = \int_{K^*_{P_0}} f\mathrm{d}x\mathrm{d}y.
\end{aligned} \tag{3.2.18}$$

On the other hand, we note that the integral is independent of the position of the point $Q_i$. Therefore we can take $Q_i$ as the circumcenter of the triangle. Then it follows from the piecewise linearity of $u_h$ that

$$-\int_{\partial K^*_{P_0}} \frac{\partial u_h}{\partial n}\mathrm{d}s = -\sum_{i=1}^{6}\int_{\overline{Q_iQ_{i+1}}} \frac{\partial u_h}{\partial n}\mathrm{d}s = -\sum_{i=1}^{6} \frac{u_{P_{i+1}} - u_{P_0}}{\overline{P_{i+1}P_0}} \cdot \overline{Q_{i+1}Q_i}.$$

So the difference equation related to $P_0$ becomes

$$a(u_h, \phi_{P_0}) = -\sum_{i=1}^{6} \frac{\overline{Q_{i+1}Q_i}}{\overline{P_{i+1}P_0}}(u_{P_{i+1}} - u_{P_0}) = \int_{K^*_{P_0}} f\mathrm{d}x\mathrm{d}y. \tag{3.2.19}$$

(3.2.18) and (3.2.19) are identical. In fact, it can be verified that $\frac{c_i^2+a_i^2-b_i^2}{8S_{Q_i}}|\overline{P_iP_0}|$ and $\frac{b_i^2+c_i^2-a_i^2}{8S_{Q_i}}|\overline{P_{i+1}P_0}|$ are the distances from the circumcenter of $\Delta P_0P_iP_{i+1}$ to the sides $\overline{P_iP_0}$ and $\overline{P_{i+1}P_0}$ respectively.

For the triangulation over a non-uniform rectangular mesh as in Fig. 3.2.5, a direct calculation leads to

$$
\begin{aligned}
a(u_h, \phi_{P_0}) &= -\int_{\partial K^*_{P_{ij}}} \frac{\partial u_h}{\partial n} \mathrm{d}s \\
&= -\frac{k_1+k_2}{2}\Big(\frac{\partial u_h}{\partial x}\Big|_{p_{i+1/2,j}} - \frac{\partial u_h}{\partial x}\Big|_{p_{i-1/2,j}}\Big) \\
&\quad -\frac{h_1+h_2}{2}\Big(\frac{\partial u_h}{\partial y}\Big|_{p_{i,j+1/2}} - \frac{\partial u_h}{\partial y}\Big|_{p_{i,j-1/2}}\Big) \\
&= -\frac{k_1+k_2}{2}\Big(\frac{u_{i+1,j}-u_{i,j}}{h_2} + \frac{u_{i+1,j}-u_{i,j}}{h_1}\Big) \\
&\quad -\frac{h_1+h_2}{2}\Big(\frac{u_{i,j+1}-u_{i,j}}{k_2} + \frac{u_{i,j-1}-u_{i,j}}{k_1}\Big).
\end{aligned}
$$

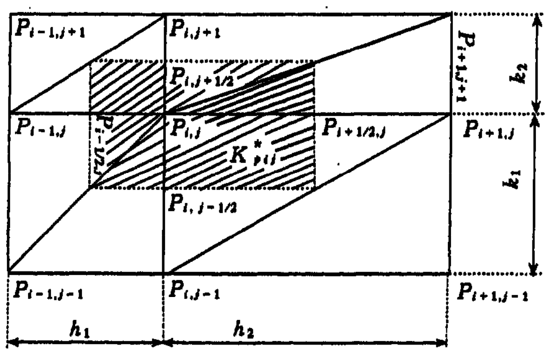



Fig. 3.2.5

So the difference equation corresponding to $P_{ij}$ is

$$\Big[\frac{k_1+k_2}{2}\Big(\frac{1}{h_1}+\frac{1}{h_2}\Big)+\frac{h_1+h_2}{2}\Big(\frac{1}{k_1}+\frac{1}{k_2}\Big)\Big]u_{ij}-$$
$$\frac{k_1+k_2}{2h_1}u_{i-1,j}-\frac{k_1+k_2}{2h_2}u_{i+1,j}-\frac{h_1+h_2}{2k_1}u_{i,j-1}-\frac{h_1+h_2}{2k_2}u_{i,j+1}$$
$$=\int_{K^*_{P_{ij}}} f\mathrm{d}x\mathrm{d}y. \tag{3.2.20}$$

For the uniform decomposition ($h_1 = h_2 = k_1 = k_2$), (3.2.20) reads

$$4u_{ij}-u_{i-1,j}-u_{i+1,j}-u_{i,j-1}-u_{i,j+1}=\int_{K^*_{P_{ij}}} f\mathrm{d}x\mathrm{d}y. \tag{3.2.21}$$

This is precisely the five point difference scheme.

Now let us consider an equilateral triangulation as in Fig. 3.2.6. Write $|\overline{P_0P_i}| = h$ $(i = 1,\cdots,6)$, then $|\overline{Q_iQ_{i+1}}| = \frac{h}{\sqrt{3}}$, and the difference equation related to $P_0$ reads

$$\frac{1}{\sqrt{3}}\Big(6u_{P_0}-\sum_{i=1}^{6}u_{P_i}\Big)=\int_{K^*_{P_0}} f\mathrm{d}x\mathrm{d}y.$$

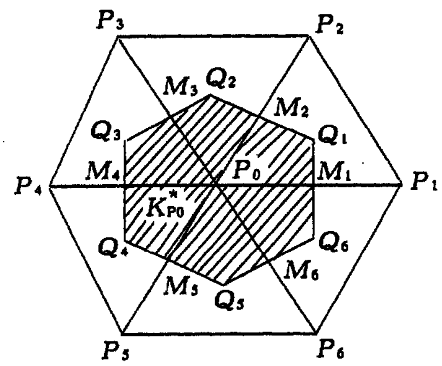



Fig. 3.2.6

### 3.2.3 *a priori* estimates

Let us introduce the following discrete zero norm, semi-norm and full-norm:

$$||u_h||_{0,h} = \Big(\sum_{K\in T_h} |u_h|^2_{0,h,K}\Big)^{1/2}, \tag{3.2.22a}$$

$$|u_h|_{1,h} = \Big(\sum_{K\in T_h} |u_h|^2_{1,h,K}\Big)^{1/2}, \tag{3.2.22b}$$

$$||u_h||_{1,h} = (||u_h||^2_{0,h} + |u_h|^2_{1,h})^{1/2}, \tag{3.2.22c}$$

where $K = K_Q = \triangle P_i P_j P_k$ and

$$|u_h|_{0,h,K} = [\tfrac{1}{3}(u_i^2 + u_j^2 + u_k^2)S_Q]^{1/2},$$

$$|u_h|_{1,h,K} = \Big\{\Big[\Big(\frac{\partial u_h(Q)}{\partial x}\Big)^2 + \Big(\frac{\partial u_h(Q)}{\partial y}\Big)^2\Big]S_Q\Big\}^{1/2}.$$

These discrete norms and the continuous norms of the Sobolev spaces have the following relations.

**Lemma 3.2.1** *For $u_h \in U_h$, $|\cdot|_{1,h}$ and $|\cdot|_1$ are identical; $||\cdot||_{0,h}$ and $||\cdot||_{1,h}$ are equivalent with $||\cdot||_0$ and $||\cdot||_1$ respectively, that is, there exist positive constants $c_1, \cdots, c_4$ independent of $U_h$ such that*

$$c_1||u_h||_{0,h} \le ||u_h||_0 \le c_2||u_h||_{0,h}, \ \forall u_h \in U_h, \tag{3.2.23a}$$

$$c_3||u_h||_{1,h} \le ||u_h||_1 \le c_4||u_h||_{1,h}, \ \forall u_h \in U_h. \tag{3.2.23b}$$

**Proof** The identification of the two norms $|\cdot|_{1,h}$ and $|\cdot|_1$ results from the fact that $\frac{\partial u_h}{\partial x}$ and $\frac{\partial u_h}{\partial y}$ are constants on each element. Since $u_h$ is linear in $K$ we can use the numerical integration formula with second order accuracy to compute that

$$\begin{aligned}
&\int_K u_h^2 \mathrm{d}x\mathrm{d}y \\
&= \frac{1}{3}[u_h^2(M_i) + u_h^2(M_j) + u_h^2(M_k)]S_Q \\
&= \frac{1}{6}(u_i^2 + u_j^2 + u_k^2 + u_iu_j + u_iu_k + u_ju_k)S_Q \\
&= \frac{1}{12}[(u_i^2 + u_j^2 + u_k^2) + (u_i + u_j + u_k)^2]S_Q,
\end{aligned}$$

where $M_i, M_j$ and $M_k$ are the midpoints of $\overline{P_jP_k}$, $\overline{P_kP_i}$ and $\overline{P_iP_j}$ respectively (cf. Fig. 3.2.7). This equality gives

$$\frac{1}{4}||u_h||_{0,h}^2 \leq ||u_h||_0^2 \leq ||u_h||_{0,h}^2.$$

This gives (3.2.23a), and leads to (3.2.23b) thanks to the identification of $|\cdot|_1$ and $|\cdot|_{1,h}$. □

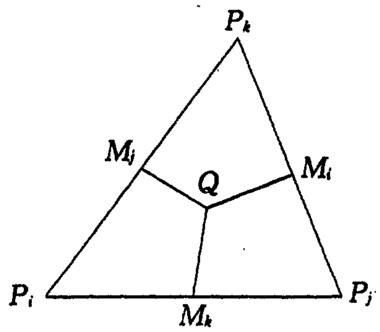



Fig. 3.2.7

**Theorem 3.2.1** *$a(u_h, \Pi_h^* u_h)$ is positive definite for small enough $h$, namely, there exist $h_0 > 0$, $\alpha > 0$ such that for $0 < h \leq h_0$*

$$a(u_h, \Pi_h^* u_h) \geq \alpha ||u_h||_1^2, \ \forall u_h \in U_h. \tag{3.2.24}$$

**Proof** It follows from (3.2.11) and (3.2.14) that

$$a(u_h, \Pi_h^* u_h) = \sum_{K \in T_h} I_K(u_h, \Pi_h^* u_h), \tag{3.2.25}$$

where

$$\begin{aligned} & I_K(u_h, \Pi_h^* u_h) \\ = & \sum_{P \in \dot{K}} \Big[ - \int_{\partial K_P^* \cap K} (W_h^{(1)} \mathrm{d}y - W_h^{(2)} \mathrm{d}x) + \int_{K_P^* \cap K} q u_h \mathrm{d}x\mathrm{d}y \Big] u_h(P), \end{aligned} \tag{3.2.26}$$

where $\dot{K}$ denotes the set of the three vertexes of $K = \triangle P_iP_jP_k$.

First let us prove the positive definiteness of the approximate bilinear form

$$a_h(u_h, \Pi_h^* u_h) = \sum_{K\in T_h} \tilde{I}_K(u_h, \Pi_h^* u_h),$$

where

$$\begin{aligned} &\tilde{I}_K(u_h, \Pi_h^* u_h) \\ = \ & [W_h^{(1)}(Q)(y_{M_k} - y_{M_j}) + W_h^{(2)}(Q)(x_{M_j} - x_{M_k})]u_h(P_i) \\ &+[W_h^{(1)}(Q)(y_{M_i} - y_{M_k}) + W_h^{(2)}(Q)(x_{M_k} - x_{M_i})]u_h(P_j) \\ &+[W_h^{(1)}(Q)(y_{M_j} - y_{M_i}) + W_h^{(2)}(Q)(x_{M_i} - x_{M_j})]u_h(P_k) \\ &+ \sum_{P\in\dot{K}} q(P)u_h(P)S_{K_P^*\cap K}\cdot u_h(P), \end{aligned} \tag{3.2.27}$$

where $S_{K_P^*\cap K}$ denotes the area of $K_P^* \cap K$. It follows from (3.2.6) and (3.2.7) that

$$\begin{aligned} &\tilde{I}_K(u_h, \Pi_h^* u_h) \\ = \ & \Big[a_{11}(Q)\Big(\frac{\partial u_h(Q)}{\partial x}\Big)^2 + (a_{12}(Q) + a_{21}(Q))\frac{\partial u_h(Q)}{\partial x}\frac{\partial u_h(Q)}{\partial y} \\ &+a_{22}(Q)\Big(\frac{\partial u_h(Q)}{\partial x}\Big)^2\Big]S_Q + \sum_{P\in\dot{K}} q(P)u_h^2(P)S_{K_P^*\cap K}. \end{aligned} \tag{3.2.28}$$

By the elliptic condition we have

$$\tilde{I}_K(u_h, \Pi_h^* u_h) \geq r\Big[\Big(\frac{\partial u_h(Q)}{\partial x}\Big)^2 + \Big(\frac{\partial u_h(Q)}{\partial y}\Big)^2\Big]S_Q.$$

Hence, by Lemma 3.2.1 and the equivalence of the semi-norm and the full norm on $H_0^1$, there exists a constant $r' > 0$ such that

$$a_h(u_h, \Pi_h^* u_h) \geq r'\|u_h\|_1^2, \ \forall u_h \in U_h. \tag{3.2.29}$$

Next we show the positive definiteness of $a(u_h, \Pi_h^* u_h)$. It is easy to see that

$$
\begin{aligned}
& I_K(u_h, \Pi_h^* u_h) - \tilde{I}_K(u_h, \Pi_h^* u_h) \\
&= \sum_{P\in \dot{K}} \Big\{ -\int_{\partial K_P^* \cap K} [(W_h^{(1)} - W_h^{(1)}(Q))\mathrm{d}y \\
& -(W_h^{(2)} - W_h^{(2)}(Q))\mathrm{d}x] \\
& + \int_{K_P^* \cap K} (q u_h - q(P) u_h(P))\mathrm{d}x\mathrm{d}y \Big\} u_h(P) \\
&= \Big\{ \sum_{l=i,j,k} \int_{\overline{M_l Q}} [(W_h^{(1)} - W_h^{(1)}(Q))\mathrm{d}y \\
& -(W_h^{(2)} - W_h^{(2)}(Q))\mathrm{d}x](u_{l+2} - u_{l+1}) \\
& + \sum_{P\in \dot{K}} \int_{K_P^* \cap K} (q u_h - q(P) u_h(P))\mathrm{d}x\mathrm{d}y \Big\} \cdot u_h(P),
\end{aligned}
\tag{3.2.30}
$$

where we set $u_{i+1} = u_j$, $u_{j+1} = u_k$, $u_{k+1} = u_i$, and $u_l = u_h(P_l)$. Since $\frac{\partial u_h}{\partial x}$ and $\frac{\partial u_h}{\partial y}$ are constants in $K$ we have

$$
\begin{aligned}
& |W_h^{(i)} - W_h^{(i)}(Q)| \\
= \ & \Big|(a_{i1} - a_{i1}(Q))\frac{\partial u_h}{\partial x} + (a_{i2} - a_{i2}(Q))\frac{\partial u_h}{\partial y}\Big| \\
\leq \ & Ch\Big(\Big|\frac{\partial u_h}{\partial y}\Big| + \Big|\frac{\partial u_h}{\partial y}\Big|\Big),\ i = 1, 2.
\end{aligned}
\tag{3.2.31}
$$

Noticing the linearity of $u_h$ in $K$ and employing the Taylor expansion we have

$$
\begin{aligned}
& |u_{l+2} - u_{l+1}| \\
= \ & \Big|\frac{\partial u_h}{\partial x}(x_{P_{l+2}} - x_{P_{l+1}}) + \frac{\partial u_h}{\partial y}(y_{P_{l+2}} - y_{P_{l+1}})\Big| \\
\leq \ & h\Big(\Big|\frac{\partial u_h}{\partial x}\Big| + \Big|\frac{\partial u_h}{\partial y}\Big|\Big),\ l = i, j, k.
\end{aligned}
\tag{3.2.32}
$$

By (3.2.31), (3.2.32) and the quasi-uniformity of the decomposition we have

$$
\begin{aligned}
&\Big|\int_{\overline{M_lQ}}[(W_h^{(1)}-W_h^{(1)}(Q))\mathrm{d}y\\
&-(W_h^{(2)}-W_h^{(2)}(Q))\mathrm{d}x](u_{l+2}-u_{l+1})\Big|\\
\leq\ & Ch^3\Big(\Big|\frac{\partial u_h}{\partial x}\Big|+\Big|\frac{\partial u_h}{\partial y}\Big|\Big)^2\\
\leq\ & Ch\Big[\Big(\frac{\partial u_h}{\partial x}\Big)^2+\Big(\frac{\partial u_h}{\partial y}\Big)^2\Big]S_Q.
\end{aligned}
\tag{3.2.33}
$$

On $K_{P_l}^*\cap K$

$$u_h=u_h(P_l)+\frac{\partial u_h}{\partial x}(x-x_{P_l})+\frac{\partial u_h}{\partial y}(y-y_{P_l}),$$

$$|qu_h-q(P)u_h(P)|\leq|(q-q(P))u_h|+|q(P)(u_h-u_h(P))|,$$

$$|(q-q(P_l))u_h|\leq Ch\Big(|u_h(P_l)|+h\Big|\frac{\partial u_h}{\partial x}\Big|+h\Big|\frac{\partial u_h}{\partial y}\Big|\Big),$$

$$|q(P_l)(u_h-u_h(P_l))|\leq Ch\Big(\Big|\frac{\partial u_h}{\partial x}\Big|+\Big|\frac{\partial u_h}{\partial y}\Big|\Big),\ l=i,j,k.$$

So

$$
\begin{aligned}
&\Big|\int_{K_{P_l}^*\cap K}(qu_h-q(P_l)u_h(P_l))\mathrm{d}x\mathrm{d}y\cdot u_h(P_l)\Big|\\
\leq\ & Ch\Big[(u_l)^2+\Big(\frac{\partial u_h}{\partial x}\Big)^2+\Big(\frac{\partial u_h}{\partial y}\Big)^2\Big]S_Q.
\end{aligned}
\tag{3.2.34}
$$

It follows from (3.2.30), (3.2.33), (3.2.34) and Lemma 3.2.1 that

$$
\begin{aligned}
&|a(u_h,\Pi_h^*u_h)-a_h(u_h,\Pi_h^*u_h)|\\
=\ & \Big|\sum_{k\in T_h}[I_K(u_h,\Pi_h^*u_h)-\tilde{I}_K(u_h,\Pi_h^*u_h)]\Big|\\
\leq\ & Ch\|u_h\|_1^2.
\end{aligned}
\tag{3.2.35}
$$

Combining (3.2.29) and (3.2.35) leads to (3.2.24). □

From Theorem 3.2.1 it is easy to deduce the existence and uniqueness of the solution to the generalized difference scheme (3.2.13).

### 3.2.4 Error estimates

**Theorem 3.2.2** *Let $u$ be the generalized solution to (3.1.1) and $u_h$ the solution to the generalized difference scheme (3.2.13). If $u \in H^2(\Omega)$, then the following error estimate holds:*

$$||u - u_h||_1 \le Ch|u|_2. \tag{3.2.36}$$

**Proof** It is obvious that

$$a(u - u_h, \psi_{P_0}) = 0, \forall P_0 \in \dot{\Omega}_h, \tag{3.2.37}$$

which together with Theorem 3.2.1 yields

$$\begin{aligned} & ||u_h - \Pi_h u||_1^2 \\ \le\ & \frac{1}{\alpha} a(u_h - \Pi_h u_h, \Pi_h^*(u_h - \Pi_h u)) \\ =\ & \frac{1}{\alpha} a(u - \Pi_h u_h, \Pi_h^*(u_h - \Pi_h u)). \end{aligned}$$

So

$$||u_h - \Pi_h u||_1 \le \frac{1}{\alpha} \sup_{\overline{u}_h \in U_h} \frac{|a(u - \Pi_h u_h, \Pi_h^* \overline{u}_h)|}{||\overline{u}_h||_1}.$$

By (3.2.25) and (3.2.26) we have

$$a(u - \Pi_h u_h, \Pi_h^* \overline{u}_h) = \sum_{K \in T_h} I_K(u - \Pi_h u_h, \Pi_h^* \overline{u}_h), \tag{3.2.39}$$

$$\begin{aligned} & I_K(u - \Pi_h u_h, \Pi_h^* \overline{u}_h) \\ =\ & \sum_{l=i,j,k} \Big\{ \int_{\overline{M_l Q}} [\overline{W}_h^{(1)} \cos\langle n_l, x\rangle + \overline{W}_h^{(2)} \cos\langle n_l, y\rangle] \mathrm{d}s \\ & \cdot(\overline{u}_{l+2} - \overline{u}_{l+1}) + \int_{K_{P_l}^* \cap K} q(u - \Pi_h u) \mathrm{d}x \mathrm{d}y \cdot \overline{u}_h(P_l) \Big\}, \end{aligned} \tag{3.2.40}$$

where $\overline{W}_h^{(i)} = a_{i1} \frac{\partial(u - \Pi_h u)}{\partial x} + a_{i2} \frac{\partial(u - \Pi_h u)}{\partial y}$ $(i = 1, 2)$ and $n_l$ is the unit outer normal vector of $K_{P_l}^* \cap K$ along $\overline{M_l Q}$ $(l = i, j, k)$. It follows from (3.2.32) that

$$|\overline{u}_{l+2} - \overline{u}_{l+1}| \le h\Big(\Big|\frac{\partial \overline{u}_h}{\partial x}\Big| + \Big|\frac{\partial \overline{u}_h}{\partial y}\Big|\Big) \le C|\overline{u}_h|_{1,h,K}. \tag{3.2.41}$$

On the other hand,

$$\begin{aligned}&\Big|\int_{\overline{M_lQ}}[\overline{W}_h^{(1)}\cos\langle n_l,x\rangle+\overline{W}_h^{(2)}\cos\langle n_l,y\rangle]\mathrm{d}s\Big|\\ \le\ & C\int_{\overline{M_lQ}}\Big(\Big|\frac{\partial(u-\Pi_hu)}{\partial x}\Big|+\Big|\frac{\partial(u-\Pi_hu)}{\partial y}\Big|\Big)\mathrm{d}s\\ \le\ & Ch^{1/2}\Big\{\int_{\overline{M_lQ}}\Big(\Big|\frac{\partial(u-\Pi_hu)}{\partial x}\Big|^2+\Big|\frac{\partial(u-\Pi_hu)}{\partial y}\Big|^2\Big)\mathrm{d}s\Big\}^{1/2}.\end{aligned}\tag{3.2.42}$$

Set $\phi_1=\frac{\partial(u-\Pi_hu)}{\partial x}$ and $\phi_2=\frac{\partial(u-\Pi_hu)}{\partial y}$. The mapping $(x,y)\to(\lambda_j,\lambda_k)$ maps the element $K$ onto the reference element $\hat{K}$; the function $\phi_m$ on $K$ into the function $\hat{\phi}_m(\lambda_j,\lambda_k)=\phi_m(x,y)$ $(m=1,2)$; and the points $M_l,P_l,Q$ into $\hat{M}_l,\hat{P}_l,\hat{Q}$ $(l=i,j,k)$ respectively. It is obvious that

$$\int_{\overline{M_lQ}}|\phi_m|^2\mathrm{d}s\le h\int_{\overline{\hat{M}_l\hat{Q}}}|\hat{\phi}_m|^2\mathrm{d}\hat{s},\ m=1,2.$$

By the trace theorem on $\hat{K}^*_{P_{l+1}}\cap\hat{K}$ we have a constant $C>0$ independent of $K$ such that

$$\int_{\overline{\hat{M}_l\hat{Q}}}|\hat{\phi}_m|^2\mathrm{d}\hat{s}\le C\|\hat{\phi}_m\|^2_{1,\hat{K}},\ m=1,2.$$

Using Theorem 1.1.12 we have

$$|\hat{\phi}_m|_{0,\hat{K}}\le Ch^{-1}|\phi_m|_{0,K},\ \ |\hat{\phi}_m|_{1,\hat{K}}\le C|\phi_m|_{1,K},\ m=1,2.$$

Hence

$$\begin{aligned}&\int_{\overline{M_lQ}}|\phi_m|^2\mathrm{d}s\le Ch(h^{-1}|\phi_m|_{0,K}+|\phi_m|_{1,K})^2\\ \le\ & Ch(h^{-1}|u-\Pi_hu|_{1,K}+|u-\Pi_hu|_{2,K})^2\le Ch|u|^2_{2,K},\ m=1,2.\end{aligned}\tag{3.2.43}$$

It follows from (3.2.41)-(3.2.43) that

$$\begin{aligned}&\Big|\sum_{l=i,j,k}\int_{\overline{M_lQ}}[\overline{W}_h^{(1)}\cos\langle n_l,x\rangle+\overline{W}_h^{(2)}\cos\langle n_l,y\rangle]\mathrm{d}s(\overline{u}_{l+2}-\overline{u}_{l+1})\Big|\\ \le\ & Ch|u|_{2,K}|\overline{u}_h|_{1,K}.\end{aligned}\tag{3.2.44}$$

It is easy to deduce that

$$\begin{aligned}
&\Big|\sum_{l=i,j,k}\int_{K^*_{P_l}\cap K} q(u-\Pi_h u)\mathrm{d}x\mathrm{d}y\cdot \overline{u}_h(P_l)\Big| \\
\le\ & C\sum_{l=i,j,k}\int_{K^*_{P_l}\cap K} |u-\Pi_h u|\mathrm{d}x\mathrm{d}y\cdot |\overline{u}_h(P_l)| \qquad (3.2.45)\\
\le\ & Ch^2|u|_{2,K}|\overline{u}_h|_{0,K}.
\end{aligned}$$

It follows from (3.2.39), (3.2.40), (3.2.44) and (3.2.45) that

$$|a(u-\Pi_h u, \Pi_h^*\overline{u}_h)| \le Ch|u|_2||\overline{u}_h||_1. \qquad (3.2.46)$$

A combination of (3.2.38) and (3.2.46) leads to

$$||u_h-\Pi_h u||_1 \le Ch|u|_2.$$

This together with (3.2.8) implies (3.2.36) and completes the proof. □

## 3.3 Generalized Difference Methods on Quadrilateral Meshes

### 3.3.1 Trial and test function spaces

Suppose $\Omega$ is a polygonal region, of which the boundary $\partial\Omega$ is a simple closed fold line. Divide $\overline{\Omega}$ into a sum of finite number of strictly convex quadrilaterals such that different quadrilaterals have no common interior point, that a vertex of any quadrilateral does not lie on an interior of a side of any other quadrilateral and that any vertex of the boundary is a vertex of some quadrilateral. Each quadrilateral is called an element and denoted by $K$. All the elements constitute a quadrilateral decomposition of $\overline{\Omega}$, denoted by $T_h$ where $h$ is the largest diameter of all the quadrilaterals. The vertexes of the quadrilaterals are called the nodes of the decomposition. Two nodes are adjacent if they are the two endpoints of a certain side of an element. Two elements are adjacent if they share a common side.

Next we construct the dual decomposition related to $T_h$. As in Fig. 3.3.1, let $P_0$ be a node of the decomposition $T_h$, $P_i$ $(i = 1, \cdots, 4)$ the adjacent nodes of $P_0$, $M_i$ the midpoint of $\overline{P_0P_i}$, and $P_{i,i+1}$ (convention: $P_{45} = P_{41} = P_{14}$) the vertex facing $P_0$, in the quadrilateral with sides $\overline{P_0P_i}$ and $\overline{P_0P_{i+1}}$. Take any point $Q_i$ in the quadrilateral $P_0P_iP_{i,i+1}P_{i+1}$ $(P_5 = P_1)$, connect successively $M_1, Q_1, M_2, Q_2, \cdots,$ $M_4, Q_4, M_1$ to form a polygonal region $K^*_{P_0}$, called a dual element. All the dual elements constitute a new decomposition, the dual decomposition, $T^*_h$ of $\overline{\Omega}$. $Q_i$ and $M_i$ are called the nodes of the dual decomposition. The most important dual decomposition takes $Q_i$ as the joint of the two lines connecting the midpoints of the opposite sides of the quadrilateral element. This is called the central dual decomposition.

As in §3.2, let $\overline{\Omega}_h$ be the node set of the decomposition $T_h$ and $\dot{\Omega}_h = \overline{\Omega}_h \setminus \partial\Omega$ the set of all interior nodes. $\Omega^*_h$ denotes the node set of the dual decomposition $T^*_h$. For $Q \in \Omega^*_h$, we denote by $K_Q$ the quadrilateral element containing $Q$. $S_Q$ (or $S_{K_Q}$) and $S^*_{P_0}$ stand for the areas of the quadrilateral element $K_Q$ and the dual element $K^*_{P_0}$ respectively. We shall always assume that $T_h$ and $T^*_h$ are quasi-uniform such that there exist constants $c_1, c_2, c_3 > 0$ independent of $h$ such that

$$c_1h^2 \leq S_Q \leq h^2, \ Q \in \Omega^*_h, \tag{3.3.1a}$$

$$c_2h^2 \leq S^*_{P_0} \leq c_3h^2, \ P_0 \in \overline{\Omega}_h. \tag{3.3.1b}$$

We point out that for the central dual decomposition, (3.3.1b) can be deduced from (3.3.1a).

The trial function space $U_h$ is chosen as the isoparametric element space of the bilinear functions on a quadrilateral decomposition $T_h$ (cf. [B-17]). Its construction is as follows. Take the unit square $\hat{K} = \{(\xi, \eta) : \ 0 \leq \xi, \eta \leq 1\}$ of the $(\xi, \eta)$ plane as a reference element. For a quadrilateral element $K_Q$, suppose its vertexes are $P_i(x_i, y_i)$ $(i = 1, \cdots, 4)$. Then there exists a unique invertible bilinear mapping

$$F_{K_Q} : \begin{cases} x = x_1 + a_1\xi + a_2\eta + a_3\xi\eta, \\ y = y_1 + b_1\xi + b_2\eta + b_3\xi\eta, \end{cases} \tag{3.3.2}$$

where

$$a_1 = x_2 - x_1,\ a_2 = x_3 - x_1,\ a_3 = x_4 - x_3 - x_2 + x_1,$$
$$b_1 = y_2 - y_1,\ b_2 = y_3 - y_1,\ b_3 = y_4 - y_3 - y_2 + y_1,$$

which maps $\hat{K}$ onto $K_Q$ (cf. Fig. 3.3.2).

For any $u_h \in U_h$ we have on $K_Q$ that

$$\begin{aligned} u_h = &\ P_{\hat{K}}(\xi, \eta) \\ = &\ u_1(1-\xi)(1-\eta) + u_2\xi(1-\eta) + u_3(1-\xi)\eta + u_4\xi\eta \\ = &\ u_1 + (u_2 - u_1)\xi + (u_3 - u_1)\eta + (u_4 - u_3 - u_2 + u_1)\xi\eta. \end{aligned} \tag{3.3.3}$$

So we have

$$U_h = \{u_h \in C(\bar{\Omega}) : u_h\Big|_{K_Q} = P_{\hat{K}} \circ F_{K_Q}^{-1}, u_h\Big|_{\partial\Omega} = 0, P_{\hat{K}} \in \mathcal{P}_{11}\},$$

where $\mathcal{P}_{11}$ is the family of bilinear functions.

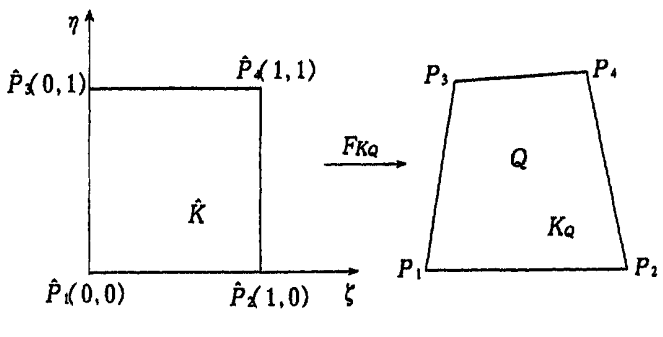



Fig. 3.3.2

The finite element obtained through the transformation $F_{K_Q}$ is called the quadrilateral isoparametric element with four nodes. If $u \in H^2(\Omega)$ and $\Pi_h u$ is its interpolation projection onto $U_h$, then

$$|u - \Pi_h u|_m \le Ch^{2-m}|u|_2,\ m = 0, 1. \tag{3.3.4}$$

The test space $V_h$ is chosen as the piecewise constant function space related to the dual decomposition, with the following basis functions: For each $P_0 \in \dot{\Omega}_h$,

$$\psi_{P_0}(P) = \begin{cases} 1, & P \in K^*_{P_0}, \\ 0, & P \notin K^*_{P_0}. \end{cases}$$

So $\psi_{P_0}$ is the characteristic function of the dual element $K^*_{P_0}$. For any $v_h \in V_h$ we have

$$v_h = \sum_{P_0 \in \dot{\Omega}_h} v_h(P_0)\psi_{P_0}. \tag{3.3.5}$$

### 3.3.2 Generalized difference equation

The generalized difference equation corresponding to the above trial and test function spaces is: Find $u_h \in U_h$ such that

$$a(u_h, \psi_{P_0}) = (f, \psi_{P_0}), \ \forall P_0 \in \dot{\Omega}_h, \tag{3.3.6}$$

where

$$a(u_h, \psi_{P_0}) = -\int_{\partial K^*_{P_0}} (W_h^{(1)} \mathrm{d}y - W_h^{(2)} \mathrm{d}x) + \int_{K^*_{P_0}} q u_h \mathrm{d}x\mathrm{d}y, \tag{3.3.7}$$

where $\partial K^*_{P_0}$ is the boundary of $K^*_{P_0}$, possessing a counterclockwise direction, and

$$W_h^{(i)} = a_{i1}\frac{\partial u_h}{\partial x} + a_{i2}\frac{\partial u_h}{\partial y}, \ i = 1, 2.$$

Using different numerical integration formulas to compute the integrals of the right-hand side of (3.3.7) leads to different approximations $a_h(u_h, \psi_{P_0})$ of $a(u_h, \psi_{P_0})$ and results in different difference equations:

$$a_h(u_h, \psi_{P_0}) = (f, \psi_{P_0}), \ P_0 \in \dot{\Omega}_h. \tag{3.3.8}$$

Let the dual decomposition be as in Fig. 3.3.1. Then the first integral of the right-hand side of (3.3.7) can be divided into a sum of line integrals along $\overline{M_1Q_1}$, $\overline{Q_1M_2}, \cdots, \overline{Q_4M_1}$. For each line integral on

$\overline{M_jQ_j}$ (or $\overline{Q_jM_{j+1}}$), if we use its value at $M_i$, or $Q_i$, or their average respectively to replace the integral function, then we obtain three approximations $a_h^k(u_h, \psi_{P_0})$ ($k = 1, 2, 3$) of $a(u_h, \psi_{P_0})$, ending up with three difference equations. One can also divide the line integral in (3.3.7) into a sum of integrals on $\overline{M_1Q_1M_2}$, $\overline{M_2Q_2M_3}$, $\overline{M_3Q_3M_4}$ and $\overline{M_4Q_4M_1}$, and approximate $W_h^{(i)}$ by $W_h^{(i)}(Q_j)$ ($i = 1, 2,\ j = 1, 2, 3, 4$), then we have the following difference scheme (cf. [B-32]):

$$\begin{aligned}
&-W_h^{(1)}(Q_1)(y_{M_2} - y_{M_1}) - W_h^{(1)}(Q_2)(y_{M_3} - y_{M_2}) \\
&-W_h^{(1)}(Q_3)(y_{M_4} - y_{M_3}) - W_h^{(1)}(Q_4)(y_{M_1} - y_{M_4}) \\
&+W_h^{(2)}(Q_1)(x_{M_2} - x_{M_1}) + W_h^{(2)}(Q_2)(x_{M_3} - x_{M_2}) \\
&+W_h^{(2)}(Q_3)(x_{M_4} - x_{M_3}) + W_h^{(2)}(Q_4)(x_{M_1} - x_{M_4}) \\
&+q(P_0)u_h(P_0)S_{P_0}^* \\
= \ & f(P_0)S_{P_0}^*, \quad \forall P_0 \in \dot{\Omega}_h.
\end{aligned} \tag{3.3.9}$$

Define the following difference operators

$$\begin{aligned}
(\nabla_1\phi)_{P_0} = \ & [\phi(Q_1)(y_{M_2} - y_{M_1}) + \phi(Q_2)(y_{M_3} - y_{M_2}) \\
& +\phi(Q_3)(y_{M_4} - y_{M_3}) + \phi(Q_4)(y_{M_1} - y_{M_4})]/S_{P_0}^*, \\
(\nabla_2\phi)_{P_0} = \ & [\phi(Q_1)(x_{M_1} - x_{M_2}) + \phi(Q_2)(x_{M_2} - x_{M_3}) \\
& +\phi(Q_3)(x_{M_3} - x_{M_4}) + \phi(Q_4)(x_{M_4} - x_{M_1})]/S_{P_0}^*.
\end{aligned}$$

Then (3.3.9) can be rewritten as a conservation form:

$$-\sum_{i=1}^{2}\Big(\nabla_i\Big(\sum_{j=1}^{2} a_{ij}\frac{\partial u_h}{\partial x_j}\Big)\Big)_{P_0} + q(P_0)u_h(P_0) = f(P_0), \ \forall P_0 \in \dot{\Omega}_h, \tag{3.3.9$'$}$$

where $\frac{\partial}{\partial x_1} = \frac{\partial}{\partial x}$ and $\frac{\partial}{\partial x_2} = \frac{\partial}{\partial y}$.

If $T_h$ is a rectangular decomposition and the sides of the rectangles are parallel to the coordinate axes, then the dual decomposition is also a rectangular decomposition (cf. Fig. 3.3.3), and the above

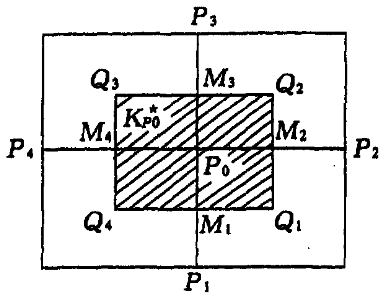



Fig. 3.3.3

mentioned three kinds of numerical integrations lead to the following three difference schemes respectively.

**Scheme I:**

$$\begin{aligned}
&-[W_h^{(1)}(M_2^-) - W_h^{(1)}(M_4^-)](y_{P_0} - y_{M_1}) \\
&-[W_h^{(1)}(M_2^+) - W_h^{(1)}(M_4^+)](y_{M_3} - y_{P_0}) \\
&+[W_h^{(2)}(M_1^-) - W_h^{(2)}(M_3^-)](x_{P_0} - x_{M_4}) \\
&+[W_h^{(2)}(M_1^+) - W_h^{(2)}(M_3^+)](x_{M_2} - x_{P_0}) \\
&+q(P_0)u_h(P_0)S^*_{P_0} = f(\overline{P})S^*_{P_0}, \quad \forall P_0 \in \dot{\Omega}_h,
\end{aligned}$$

where $W_h^{(1)}(M_i^-)$ and $W_h^{(1)}(M_i^+)$ $(i = 2,4)$ denote the single-sided limits of $W_h^{(1)}$ at $M_i$ from left and right sides respectively along $\overline{Q_iQ_{i+1}}$ $(i = 1,3)$; $W_h^{(2)}(M_i^-)$ and $W_h^{(2)}(M_i^+)$ $(i = 1,3)$ stand for the single-sided limits of $W_h^{(2)}$ at $M_i$ from down side and upside respectively along $\overline{Q_4Q_1}$ and $\overline{Q_2Q_3}$; $\overline{P}$ can be viewed as an averaging center of the rectangle $K^*_P$; and the meanings of the other notations are self-evident. In particular, Scheme I becomes the well-known five-point difference scheme when $A$ is the Laplacian operator.

**Scheme II:**

$$-[W_h^{(1)}(Q_1) - W_h^{(1)}(Q_4)](y_{P_0} - y_{M_1})$$
$$-[W_h^{(1)}(Q_2) - W_h^{(1)}(Q_3)](y_{M_3} - y_{P_0})$$
$$+[W_h^{(2)}(Q_4) - W_h^{(2)}(Q_3)](x_{P_0} - x_{M_4})$$
$$+[W_h^{(2)}(Q_1) - W_h^{(2)}(Q_2)](x_{M_2} - x_{P_0})$$
$$+q(P_0)u_h(P_0)S_{P_0}^* = f(\overline{P})S_{P_0}^*, \quad \forall P_0 \in \dot{\Omega}_h,$$

**Scheme III:**

$$\frac{1}{2}[W_h^{(1)}(Q_1) + W_h^{(1)}(M_2^-) - W_h^{(1)}(Q_4) + W_h^{(1)}(M_4^-)](y_{P_0} - y_{M_1})$$
$$-\tfrac{1}{2}[W_h^{(1)}(Q_2) + W_h^{(1)}(M_2^+) - W_h^{(1)}(Q_3) - W_h^{(1)}(M_4^+)](y_{M_3} - y_{P_0})$$
$$+\frac{1}{2}[W_h^{(2)}(Q_4) + W_h^{(2)}(M_1^-) - W_h^{(2)}(Q_3) - W_h^{(2)}(M_3^-)](x_{P_0} - x_{M_4})$$
$$+\frac{1}{2}[W_h^{(2)}(Q_1) + W_h^{(2)}(M_1^+) - W_h^{(2)}(Q_2) - W_h^{(2)}(M_3^+)](x_{M_2} - x_{P_0})$$
$$+\frac{1}{4}\sum_{i=1}^{4} q(Q_i)u_h(Q_i)S_{P_0}^* = f(\overline{P})S_{P_0}^*, \quad \forall P_0 \in \dot{\Omega}_h.$$

### 3.3.3 Convergence order estimates

Suppose that $T_h$ and $T_h^*$ are a quasi-uniform arbitrary quadrilateral grid and its central dual grid respectively, and that $u$ and $u_h$ are the solutions to the Poisson equation and the corresponding generalized difference scheme (3.3.6) respectively. Then under certain geometrical restrictions on the quadrilateral grid, there holds the following error estimate (see [B-62])

$$\|u - u_h\|_1 \leq Ch|u|_2. \tag{3.3.10}$$

In the case of rectangular grid and under stronger assumptions on the smoothness of the solutions, a higher order convergence estimate, namely a superconvergence result, in a discrete norm as follows can be obtained. (See [A-62] for details.)

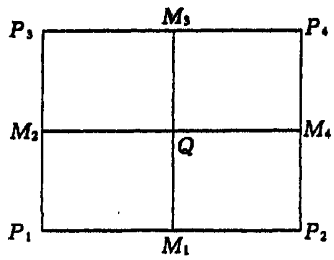



Fig. 3.3.4

**Theorem 3.3.1** *Let $T_h$ and $T_h^*$ be a quasi-uniform rectangular grid and its central dual grid respectively, and let $u$ and $u_h$ be the solutions to (3.1.1) and the difference scheme (3.3.6) respectively. If $u \in C^3(\overline{\Omega})$, then the following error estimate holds:*

$$||u - u_h||_{1,h} \le Ch^2 M_{23}, \tag{3.3.11}$$

*where $M_{23} = \max\{|D^2 u|_{\max}, |D^3 u|_{\max}\}$, and the discrete norm is defined by (cf. Fig. 3.3.4)*

$$||u||_{1,h} = (|u|_{0,h}^2 + |u|_{1,h}^2)^{1/2},$$

$$|u|_{m,h} = \Big(\sum_{K \in T_h} |u|_{m,h,K}^2\Big)^{1/2}, \; m = 0, 1,$$

$$|u|_{0,h,K} = \Big\{\frac{1}{4}[u^2(P_1) + u^2(P_2) + u^2(P_3) + u^2(P_4)]S_Q\Big\}^{1/2},$$

$$\begin{aligned} |u|_{1,h,K} = \; & \Big\{\Big[\Big(\frac{\partial u(M_1)}{\partial x}\Big)^2 + \Big(\frac{\partial u(M_3)}{\partial x}\Big)^2 \\ & + \Big(\frac{\partial u(M_2)}{\partial y}\Big)^2 + \Big(\frac{\partial u(M_4)}{\partial y}\Big)^2\Big] S_Q\Big\}^{1/2}. \end{aligned}$$

For the above Schemes I and II, if the difference of the squares of any two successive step-lengths on $x$ and $y$ directions is $O(h^{2+d})$, we have the following error estimate:

$$||u - u_h||_{1,h} \le Ch^{1+d} \max_{1 \le l \le 4} |D^l u|_{\max}, \; 0 \le d \le 1. \tag{3.3.12}$$

For Scheme III, the following error estimate holds for the quasi-uniform rectangular mesh:

$$||u - u_h||_{1,h} \le Ch^2 \max_{1\le l\le 4} |D^l u|_{\max}. \tag{3.3.13}$$

For Scheme (3.3.9) on arbitrary quadrilateral meshes, the convergence order is $O(h)$ under suitable assumptions on the decomposition. (cf. [B-32] and [C-7]).

## 3.4 Quadratic Element Difference Schemes

The following two sections will be devoted to the generalized difference methods based on higher order elements. For simplicity, we take the boundary value problem of the Poisson equation as an example to illustrate the idea.

Let $\Omega$ be a planar polygonal region with boundary $\partial\Omega$ and $f \in L^2(\Omega)$. Consider the first boundary value problem of the Poisson equation:

$$\begin{cases} -\Delta u = f, \text{ in } \Omega, & (3.4.1a) \\ u|_{\partial\Omega} = 0. & (3.4.1b) \end{cases}$$

The corresponding variational problem is: Find $u \in H_0^1(\Omega)$ such that

$$a(u, v) = (f, v), \ \forall v \in H_0^1(\Omega), \tag{3.4.2}$$

where

$$a(u, v) = \int_\Omega \Big(\frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y}\Big) dx dy. \tag{3.4.3}$$

### 3.4.1 Trial and test function spaces

As in §3.2, let $T_h$ be a quasi-uniform triangulation of $\overline{\Omega}$. $T_h$ consists of finite number of triangular elements $K_Q$, $Q$ being the barycenter of the triangle. The vertexes of the triangles and the midpoints of the sides are taken as the nodes. $\overline{\Omega}_h$ denotes the set of the vertexes of all the triangular elements, $\overline{M}_h$ the set of the midpoints of the sides

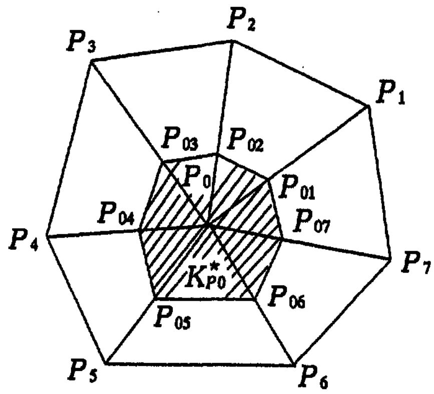



Fig. 3.4.1

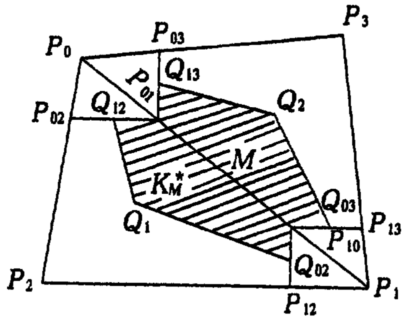



Fig. 3.4.2

of all the elements, $\Omega_h^*$ the set of the barycenters of all the elements, and $\dot{\Omega}_k = \overline{\Omega}_h \setminus \partial\Omega$, $\dot{M}_h = \overline{M}_h \setminus \partial\Omega$.

The dual decomposition of $T_h$ is denoted by $T_h^*$, consisting of the polygons $K_{P_0}^*$ surrounding the node $P_0 \in \overline{\Omega}_h$ and $K_M^*$ surrounding $M \in \overline{M}_h$. These small polygons are called dual elements. Their detailed construction is as follows.

1) Construction of $K_{P_0}^*$. Suppose that $P_0 \in \overline{\Omega}_h$, that $P_i$ $(i = 1, 2, \cdots, 7)$ are its adjacent vertexes, and that $P_{0i}$ is a point on $\overline{P_0P_i}$ such that $\overline{P_0P_{0i}} = \frac{1}{3}\overline{P_0P_i}$. Connect successively $P_{0i}$ $(i = 1, 2, \cdots, 7)$ to obtain a polygon $K_{P_0}$ surrounding $P_0$. (See Fig. 3.4.1.)

2) Construction of $K_M^*$. Let $M \in \overline{M}_h$ be a midpoint of a common side of two adjacent triangular elements $K_{Q_1} = \triangle P_0P_1P_2$ and $K_{Q_2} = \triangle P_0P_1P_3$. Denote by $Q_{12}, Q_{13}, Q_{02}, Q_{03}$ the midpoints of $\overline{P_{01}P_{02}}, \overline{P_{01}P_{03}}, \overline{P_{10}P_{12}}$ and $\overline{P_{10}P_{13}}$ respectively. A polygon $K_M^*$ surrounding $M$ is obtained by connecting successively $P_{10}, Q_{03}, Q_2, Q_{13}, P_{01}, Q_{12}, Q_1, Q_{02}, P_{10}$ (see Fig. 3.4.2).

The trial space $U_h$ is chosen as the Lagrangian quadratic element space related to the triangulation $T_h$. For each $P_0 \in \dot{\Omega}_h$ and $M_0 \in \dot{M}_h$, the corresponding basis functions are the piecewise quadratic polynomials satisfying the following interpolation conditions respectively:

$$\phi_{P_0}(P) = \begin{cases} 1, & P = P_0, \\ 0, & P \in \overline{\Omega}_h \cup \overline{M}_h \setminus \{P_0\}, \end{cases} \tag{3.4.4a}$$

$$\phi_{M_0}(P) = \begin{cases} 1, & P = M_0, \\ 0, & P \in \overline{\Omega}_h \cup \overline{M}_h \setminus \{M_0\}. \end{cases} \tag{3.4.4b}$$

So $U_h = \text{span}\{\phi_{P_0}, \phi_{M_0};\ P_0 \in \dot{\Omega}_h, M \in \dot{M}_h\}$.

The test function space $V_h$ is taken as the piecewise constant function space related to the dual decomposition $T_h^*$. For each $P_0 \in \dot{\Omega}_h$ and $M_0 \in \dot{M}_h$, the corresponding basis functions are the characteristic functions of $K_{P_0}^*$ and $K_{M_0}^*$ respectively:

$$\psi_{P_0}(P) = \begin{cases} 1, & P \in K_{P_0}^*, \\ 0, & P \notin K_{P_0}^*, \end{cases} \tag{3.4.5a}$$

$$\psi_{M_0}(P) = \begin{cases} 1, & P \in K_{M_0}^*, \\ 0, & P \notin K_{M_0}^*. \end{cases} \tag{3.4.5b}$$

Hence $V_h = \text{span}\{\psi_{P_0}, \psi_{M_0};\ P_0 \in \dot{\Omega}_h, M \in \dot{M}_h\}$.

### 3.4.2 Generalized difference equation

The quadratic element difference scheme corresponding to $U_h$ and $V_h$ constructed above is: Find $u_h \in U_h$ such that

$$a(u_h, v_h) = (f, v_h),\ \forall v_h \in V_h, \tag{3.4.6}$$

or equivalently

$$\begin{cases} a(u_h, \psi_{P_0}) = (f, \psi_{P_0}), & P_0 \in \dot{\Omega}_h, & (3.4.6\text{a})' \\ a(u_h, \psi_M) = (f, \psi_M), & M \in \dot{M}_h, & (3.4.6\text{b})' \end{cases}$$

where

$$a(u_h, v_h) = \sum_{P_0 \in \dot{\Omega}_h} v_h(P_0) a(u_h, \psi_{P_0}) + \sum_{M \in \dot{M}_h} v_h(M) a(u_h, \psi_M), \tag{3.4.7a}$$

$$a(u_h, \psi_{P_0}) = -\int_{\partial K^*_{P_0}} \frac{\partial u_h}{\partial x}\mathrm{d}y - \frac{\partial u_h}{\partial y}\mathrm{d}x, \tag{3.4.7b}$$

$$a(u_h, \psi_M) = -\int_{\partial K^*_M} \frac{\partial u_h}{\partial x}\mathrm{d}y - \frac{\partial u_h}{\partial y}\mathrm{d}x. \tag{3.4.7c}$$

Note (cf. Figs. 3.4.1 and 3.4.2) that $\partial K^*_{P_0} = \overline{P_{01}P_{02}} \cup \overline{P_{02}P_{03}} \cup \cdots \cup \overline{P_{07}P_{01}}$ and $\partial K^*_M = \overline{Q_1Q_{02}} \cup \overline{Q_{02}P_{10}} \cup \cdots \cup \overline{Q_{12}Q_1}$. The right-hand side integrals of (3.4.7b) and (3.4.7c) can be divided into a sum of the easy-to-compute integrals on these segments, resulting in a linear algebraic system with unknowns $u_h(P_0)$ ($P_0 \in \dot{\Omega}_h$) and $u_h(M)$ ($M \in \dot{M}_h$).

There are two approaches to form the generalized difference equation: Directly compute the equation for each node, or first compute the *stiff* matrix (see (3.4.13) and (3.4.17) below) for each element and then form the whole matrix of the equation by summation of all the stiff matrices. The latter approach is more convenient and suitable by computer, especially for two-dimensional high order difference schemes and irregular meshes.

Take any a triangular element $K_Q$. Let $P_l(x_l, y_l)$ ($l = i, j, k$, counterclockwise) be the vertexes, $M_i$ the midpoint of $\overline{P_jP_k}$, $P_{ij}$ the point on $\overline{P_iP_j}$ such that $|\overline{P_iP_{ij}}| = \frac{1}{3}|\overline{P_iP_j}|$, $Q_i$ the midpoint of $\overline{P_{ij}P_{ik}}$ etc. (cf. Fig. 3.4.3).

For $u_h \in U_h$, write $u_P = u_h(P)$. Then on $K_Q$

$$u_h = \sum_{l=i,j,k} (u_{P_l}\phi_{P_l} + u_{M_l}\phi_{M_l}).$$

Perform a linear transformation

$$\lambda_j = \frac{1}{2S_Q}\begin{vmatrix} 1 & x & y \\ 1 & x_k & y_k \\ 1 & x_i & y_i \end{vmatrix}, \ \lambda_k = \frac{1}{2S_Q}\begin{vmatrix} 1 & x & y \\ 1 & x_i & y_i \\ 1 & x_j & y_j \end{vmatrix}. \tag{3.4.8}$$

Then $K_Q$ is transformed into a reference element $\hat{K}_Q$ with vertexes $\hat{P}_i(0,0)$, $\hat{P}_j(1,0)$, $\hat{P}_k(0,1)$ on $\lambda_j\lambda_k$ plane; $M_i, P_{ij}, Q_i$ become $\hat{M}_i, \hat{P}_{ij}$,

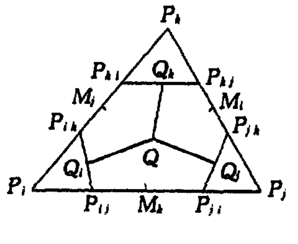



**Fig. 3.4.3**

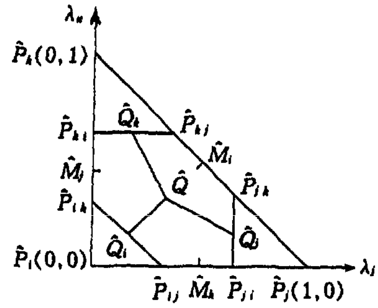



**Fig. 3.4.4**

$\hat{Q}_i$ etc. (cf. Fig. 3.4.4); and

$$\begin{aligned} u_h = \; & u_{P_i}(1-\lambda_j-\lambda_k)(1-2\lambda_j-2\lambda_k)+u_{P_j}(2\lambda_j-1)\lambda_j \\ & +u_{P_k}(2\lambda_k-1)\lambda_k+4u_{M_i}\lambda_j\lambda_k \\ & +4u_{M_j}(1-\lambda_j-\lambda_k)\lambda_k+4u_{M_k}(1-\lambda_j-\lambda_k)\lambda_j. \end{aligned} \tag{3.4.9}$$

By (3.4.8) and (3.4.9) we have

$$\frac{\partial u_h}{\partial x}=\frac{1}{2S_Q}\Big[\frac{\partial u_h}{\partial \lambda_j}(y_k-y_i)-\frac{\partial u_h}{\partial \lambda_k}(y_j-y_i)\Big], \tag{3.4.10a}$$

$$\frac{\partial u_h}{\partial y}=\frac{1}{2S_Q}\Big[-\frac{\partial u_h}{\partial \lambda_j}(x_k-x_i)+\frac{\partial u_h}{\partial \lambda_k}(x_j-x_i)\Big], \tag{3.4.10b}$$

$$\begin{aligned} \frac{\partial u_h}{\partial \lambda_j}= \; & u_{P_i}(4\lambda_j+4\lambda_k-3)+u_{P_j}(4\lambda_j-1) \\ & +4u_{M_i}\lambda_k-4u_{M_j}\lambda_k+4u_{M_k}(1-2\lambda_j-\lambda_k), \end{aligned} \tag{3.4.11a}$$

$$\begin{aligned} \frac{\partial u_h}{\partial \lambda_k}= \; & u_{P_i}(4\lambda_j+4\lambda_k-3)+u_{P_k}(4\lambda_k-1) \\ & +4u_{M_i}\lambda_j+4u_{M_j}(1-\lambda_j-2\lambda_k)-4u_{M_k}\lambda_j. \end{aligned} \tag{3.4.11b}$$

The bilinear form $a(u_h, v_h)$ on $U_h \times V_h$ reads

$$a(u_h, v_h) = \sum_{K \in T_h} I_K(u_h, v_h), \tag{3.4.12}$$

where

$$\begin{aligned} I_K(u_h, v_h) = & \sum_{l=i,j,k} \Big[ v_{P_l} \int_{L_1} \Big( -\frac{\partial u_h}{\partial x} \mathrm{d}y + \frac{\partial u_h}{\partial y} \mathrm{d}x \Big) \\ & + v_{M_l} \int_{L_2} \Big( -\frac{\partial u_h}{\partial x} \mathrm{d}y + \frac{\partial u_h}{\partial y} \mathrm{d}x \Big) \Big], \end{aligned} \tag{3.4.13}$$

where $L_1 = \overline{P_{l,l+1} P_{l,l+2}}$, $L_2 = \overline{P_{l+2,l+1} Q_{l+2} Q Q_{l+1} P_{l+1,l+2}}$, $i+1 = j$, $j+1 = k$, $k+1 = i$. It follows from (3.4.8) that

$$\left\{ \begin{aligned} \mathrm{d}\lambda_j &= \frac{1}{2S_Q} [(y_k - y_i)\mathrm{d}x - (x_k - x_i)\mathrm{d}y], \\ \mathrm{d}\lambda_k &= \frac{1}{2S_Q} [(y_j - y_i)\mathrm{d}x - (x_j - x_i)\mathrm{d}y], \end{aligned} \right. \tag{3.4.14}$$

$$\left\{ \begin{aligned} \mathrm{d}x &= (x_j - x_i)\mathrm{d}\lambda_j + (x_k - x_i)\mathrm{d}\lambda_k, \\ \mathrm{d}y &= (y_j - y_i)\mathrm{d}\lambda_j + (y_k - y_i)\mathrm{d}\lambda_k. \end{aligned} \right. \tag{3.4.15}$$

So

$$\begin{aligned} & \int_L \Big( -\frac{\partial u_h}{\partial x} \mathrm{d}y + \frac{\partial u_h}{\partial y} \mathrm{d}x \Big) \\ = & \frac{1}{2S_Q} \int_{\hat{L}} \Big\{ \Big[ -\frac{\partial u_h}{\partial \lambda_j} (y_k - y_i) + \frac{\partial u_h}{\partial \lambda_k} (y_j - y_i) \Big] \Big[ (y_j - y_i)\mathrm{d}\lambda_j \\ & + (y_k - y_i)\mathrm{d}\lambda_k \Big] + \Big[ -\frac{\partial u_h}{\partial \lambda_j} (x_k - x_i) \\ & + \frac{\partial u_h}{\partial \lambda_k} (x_j - x_i) \Big] \Big[ (x_j - x_i)\mathrm{d}\lambda_j + (x_k - x_i)\mathrm{d}\lambda_k \Big] \Big\} \\ = & \frac{1}{2S_Q} \int_{\hat{L}} \Big[ -a^2 \frac{\partial u_h}{\partial \lambda_j} \mathrm{d}\lambda_k + b^2 \frac{\partial u_h}{\partial \lambda_k} \mathrm{d}\lambda_j \\ & + \frac{a^2 + b^2 - c^2}{2} \Big( -\frac{\partial u_h}{\partial \lambda_j} \mathrm{d}\lambda_j + \frac{\partial u_h}{\partial \lambda_k} \mathrm{d}\lambda_k \Big) \Big], \end{aligned} \tag{3.4.16}$$

where $a = |P_iP_k|$, $b = |P_iP_j|$, $c = |P_jP_k|$ and $\hat{L}$ is the image of $L$ by transformation (3.4.8). Using (3.4.16) to compute (3.4.13) results in

$$\begin{aligned} I_K(u_h, v_h) = & \frac{1}{36S_Q}[v_{P_i}, v_{P_j}, v_{P_k}, v_{M_i}, v_{M_j}, v_{M_k}]A \\ & \cdot[v_{P_i}, v_{P_j}, v_{P_k}, v_{M_i}, v_{M_j}, v_{M_k}]^T, \end{aligned} \tag{3.4.17}$$

where $A = [a_{ij}]$ is a $6\times6$ matrix with

$$\begin{aligned}
a_{11} &= 10c^2,\ a_{12} = a_{21} = a^2 - b^2 + c^2, \\
a_{13} &= a_{31} = -a^2 + b^2 + c^2,\ a_{14} = -4c^2, \\
a_{15} &= a_{46} = 8a^2 - 8b^2 - 4c^2,\ a_{16} = a_{45} = -8a^2 + 8b^2 - 4c^2, \\
a_{22} &= 10a^2,\ a_{23} = a_{32} = a^2 + b^2 - c^2, \\
a_{24} &= a_{56} = -4a^2 - 8b^2 + 8c^2,\ a_{25} = -4a^2, \\
a_{26} &= a_{54} = -4a^2 + 8b^2 - 8c^2,\ a_{33} = 10b^2, \\
a_{34} &= a_{65} = -8a^2 - 4b^2 + 8c^2,\ a_{35} = a_{64} = 8a^2 - 4b^2 - 8c^2, \\
a_{36} &= -4b^2,\ a_{41} = -2c^2,\ a_{42} = -5a^2 - 3b^2 + 3c^2, \\
a_{43} &= -3a^2 - 5b^2 + 3c^2,\ a_{44} = 8a^2 + 8b^2 + 4c^2, \\
a_{51} &= 3a^2 - 3b^2 - 5c^2,\ a_{52} = -2a^2,\ a_{53} = 3a^2 - 5b^2 - 3c^2, \\
a_{55} &= 4a^2 + 8b^2 + 8c^2,\ a_{61} = -3a^2 + 3b^2 - 5c^2, \\
a_{62} &= -5a^2 + 3b^2 - 3c^2,\ a_{63} = -2b^2,\ a_{66} = 8a^2 + 4b^2 + 8c^2.
\end{aligned}$$

Here $\frac{1}{36S_Q}A$ is the stiff matrix of the element.

### 3.4.3 *a priori* estimates

Let us introduce the discrete semi- and full-norms:

$$||u_h||_{0,h} = \Big(\sum_{K\in T_h} |u_h|^2_{0,h,K}\Big)^{1/2}, \tag{3.4.18}$$

$$|u_h|_{1,h} = \Big(\sum_{K\in T_h} |u_h|^2_{1,h,K}\Big)^{1/2}, \tag{3.4.19}$$

$$||u_h||_{1,h} = (||u_h||_{0,h}^2 + |u_h|_{1,h}^2)^{1/2}, \tag{3.4.20}$$

where

$$\begin{aligned} |u_h|_{0,h,K} = & \ [(u_{P_i}^2 + u_{P_j}^2 + u_{P_k}^2 + u_{M_i}^2 + u_{M_j}^2 + u_{M_k}^2)S_Q/6]^{1/2}, \\ |u_h|_{1,h,K} = & \ [(u_{P_i} - u_{M_i})^2 + (u_{P_j} - u_{M_j})^2 + (u_{P_k} - u_{M_k})^2 \\ & +(u_{M_i} - u_{M_j})^2 + (u_{M_i} - u_{M_k})^2]^{1/2}. \end{aligned}$$

**Lemma 3.4.1** *On the space $U_h$, $||\cdot||_{0,h}$ is equivalent with the $L^2$-norm $||\cdot||_0$, and $|\cdot|_{1,h}$ is equivalent with the $H^1$−semi-norm (and hence with the $H_1$−norm $||\cdot||_1$), namely there exist constants $c_i$ ($i = 1, 2, 3, 4$) independent of $U_h$ such that*

$$c_1||u_h||_{0,h} \le ||u_h||_0 \le c_2||u_h||_{0,h}, \ \forall u_h \in U_h, \tag{3.4.21}$$

$$c_3|u_h|_{1,h} \le |u_h|_1 \le c_4|u_h|_{1,h}, \ \forall u_h \in U_h. \tag{3.4.22}$$

**Proof** For $u_h \in U_h$

$$||u_h||_0^2 = \sum_{K\in T_h} \int_K u_h^2 \mathrm{d}x\mathrm{d}y = \sum_{K\in T_h} 2S_Q \int_{\hat{K}} u_h^2 \mathrm{d}\lambda_j \mathrm{d}\lambda_k.$$

It is easy to show that $\int_{\hat{K}} u_h^2 \mathrm{d}\lambda_j \mathrm{d}\lambda_k$ is a positive definite bilinear form of $u_{P_i}, u_{P_j}, u_{P_k}, u_{M_i}, u_{M_j}, u_{M_k}$. Thus (3.4.21) holds.

Next let us turn to (3.4.22). Obviously we only have to prove the equivalence of $|\cdot|_{1,K}$ and $|\cdot|_{1,h,K}$. Since $u_h$ is a quadratic polynomial on $K$, $\left(\frac{\partial u_h}{\partial x}\right)^2 + \left(\frac{\partial u_h}{\partial y}\right)^2$ is also a quadratic polynomial. Hence

$$\begin{aligned} |u_h|_{1,K}^2 & = \int_K \Big[\Big(\frac{\partial u_h}{\partial x}\Big)^2 + \Big(\frac{\partial u_h}{\partial y}\Big)^2\Big]\mathrm{d}x\mathrm{d}y \\ & = \frac{1}{3}\sum_{l=i,j,k} \Big[\Big(\frac{\partial u_h(M_l)}{\partial x}\Big)^2 + \Big(\frac{\partial u_h(M_l)}{\partial y}\Big)^2\Big] S_Q. \end{aligned} \tag{3.4.23}$$

It follows from (3.4.10) that

$$\begin{aligned} & \Big(\frac{\partial u_h}{\partial x}\Big)^2 + \Big(\frac{\partial u_h}{\partial y}\Big)^2 \\ = & \ \frac{1}{4S_Q}\Big[a^2\Big(\frac{\partial u_h}{\partial \lambda_j}\Big)^2 + b^2\Big(\frac{\partial u_h}{\partial \lambda_k}\Big)^2 - 2ab\cos\angle P_jP_iP_k \frac{\partial u_h}{\partial \lambda_j}\frac{\partial u_h}{\partial \lambda_k}\Big]. \end{aligned} \tag{3.4.24}$$

The quasi-uniformness of the decomposition leads to the existence of $\sigma > 0$ and $\theta_0 > 0$ satisfying

$$\frac{h_K}{\rho_K} \le \sigma,\ \theta_K \ge \theta_0,\ \forall K \in T_h, \tag{3.4.25}$$

where $\rho_K$ is the diameter of the inscribed circle of $K$, $h_K$ the maximum side-length of $K$, and $\theta_K$ the minimum interior angle of $K$. By (3.4.24) and (3.4.25) we have

$$\begin{aligned} & \frac{1-\cos\theta_0}{\sigma^2}\Big[\Big(\frac{\partial u_h}{\partial \lambda_j}\Big)^2 + \Big(\frac{\partial u_h}{\partial \lambda_k}\Big)^2\Big] \\ \le\ & \Big[\Big(\frac{\partial u_h}{\partial x}\Big)^2 + \Big(\frac{\partial u_h}{\partial y}\Big)^2\Big] S_Q \le 2\sigma^2\Big[\Big(\frac{\partial u_h}{\partial \lambda_j}\Big)^2 + \Big(\frac{\partial u_h}{\partial \lambda_k}\Big)^2\Big]. \end{aligned} \tag{3.4.26}$$

By (3.4.23) and (3.4.26) there exist constants $c_3', c_4' > 0$ such that

$$\begin{aligned} & c_3' \sum_{l=i,j,k} \Big[\Big(\frac{\partial u_h(\hat{M}_l)}{\partial \lambda_j}\Big)^2 + \Big(\frac{\partial u_h(\hat{M}_l)}{\partial \lambda_k}\Big)^2\Big] \\ \le\ & |u_h|_{1,K}^2 \le c_4' \sum_{l=i,j,k} \Big[\Big(\frac{\partial u_h(\hat{M}_l)}{\partial \lambda_j}\Big)^2 + \Big(\frac{\partial u_h(\hat{M}_l)}{\partial \lambda_k}\Big)^2\Big], \end{aligned} \tag{3.4.27}$$

where $\hat{M}_l$ is the image of $M_l$ by the transformation (3.4.8). Write

$$\begin{cases} z_1 = u_{P_i} - u_{M_i},\ z_2 = u_{P_j} - u_{M_j}, \\ z_3 = u_{P_k} - u_{M_k},\ z_4 = u_{M_i} - u_{M_j}, \\ z_5 = u_{M_i} - u_{M_k}. \end{cases} \tag{3.4.28}$$

By (3.4.11) we have

$$\begin{aligned} & \sum_{l=i,j,k} \Big[\Big(\frac{\partial u_h(\hat{M}_l)}{\partial \lambda_j}\Big)^2 + \Big(\frac{\partial u_h(\hat{M}_l)}{\partial \lambda_k}\Big)^2\Big] \\ =\ & (z_1 + z_2 + z_3 + 2z_5)^2 + (z_1 + z_3 + 2z_4 + z_5)^2 \\ & +(-z_1 - z_2 + 3z_4 - 2z_5)^2 + (-z_1 + z_3 - z_5)^2 \\ & +(-z_1 + z_2 - z_4)^2 + (-z_1 - z_3 + 3z_5 - 2z_4)^2. \end{aligned} \tag{3.4.29}$$

It is easy to check that the right-hand side of the above equation is a positive definite bilinear form of $z_1, z_2, \cdots, z_5$, and hence it is equivalent to $\sum_{i=1}^{5} z_i^2$. Now by (3.4.27) $|u_h|_{1,K}$ is equivalent to $|u_h|_{1,h,K}$. □

Denote by $\Pi_h w$ and $\Pi_h^* w$ the interpolations of $w$ in $U_h$ and $V_h$ respectively:

$$\Pi_h w = \sum_{P_0 \in \dot{\Omega}_h} w(P_0)\phi_{P_0} + \sum_{M_0 \in \dot{M}_h} w(M_0)\phi_{M_0}, \tag{3.4.30a}$$

$$\Pi_h^* w = \sum_{P_0 \in \dot{\Omega}_h} w(P_0)\psi_{P_0} + \sum_{M_0 \in \dot{M}_h} w(M_0)\psi_{M_0}, \tag{3.4.30b}$$

**Theorem 3.4.1** *Suppose that the maximum angle of each element of the triangulation $T_h$ is not greater than $\frac{\pi}{2}$, and that the ratio $\tau$ of the lengths of the two sides of the maximum angle satisfies $\tau \in [\sqrt{\frac{2}{3}}, \sqrt{\frac{3}{2}}]$. Then there exists a constant $\alpha > 0$ independent of $U_h$ such that*

$$a(u_h, \Pi_h^* u_h) \geq \alpha \|u_h\|_1^2, \ \forall u_h \in U_h. \tag{3.4.31}$$

**Proof** By (3.4.17)

$$I_K(u_h, \Pi_h^* u_h) = \frac{1}{36 S_Q} Z^T \tilde{A} Z, \tag{3.4.32}$$

where $Z = [z_1, z_2, z_3, z_4, z_5]^T$, and $\tilde{A} = [\tilde{a}_{ij}]$ is a symmetric $5 \times 5$ matrix with

$$\tilde{a}_{11} = 10c^2, \qquad \tilde{a}_{12} = a^2 - b^2 + c^2,$$

$$\tilde{a}_{13} = -a^2 + b^2 + c^2, \qquad \tilde{a}_{14} = \frac{-13a^2 + 13b^2 + 7c^2}{2},$$

$$\tilde{a}_{15} = \frac{13a^2 - 13b^2 + 7c^2}{2}, \qquad \tilde{a}_{22} = 10a^2,$$

$$\tilde{a}_{23} = a^2 + b^2 - c^2, \qquad \tilde{a}_{24} = -7a^2,$$

$$\tilde{a}_{25} = \frac{7a^2 - 13b^2 + 13c^2}{2}, \qquad \tilde{a}_{33} = 10b^2,$$
$$\tilde{a}_{34} = \frac{-13a^2 + 7b^2 + 13c^2}{2}, \quad \tilde{a}_{35} = -7b^2,$$
$$\tilde{a}_{44} = 8(a^2 + b^2 + c^2), \qquad \tilde{a}_{45} = -4(a^2 + b^2 + c^2),$$
$$\tilde{a}_{55} = 8(a^2 + b^2 + c^2).$$

Next we prove the existence of a constant $\tilde{\alpha} > 0$ independent of $K$ such that

$$I_K(u_h, \Pi_h^* u_h) \geq \tilde{\alpha} Z^T Z = \tilde{\alpha} |u|_{1,h,K}^2, \ \forall u_h \in U_h. \tag{3.4.33}$$

Define

$$B = [b_{ij}] = G^T \tilde{A} G,$$

where

$$G = \begin{bmatrix} 1 & 0 & 0 & -5/8 & 3/8 \\ 0 & 1 & 0 & 3/8 & -5/8 \\ 0 & 0 & 1 & 3/8 & 3/8 \\ 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

Then $B$ is a symmetric $5 \times 5$ matrix with

$$b_{11} = 10c^2, \ b_{12} = -b_{13} = a^2 - b^2 + c^2, \ b_{14} = \frac{3}{2}c^2,$$
$$b_{15} = \frac{11}{2}a^2 - \frac{11}{2}b^2, \ b_{22} = 10a^2, \ b_{23} = a^2 + b^2 - c^2,$$
$$b_{24} = -\frac{11}{2}b^2 + \frac{11}{2}c^2, \ b_{25} = \frac{3}{2}a^2, \ b_{33} = 10b^2,$$
$$b_{34} = -b_{35} = -\frac{11}{2}a^2 + \frac{11}{2}c^2,$$
$$b_{44} = \frac{35}{16}a^2 + \frac{35}{16}b^2 + \frac{187}{16}c^2,$$
$$b_{45} = -\frac{81}{16}a^2 + \frac{35}{16}b^2 - \frac{81}{16}c^2,$$
$$b_{55} = \frac{187}{16}a^2 + \frac{35}{16}b^2 + \frac{35}{16}c^2.$$

Without loss of generality, we assume $b$ is the largest side of $K$. Then

$$\tau_0 \leq \frac{a}{c} \leq \tau_0^{-1}, \ \tau_0 = \sqrt{\frac{2}{3}}.$$

The matrix $B$ is positive definite because

$$\begin{aligned}
b_{11} - \sum_{i\neq 1} |b_{1i}| &= \frac{13}{2}c^2 - |\frac{11}{2}a^2 - \frac{11}{2}b^2| \geq c^2 \geq \tau_o ac\sin\theta = 2\tau_0 S_Q,\\
b_{22} - \sum_{i\neq 2} |b_{2i}| &= \frac{13}{2}a^2 - | - \frac{11}{2}b^2 + \frac{11}{2}c^2| \geq 2\tau_0 S_Q,\\
b_{33} - \sum_{i\neq 3} |b_{3i}| &= 8b^2 - |11a^2 - 11c^2| \geq 2\tau_0(11\tau_0 - 3)S_Q,\\
b_{44} - \sum_{i\neq 4} |b_{4i}| &= -\frac{46}{16}a^2 - \frac{18}{16}b^2 + \frac{170}{16}c^2 - |\frac{11}{2}a^2 + \frac{11}{2}c^2| \geq \tau_0 S_Q,\\
b_{55} - \sum_{i\neq 5} |b_{5i}| &= \frac{170}{16}a^2 - \frac{18}{16}b^2 - \frac{46}{16}c^2 - |\frac{11}{2}a^2 - \frac{11}{2}b^2| \geq \tau_0 S_Q.
\end{aligned}$$

By the Gerschgorin theorem the smallest eigenvalue

$$\lambda_{\min} \geq \tau_0 S_Q. \tag{3.4.34}$$

Therefore, by (3.4.32) there exists a constant $\tilde{\alpha} > 0$ such that

$$\begin{aligned}
&I_K(u_h, \Pi_h^* u_h) \geq \frac{1}{36 S_Q}(G^{-1}Z)^T B (G^{-1}Z)\\
\geq\ & \frac{\tau_0}{36}(G^{-1}Z)^T(G^{-1}Z) \geq \tilde{\alpha} Z^T Z = \tilde{\alpha}|u|_{1,h,K}^2,
\end{aligned} \tag{3.4.35}$$

which gives (3.4.33). Combining (3.4.12), (3.4.19) and Lemma 3.4.1 leads to (3.4.31). □

### 3.4.4 Error estimates

**Theorem 3.4.2** *Suppose that the triangulation $T_h$ satisfies the conditions in Theorem 3.4.1. Let $u$ be the solution to (3.4.2) and $u_h$ to the quadratic element difference scheme (3.4.6). If $u \in H^3(\Omega)$, then the following error estimate holds:*

$$||u - u_h||_1 \leq Ch^2|u|_3. \tag{3.4.36}$$

**Proof** The proof is similar to that of Theorem 3.2.2. (3.4.2), (3.4.6) and the *a priori* estimate (3.4.31) imply

$$||u-u_h||_1 \le ||u-\Pi_h u||_1 + \frac{1}{\alpha}\sup_{\overline{u}_h\in U_h}\frac{|a(u-\Pi_h u,\Pi_h^*\overline{u}_h)|}{||\overline{u}_h||_1}. \qquad (3.4.37)$$

By (3.4.12) and (3.4.13) we have

$$a(u-\Pi_h u,\Pi_h^*\overline{u}_h) = \sum_{K\in T_h} I_K(u-\Pi_h u,\Pi_h^*\overline{u}_h), \qquad (3.4.38)$$

$$\begin{aligned} & I_K(u-\Pi_h u,\Pi_h^*\overline{u}_h) \\ = & \sum_{l=i,j,k}\Big[\int_{\overline{Q_lP_{l,l+1}}}\Big(-\frac{\partial(u-\Pi_h u)}{\partial x}\mathrm{d}y+\frac{\partial(u-\Pi_h u)}{\partial y}\mathrm{d}x\Big)(\overline{u}_{M_{l+2}}-\overline{u}_{P_l}) \\ & +\int_{\overline{Q_lP_{l,l+2}}}\Big(-\frac{\partial(u-\Pi_h u)}{\partial x}\mathrm{d}y+\frac{\partial(u-\Pi_h u)}{\partial y}\mathrm{d}x\Big)(\overline{u}_{P_l}-\overline{u}_{M_{l+1}}) \\ & +\int_{\overline{QQ_l}}\Big(-\frac{\partial(u-\Pi_h u)}{\partial x}\mathrm{d}y+\frac{\partial(u-\Pi_h u)}{\partial y}\mathrm{d}x\Big)(\overline{u}_{M_{l+2}}-\overline{u}_{M_{l+1}})\Big], \end{aligned} \qquad (3.4.39)$$

where $i+1=j, j+1=k, k+1=i, \overline{u}_P=\overline{u}_h(P)$ etc. By the definition of $|\cdot|_{1,h,K}$ we have

$$|\overline{u}_{M_{l+2}}-\overline{u}_{P_l}|, |\overline{u}_{P_l}-\overline{u}_{M_{l+1}}|, |\overline{u}_{M_{l+2}}-\overline{u}_{M_{l+1}}| \le C|\overline{u}_h|_{1,h,K}. \quad (3.4.40)$$

Let $L=\overline{Q_lP_{l,l+1}}$ (or $\overline{Q_lP_{l,l+2}}$, $\overline{QQ_l}$) and $\phi_1=\frac{\partial(u-\Pi_h u)}{\partial x}$, $\phi_2=\frac{\partial(u-\Pi_h u)}{\partial y}$. Then

$$\begin{aligned} & \Big|\int_L\Big(-\frac{\partial(u-\Pi_h u)}{\partial x}\mathrm{d}y+\frac{\partial(u-\Pi_h u)}{\partial y}\mathrm{d}x\Big)\Big| \\ \le & \int_L(|\phi_1|+|\phi_2|)\mathrm{d}s \le h^{1/2}\Big[\int_L(\phi_1^2+\phi_2^2)\mathrm{d}s\Big]^{1/2}. \end{aligned} \qquad (3.4.41)$$

Assume that the linear mapping (3.4.8) maps the element $K$ onto the reference element $\hat{K}$, the segment $L$ into $\hat{L}$, and the function $\phi_i$ on $K$ into the function $\hat{\phi}_i(\lambda_j,\lambda_k)=\phi_i(x,y)$, $(i=1,2)$ on $\hat{K}$. Then we have

$$\int_L \phi_i^2\mathrm{d}s \le h\int_{\hat{L}}\hat{\phi}_i^2\mathrm{d}\hat{s},\ i=1,2. \qquad (3.4.42)$$

Let $L$ be a part of the boundary of $K^*_{M_l}$. Employing the trace theorem on $\hat{K}^*_{\hat{M}_l} \cap \hat{K}$ we have a constant $C > 0$ independent of $K$ such that

$$\int_{\hat{L}} \hat{\phi}_i^2 \mathrm{d}\hat{s} \le C\|\hat{\phi}_i\|^2_{1,\hat{K}}, \ i = 1, 2. \tag{3.4.43}$$

After the affine transformation, the Sobolev semi-norms have the following relationships:

$$\begin{cases} |\hat{\phi}_i|_{0,\hat{K}} \le Ch^{-1}|\phi_i|_{0,K}, \\ |\hat{\phi}_i|_{1,\hat{K}} \le C|\phi_i|_{1,K}, \end{cases} \quad i = 1, 2. \tag{3.4.44}$$

By (3.4.42)-(3.4.44) and the interpolation approximation theorem we have

$$\begin{aligned} \int_L \phi_i^2 \mathrm{d}s \ &\le Ch(h^{-1}|\phi_i|_{0,K} + |\phi_i|_{1,K})^2 \\ &\le Ch(h^{-1}|u - \Pi_h u|_{1,K} + |u - \Pi_h u|_{2,K})^2 \\ &\le Ch^3|u|^2_{3,K}. \end{aligned} \tag{3.4.45}$$

A combination of (3.4.39)-(3.4.41) and (3.4.45) yields

$$I_K(u - \Pi_h u, \Pi_h^* \overline{u}_h) \le Ch^2|u|_{3,K}|\overline{u}_h|_{1,h,K}. \tag{3.4.46}$$

This together with (3.4.38) and Lemma 3.4.1 leads to

$$a(u - \Pi_h u, \Pi_h^* \overline{u}_h) \le Ch^2|u|_3|\overline{u}_h|_1. \tag{3.4.47}$$

Finally, (3.4.36) results from (3.4.37), (3.4.47) and the interpolation approximation property. □

### 3.4.5 Numerical example

The following problem is approximated by the five point finite difference method (FDM), the quadratic finite element method (FEM) and the quadratic element generalized difference method (GDM), respectively:

$$\begin{cases} -\Delta u = 2\sin x \sin y, \ \text{on } \Omega = (0, \pi) \times (0, \pi), \\ u|_{\partial\Omega} = 0. \end{cases} \tag{3.4.48}$$

Place a right triangular decomposition on $\Omega$ (see Fig. 3.4.5) with the right-angle-side length $h = \frac{\pi}{N}$, $x_i = ih, y_j = jh$, $i, j = 1, 2, \cdots, N$. The results of the three methods as well as the true solution (TS) $u(x, y) = \sin x \sin y$ are given in Table 3.4.1.

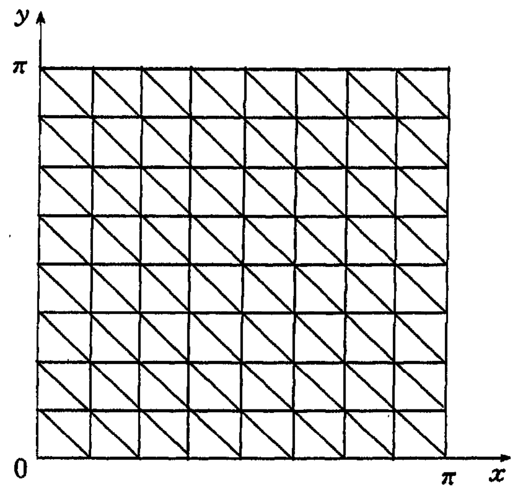



Fig. 3.4.5

**Table 3.4.1. Numerical results** ([A-41])

| | FDM(N=8) | FEM(N=8) | GDM(N=8) | TS |
|---|---|---|---|---|
| $(x_1, y_1)$ | 0.148343 | 0.146418 | 0.146178 | 0.146447 |
| $(x_2, y_1)$ | 0.274102 | 0.270546 | 0.271157 | 0.270598 |
| $(x_2, y_2)$ | 0.506475 | 0.499904 | 0.502044 | 0.500000 |
| $(x_3, y_1)$ | 0.358132 | 0.353485 | 0.355223 | 0.353553 |
| $(x_3, y_2)$ | 0.661742 | 0.653156 | 0.656951 | 0.653281 |
| $(x_3, y_3)$ | 0.864607 | 0.853389 | 0.858948 | 0.853553 |
| $(x_4, y_1)$ | 0.387639 | 0.382610 | 0.385339 | 0.382683 |
| $(x_4, y_2)$ | 0.716264 | 0.706971 | 0.712014 | 0.707107 |
| $(x_4, y_3)$ | 0.935844 | 0.923702 | 0.930291 | 0.923879 |
| $(x_4, y_4)$ | 1.012950 | 0.999808 | 1.006940 | 1.000000 |

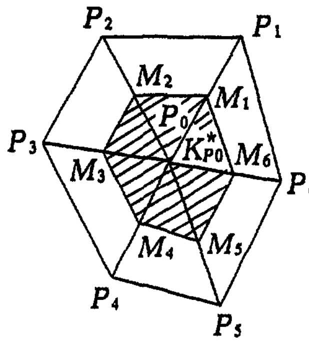



Fig. 3.5.1

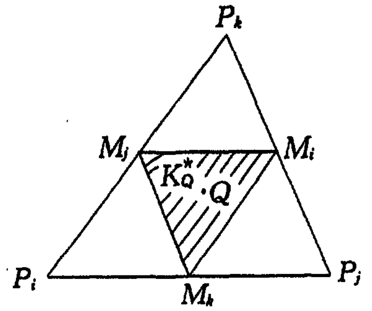



Fig. 3.5.2

## 3.5 Cubic Element Difference Schemes

In this section we discuss a generalized difference scheme based on a cubic element of Hermite type for problem (3.4.1).

### 3.5.1 Trial and test function spaces

Let $T_h$ be a quasi-uniform triangulation of $\overline{\Omega}$ as in §3.2 and $h$ the largest side length of all the triangles. $T_h$ consists of a finite number of triangular elements $K_Q$'s, $Q$ being the barycenter of the triangle. Denote by $\overline{\Omega}_h$ and $\Omega_h^*$ the sets of the vertexes and the barycenters of all the triangular elements, respectively, and $\dot{\Omega}_h = \overline{\Omega}_h \setminus \partial\Omega$. Let $S_Q$ be the area of $K_Q$.

For the dual decomposition, we consider a vertex $P_0$ of a triangular element. Suppose $P_i$ $(i = 1, 2 \cdots, 6)$ are the adjacent vertexes of $P_0$ and $M_i$ is the midpoint of $\overline{P_0P_i}$ (cf. Fig. 3.5.1). Connect $M_i$ successively to obtain a polygonal region $K_{P_0}^*$ surrounding $P_0$, as an element of the dual decomposition. Suppose that $Q$ is the barycenter of a triangular element $K = \triangle P_iP_jP_k$, and that $M_i, M_j, M_k$ are the midpoints of $\overline{P_jP_k}, \overline{P_kP_i}, \overline{P_iP_j}$ respectively (cf. Fig. 3.5.2). Connecting $M_i, M_j$ and $M_k$ results in a triangular region $K_Q^*$ surrounding $Q$, which is also taken as an element of the dual decomposition. These two kinds of dual elements form a dual decomposition, denoted by $T_h^*$.

The trial function space is chosen as the Hermitian cubic element space related to $T_h$. There are three basis functions corresponding to a node $P_0 \in \overline{\Omega}_h$, denoted by $\phi_{P_0}^{(k)}$ $(k = 0, 1, 2)$ and satisfying the following interpolation conditions:

$$\begin{cases} \phi_{P_0}^{(0)}(P_0) = 1, \\ \phi_{P_0}^{(0)}(P) = 0, \text{ if } P \in \overline{\Omega}_h \cup \Omega_h^* \setminus \{P_0\}, \\ \dfrac{\partial}{\partial x}\phi_{P_0}^{(0)}(P) = \dfrac{\partial}{\partial y}\phi_{P_0}^{(0)}(P) = 0, \text{ if } P \in \overline{\Omega}_h; \end{cases}$$

$$\begin{cases} \dfrac{\partial}{\partial x}\phi_{P_0}^{(1)}(P_0) = 1, & \\ \dfrac{\partial}{\partial x}\phi_{P_0}^{(1)}(P) = 0, & \text{if } P \in \overline{\Omega}_h \setminus \{P_0\}, \\ \phi_{P_0}^{(1)}(P) = 0, & \text{if } P \in \overline{\Omega}_h \cup \Omega_h^*, \\ \dfrac{\partial}{\partial y}\phi_{P_0}^{(1)}(P) = 0, & \text{if } P \in \overline{\Omega}_h; \end{cases}$$

$$\begin{cases} \dfrac{\partial}{\partial y}\phi_{P_0}^{(2)}(P_0) = 1, & \\ \dfrac{\partial}{\partial y}\phi_{P_0}^{(2)}(P) = 0, & \text{if } P \in \overline{\Omega}_h \setminus \{P_0\}, \\ \phi_{P_0}^{(2)}(P) = 0, & \text{if } P \in \overline{\Omega}_h \cup \Omega_h^*, \\ \dfrac{\partial}{\partial x}\phi_{P_0}^{(2)}(P) = 0, & \text{if } P \in \overline{\Omega}_h. \end{cases}$$

There is also a basis function $\phi_{Q_0}(P)$ related to the barycenter $Q$ of the triangular element, satisfying the interpolation condition

$$\begin{cases} \phi_{Q_0}(Q_0) = 1, \\ \phi_{Q_0}(P) = 0, \text{ if } P \in \overline{\Omega}_h \cup \Omega_h^* \setminus \{Q_0\}, \\ \dfrac{\partial}{\partial x}\phi_{Q_0}(P) = \dfrac{\partial}{\partial y}\phi_{Q_0}(P) = 0, \text{ if } P \in \overline{\Omega}_h. \end{cases}$$

Taking into account of the boundary condition $u|_{\partial\Omega} = 0$, we choose

$$U_h = \text{span}\{\phi_{P_0}^{(k)}, \phi_{Q_0} : \ P_0 \in \dot{\Omega}_h, k = 0; P_0 \in \overline{\Omega}_h, k = 1, 2; Q_0 \in \Omega_h^*\}.$$

The test function space is chosen as the piecewise constant and piecewise linear function space. The three basis functions related to $P_0 = (x_0, y_0) \in \overline{\Omega}_h$ are

$$\psi_{P_0}^{(0)}(P) = \begin{cases} 1, & \text{if } P \in K_{P_0}^*, \\ 0, & \text{if } P \notin K_{P_0}^*; \end{cases}$$

$$\psi_{P_0}^{(1)}(P) = \begin{cases} x - x_0, & \text{if } P \in K_{P_0}^*, \\ 0, & \text{if } P \notin K_{P_0}^*; \end{cases}$$

$$\psi_{P_0}^{(2)}(P) = \begin{cases} y - y_0, & \text{if } P \in K_{P_0}^*, \\ 0, & \text{if } P \notin K_{P_0}^*. \end{cases}$$

The basis function related to $Q_0 \in \Omega_h^*$ is

$$\psi_{Q_0}^{(0)}(P) = \begin{cases} 1, & \text{if } P \in K_{Q_0}^*, \\ 0, & \text{if } P \notin K_{Q_0}^*. \end{cases}$$

Similarly we require that the functions in $V_h$ vanish on the boundary. So we have

$$V_h = \text{span}\{\psi_{P_0}^{(k)}, \psi_{Q_0} : \ P_0 \in \dot{\Omega}_h, k = 0; P_0 \in \overline{\Omega}_h, k = 1, 2; Q_0 \in \Omega_h^*\}.$$

### 3.5.2 Generalized difference equations

The cubic element difference scheme corresponding to $U_h$ and $V_h$ defined above is: Find $u_h \in U_h$ satisfying

$$a(u_h, v_h) = (f, v_h), \ \forall v_h \in V_h, \tag{3.5.1}$$

or equivalently

$$\begin{cases} a(u_h, \psi_{P_0}^{(k)}) = (f, \psi_{P_0}^{(k)}), & (3.5.2a) \\ \quad P_0 \in \dot{\Omega}_h, k = 0; \ P_0 \in \overline{\Omega}_h, k = 1, 2, & \\ a(u_h, \psi_{Q_0}) = (f, \psi_{Q_0}), \ Q_0 \in \Omega_h^*, & (3.5.2b) \end{cases}$$

where

$$a(u_h, \psi_{P_0}^{(0)}) = \int_{\partial K_{P_0}^*} \frac{\partial u_h}{\partial y}\mathrm{d}x - \frac{\partial u_h}{\partial x}\mathrm{d}y, \tag{3.5.3a}$$

$$a(u_h, \psi_{P_0}^{(1)}) = \int_{K_{P_0}^*} \frac{\partial u_h}{\partial x}\mathrm{d}x\mathrm{d}y + \int_{\partial K_{P_0}^*} \frac{\partial u_h}{\partial y}\psi_{P_0}^{(1)}\mathrm{d}x - \frac{\partial u_h}{\partial x}\psi_{P_0}^{(1)}\mathrm{d}y, \tag{3.5.3b}$$

$$a(u_h, \psi_{P_0}^{(2)}) = \int_{K_{P_0}^*} \frac{\partial u_h}{\partial y}\mathrm{d}x\mathrm{d}y + \int_{\partial K_{P_0}^*} \frac{\partial u_h}{\partial y}\psi_{P_0}^{(2)}\mathrm{d}x - \frac{\partial u_h}{\partial x}\psi_{P_0}^{(2)}\mathrm{d}y, \tag{3.5.3c}$$

$$a(u_h, \psi_{Q_0}) = \int_{\partial K_{Q_0}^*} \frac{\partial u_h}{\partial y}\mathrm{d}x - \frac{\partial u_h}{\partial x}\mathrm{d}y. \tag{3.5.3d}$$

Define an interpolation operator $\Pi_h^* : U_h \to V_h$ by

$$\begin{aligned}\Pi_h^*\overline{u}_h = &\sum_{P_0\in\overline{\Omega}_h}\Big[\overline{u}_h(P_0)\psi_{P_0}^{(0)} + \frac{\partial\overline{u}_h(P_0)}{\partial x}\psi_{P_0}^{(1)} \\ &+\frac{\partial\overline{u}_h(P_0)}{\partial y}\psi_{P_0}^{(2)}\Big] + \sum_{Q_0\in\Omega_h^*}\overline{u}_h(Q_0)\psi_{Q_0}.\end{aligned} \tag{3.5.4}$$

Then, (3.5.1) is equivalent to

$$a(u_h, \Pi_h^*\overline{u}_h) = (f, \Pi_h^*\overline{u}_h),\ \forall \overline{u}_h \in U_h, \tag{3.5.5}$$

To compute the element stiff matrix we write $a(u_h, \Pi_h^*\overline{u}_h)$ as

$$a(u_h, \Pi_h^*\overline{u}_h) = \sum_{K\in T_h} I_K(u_h, \Pi_h^*\overline{u}_h), \tag{3.5.6}$$

where $K = \triangle P_iP_jP_k$ (cf. Fig. 3.5.2),

$$\begin{aligned}&I_K(u_h, \Pi_h^*\overline{u}_h) \\ =& \sum_{l=i,j,k}\Big[\overline{u}_h(P_l)I_K(u_h, \psi_{P_l}^{(0)}) + \frac{\partial\overline{u}_h(P_l)}{\partial x}I_K(u_h, \psi_{P_l}^{(1)}) \\ &+\frac{\partial\overline{u}_h(P_l)}{\partial y}I_K(u_h, \psi_{P_l}^{(2)})\Big] + \overline{u}_h(Q)I_K(u_h, \psi_Q),\end{aligned} \tag{3.5.7}$$

and each $I_K(u_h, \psi_P)$ is obtained by changing the integral regions $K_P^*$ and $\partial K_P^*$ in $a(u_h, \psi_P)$ (cf. (3.5.3)) into $K_P^* \cap K$ and $\partial K_P^* \cap K$ respectively.

On $K$, $u_h \in U_h$ can be expressed in terms of $(\lambda_i, \lambda_j, \lambda_k)$ as (cf. Example 3 in §1.1.4):

$$
\begin{aligned}
u_h = \; & 27\lambda_i\lambda_j\lambda_k u_h(Q) \\
& + \sum_{l=i,j,k} (-2\lambda_l^3 + 3\lambda_l^2 - 7\lambda_i\lambda_j\lambda_k) u_h(P_l) \\
& + \sum_{l=j,k} \Big[ (2\lambda_i\lambda_j\lambda_k - \lambda_l\lambda_j\lambda_k - \lambda_l^2\lambda_i) \frac{\partial u_h(P_l)}{\partial \lambda_l} \\
& + \sum_{\substack{m=i,j,k \\ m\neq l}} (\lambda_m^2\lambda_l - \lambda_i\lambda_j\lambda_k) \frac{\partial u_h(P_m)}{\partial \lambda_l} \Big].
\end{aligned}
\tag{3.5.8}
$$

After the transformation $(x, y) \to (\lambda_j, \lambda_k)$, the triangular element $K$ becomes the reference element $\hat{K}$ (cf. Fig. 3.5.3), and the points $P_l, M_l$ and $Q$ become $\hat{P}_l, \hat{M}_l$ and $\hat{Q}$ ($l = i, j, k$) respectively. Also note

$$\frac{\partial u_h}{\partial x} = \frac{1}{2S_Q}\Big[(y_k - y_i)\frac{\partial u_h}{\partial \lambda_j} + (y_i - y_j)\frac{\partial u_h}{\partial \lambda_k}\Big], \tag{3.5.9a}$$

$$\frac{\partial u_h}{\partial y} = \frac{1}{2S_Q}\Big[(x_i - x_k)\frac{\partial u_h}{\partial \lambda_j} + (x_j - x_i)\frac{\partial u_h}{\partial \lambda_k}\Big], \tag{3.5.9b}$$

$$\mathrm{d}x = (x_j - x_i)\mathrm{d}\lambda_j + (x_k - x_i)\mathrm{d}\lambda_k, \tag{3.5.10a}$$

$$\mathrm{d}y = (y_j - y_i)\mathrm{d}\lambda_j + (y_k - y_i)\mathrm{d}\lambda_k. \tag{3.5.10b}$$

The integrals in $I_K(u_h, \Pi_h^* \overline{u}_h)$ can be changed into the integrals on the reference element $\hat{K}$ on $(\lambda_j, \lambda_k)$ plane by the transformation $(x, y) \to (\lambda_j, \lambda_k)$. Then the element stiff matrix can be obtained by computing $I_K(u_h, \Pi_h^* \overline{u}_h)$ as in §3.4.2.

### 3.5.3 *a priori* estimates

Let us introduce a discrete semi-norm

$$|u_h|_{1,h} = \Big(\sum_{K\in T_h} |u_h|_{1,h,K}^2\Big)^{1/2}, \tag{3.5.11}$$

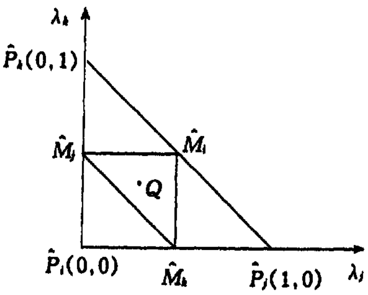



Fig. 3.5.3

where

$$|u_h|_{1,h,K}^2 = \sum_{l=i,j,k} \Big[(u_h(P_l) - u_h(Q))^2 + \Big(\frac{\partial u_h(P_l)}{\partial \lambda_j}\Big)^2 + \Big(\frac{\partial u_h(P_l)}{\partial \lambda_k}\Big)^2\Big].$$

**Lemma 3.5.1** *The discrete semi-norm* $|\cdot|_{1,h}$ *is equivalent to the* $H^1$*−semi-norm* $|\cdot|_1$ *on* $U_h$*, namely there exist constants* $c_1, c_2 > 0$ *independent of* $U_h$ *such that*

$$c_1 |u_h|_{1,h} \le |u_h|_1 \le c_2 |u_h|_{1,h}, \ \forall u_h \in U_h. \tag{3.5.12}$$

**Proof** For $u_h \in U_h$,

$$\begin{aligned} |u_h|_1^2 &= \sum_{K\in T_h} \int_K \Big[\Big(\frac{\partial u_h}{\partial x}\Big)^2 + \Big(\frac{\partial u_h}{\partial y}\Big)^2\Big] \mathrm{d}x\mathrm{d}y \\ &= \sum_{K\in T_h} \int_{\hat{K}} \Big[\Big(\frac{\partial u_h}{\partial x}\Big)^2 + \Big(\frac{\partial u_h}{\partial y}\Big)^2\Big] \cdot 2S_Q \mathrm{d}\lambda_j \mathrm{d}\lambda_k. \end{aligned} \tag{3.5.13}$$

It follows from (3.5.9) that

$$\begin{aligned} \Big(\frac{\partial u_h}{\partial x}\Big)^2 + \Big(\frac{\partial u_h}{\partial y}\Big)^2 = \ & \frac{1}{4S_Q^2}\Big[a^2\Big(\frac{\partial u_h}{\partial \lambda_j}\Big)^2 + b^2\Big(\frac{\partial u_h}{\partial \lambda_k}\Big)^2 \\ & -2ab\cos\angle P_jP_iP_k \frac{\partial u_h}{\partial \lambda_j}\frac{\partial u_h}{\partial \lambda_k}\Big], \end{aligned} \tag{3.5.14}$$

where $a = |\overline{P_kP_i}|, b = |\overline{P_iP_j}|$ and $c = |\overline{P_jP_k}|$. The regularity of the decomposition implies the existence of constants $\sigma > 0$ and $\theta_0 > 0$

such that

$$\frac{h_K}{\rho_K} \le \sigma,\ \theta_K \ge \theta_0,\ \forall K \in T_h,$$

where $\rho_K$ is the diameter of the inscribed circle of $K$, $h_K$ the maximum side length of $K$ and $\theta_K$ the minimum interior angle. By (3.5.14)

$$\begin{aligned}&\frac{1-\cos\theta_0}{\sigma^2}\Big[\Big(\frac{\partial u_h}{\partial\lambda_j}\Big)^2+\Big(\frac{\partial u_h}{\partial\lambda_k}\Big)^2\Big]\\ \le\ &\Big[\Big(\frac{\partial u_h}{\partial x}\Big)^2+\Big(\frac{\partial u_h}{\partial x}\Big)^2\Big]S_Q \le 2\sigma^2\Big[\Big(\frac{\partial u_h}{\partial\lambda_j}\Big)^2+\Big(\frac{\partial u_h}{\partial\lambda_k}\Big)^2\Big].\end{aligned} \tag{3.5.15}$$

It easily follows from (3.5.8) that $\int_{\hat{K}}\Big[\Big(\frac{\partial u_h}{\partial\lambda_j}\Big)^2+\Big(\frac{\partial u_h}{\partial\lambda_k}\Big)^2\Big]\mathrm{d}\lambda_j\mathrm{d}\lambda_k$ is a positive definite bilinear form of

$$\begin{aligned}\hat{Z} =\ &\Big[u_h(P_j)-u_h(Q), u_h(P_k)-u_h(Q), u_h(P_i)-u_h(Q),\\ &\frac{\partial u_h(P_j)}{\partial\lambda_j}, \frac{\partial u_h(P_k)}{\partial\lambda_j}, \frac{\partial u_h(P_i)}{\partial\lambda_j}, \frac{\partial u_h(P_j)}{\partial\lambda_k}, \frac{\partial u_h(P_k)}{\partial\lambda_k}, \frac{\partial u_h(P_i)}{\partial\lambda_k}\Big]^T.\end{aligned} \tag{3.5.16}$$

This together with (3.5.13) and (3.5.15) leads to the desired conclusion and completes the proof. □

The norm $|\cdot|_{1,h}$ is also equivalent to the $H^1$-norm $\|\cdot\|_1$ on $U_h$ since $U_h \subset H_0^1(\Omega)$ and $|\cdot|_1$ is an equivalent norm on $H_0^1(\Omega)$.

**Theorem 3.5.1** *Assume that the maximum angle of each element of the triangulation $T_h$ is not greater than $\frac{\pi}{2}$ and that the ratio $\tau$ of the two side lengths of the maximum angle satisfies $\tau \in [\sqrt{\frac{2}{3}}, \sqrt{\frac{3}{2}}]$. Then there exists a constant $\alpha > 0$ independent of $U_h$ such that*

$$a(u_h, \Pi_h^* u_h) \ge \alpha\|u_h\|_1^2,\ \forall u_h \in U_h. \tag{3.5.17}$$

**Proof** Write $I_K(u_h, \Pi_h^* u_h)$ into a symmetric form

$$I_K(u_h, \Pi_h^* u_h) = \frac{1}{768 S_Q} Z^T A Z, \tag{3.5.18}$$

where $A$ is a symmetric matrix

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{12}^T & A_{22} & A_{23} \\ A_{13}^T & A_{23}^T & A_{33} \end{bmatrix},$$

$$A_{11} =$$

$$\begin{bmatrix} 400a^2 & 96c^2 - 40a^2 - 40b^2 & 96b^2 - 40a^2 - 40c^2 \\ 96c^2 - 40a^2 - 40b^2 & 400b^2 & 96a^2 - 40b^2 - 40c^2 \\ 96b^2 - 40a^2 - 40c^2 & 96a^2 - 40b^2 - 40c^2 & 400c^2 \end{bmatrix},$$

$$A_{12} =$$

$$\begin{bmatrix} -96a^2 & 41c^2 - 7a^2 - 20b^2 & 41b^2 - 7a^2 - 20c^2 \\ 13a^2 + 11b^2 - 27c^2 & 41c^2 - 41a^2 + 48b^2 & 7c^2 - 14a^2 - 4b^2 \\ 13a^2 - 27b^2 + 11c^2 & 7b^2 - 14a^2 - 4c^2 & 41b^2 - 41a^2 + 48c^2 \end{bmatrix},$$

$$A_{13} =$$

$$\begin{bmatrix} 48a^2 - 41b^2 + 41c^2 & 11a^2 + 13b^2 - 27c^2 & 7c^2 - 4a^2 - 14b^2 \\ 41c^2 - 20a^2 - 7b^2 & -96b^2 & 41a^2 - 7b^2 - 20c^2 \\ 7a^2 - 14b^2 - 4c^2 & 13b^2 - 27a^2 + 11c^2 & 41a^2 - 41b^2 + 48c^2 \end{bmatrix},$$

$$A_{22} =$$

$$\begin{bmatrix} 34a^2 & a^2 + 5b^2 - 11c^2 & a^2 - 11b^2 + 5c^2 \\ a^2 + 5b^2 - 11c^2 & 7a^2 + 7b^2 + 15c^2 & -2a^2 + b^2 + c^2 \\ a^2 - 11b^2 + 5c^2 & -2a^2 + b^2 + c^2 & 7a^2 + 15b^2 + 7c^2 \end{bmatrix},$$

$$A_{23} =$$

$$\begin{bmatrix} -17a^2 + 4b^2 - 4c^2 & -3a^2 - 3b^2 + 7c^2 & 2a^2 + 4b^2 - 2c^2 \\ -2a^2 - 2b^2 + 13c^2 & 4a^2 - 17b^2 - 4c^2 & -2a^2 + b^2 - 3c^2 \\ a^2 - 2b^2 - 3c^2 & 4a^2 - 2b^2 - 2c^2 & -11a^2 - 11b^2 + 10c^2 \end{bmatrix},$$

$$A_{33} =$$

$$\begin{bmatrix} 7a^2 + 7b^2 + 15c^2 & 5a^2 + b^2 - 11c^2 & a^2 - 2b^2 + c^2 \\ 5a^2 + b^2 - 11c^2 & 34b^2 & -11a^2 + b^2 + 5c^2 \\ a^2 - 2b^2 + c^2 & -11a^2 + b^2 + 5c^2 & 15a^2 + 7b^2 + 7c^2 \end{bmatrix}.$$

Let the shortest and the longest sides of $K$ are $\overline{P_iP_j}$ and $\overline{P_jP_k}$ respectively. Divide our discussion into the following two cases.

**Case 1.** Suppose $c^2 \leq \frac{3}{4}(a^2+b^2)$. Take $a^2 = b^2 = c^2 = 1$ in the subblocks $A_{ij}$ of $A$ to obtain $\hat{A}_{ij}$. Set

$$G = \begin{bmatrix} \frac{1}{4}I & -D_{12} & D_{12}(\hat{A}_{22} - \hat{A}_{12}^T D_{12})^{-1}(\hat{A}_{23} - \hat{A}_{12}^T D_{13}) - D_{13} \\ 0 & I & -(\hat{A}_{22} - \hat{A}_{12}^T D_{12})^{-1}(\hat{A}_{23} - \hat{A}_{12}^T D_{13}) \\ 0 & 0 & I \end{bmatrix}$$

where $I$ is the $3 \times 3$ identity matrix, $D_{12} = \hat{A}_{11}^{-1}\hat{A}_{12}$, $D_{13} = \hat{A}_{11}^{-1}\hat{A}_{13}$. It is easy to show that the inverses of the above submatrices indeed exist. Perform the transformation

$$B = G^T A G,$$

then $B = [b_{ij}]$ is a symmetric matrix where

$$b_{11} = 25a^2, \qquad b_{12} = -2.5a^2 - 2.5b^2 + 6c^2,$$
$$b_{13} = -2.5a^2 + 6b^2 - 2.5c^2, \qquad b_{14} = 0.06a^2 - 0.03b^2 - 0.03c^2,$$
$$b_{15} = -4.04a^2 - 2.99b^2 + 7.03c^2, \qquad b_{16} = -4.04a^2 + 7.03b^2 - 2.99c^2,$$
$$b_{17} = -6.84b^2 + 6.84c^2, \qquad b_{18} = -0.51a^2 - 0.55b^2 + 1.06c^2,$$
$$b_{19} = 0.51a^2 - 1.06b^2 + 0.55c^2, \qquad b_{22} = 25b^2,$$
$$b_{23} = 6a^2 - 2.5b^2 - 2.5c^2, \qquad b_{24} = 0.80a^2 + 0.17b^2 - 0.97c^2,$$
$$b_{25} = -9.13a^2 - 0.03b^2 + 9.16c^2, \qquad b_{26} = -6.07a^2 + 3.87b^2 + 2.20c^2$$
$$b_{27} = -1.89a^2 - 3.06b^2 + 4.95c^2, \qquad b_{28} = -1.84a^2 - 0.28b^2 + 2.12c^2,$$
$$b_{29} = 1.74a^2 - 1.14b^2 - 0.60c^2, \qquad b_{33} = 25c^2,$$
$$b_{34} = 0.80a^2 - 0.97b^2 + 0.17c^2, \qquad b_{35} = -6.07a^2 + 2.20b^2 + 3.87c^2,$$
$$b_{36} = -9.13a^2 + 9.16b^2 - 0.03c^2, \qquad b_{37} = 1.89a^2 - 4.95b^2 + 3.06c^2,$$
$$b_{38} = -1.7a^2 + 0.60b^2 + 1.14c^2, \qquad b_{39} = 1.84a^2 - 2.12b^2 + 0.28c^2,$$
$$b_{44} = 10.93a^2 + 0.01b^2 + 0.01c^2, \qquad b_{45} = -0.85a^2 - 0.12b^2 - 0.74c^2,$$

$$b_{46} = -0.85a^2 - 0.74b^2 - 0.12c^2, \quad b_{47} = -6.08b^2 + 6.08c^2,$$
$$b_{48} = 0.12a^2 + 0.49b^2 - 0.61c^2, \quad b_{49} = -0.12a^2 + 0.61b^2 - 0.49c^2,$$
$$b_{55} = 15.75a^2 + 2.79b^2 + 3.88c^2, \quad b_{56} = 0.72a^2 + 0.88b^2 + 0.88c^2,$$
$$b_{57} = -2.48a^2 + 0.90b^2 + 1.58c^2, \quad b_{58} = -2.10a^2 - 4.84b^2 + 6.85c^2,$$
$$b_{59} = -0.04a^2 + 0.35b^2 - 0.31c^2, \quad b_{66} = 15.75a^2 + 3.88b^2 + 2.79c^2,$$
$$b_{67} = 2.48a^2 - 1.58b^2 - 0.90c^2, \quad b_{68} = 0.04a^2 + 0.31b^2 - 0.35c^2,$$
$$b_{69} = 2.01a^2 - 6.85b^2 + 4.84c^2, \quad b_{77} = 0.75a^2 + 8.42b^2 + 8.42c^2,$$
$$b_{78} = 0.91a^2 + 1.19b^2 - 2.03c^2, \quad b_{79} = 0.91a^2 - 2.03b^2 + 1.19c^2,$$
$$b_{88} = -2.11a^2 + 8.26b^2 + 3.31c^2, \quad b_{89} = 0.53a^2 - 0.36b^2 - 0.36c^2,$$
$$b_{99} = -2.11a^2 + 3.31b^2 + 8.36c^2.$$

Under our assumptions we have

$$b_{11} - \sum_{j\neq 1} |b_{1j}| \geq 20a^2 + 22b^2 - 22c^2 \geq 3.5a^2 + 3.5b^2,$$
$$b_{22} - \sum_{j\neq 2} |b_{2j}| \geq 14.95a^2 + 28.55b^2 - 23.5c^2 \geq 19.46a^2,$$
$$b_{33} - \sum_{j\neq 3} |b_{3j}| \geq -17.95a^2 + 13.5b^2 + 21.45c^2 \geq 3.5a^2 + 13.5b^2,$$
$$b_{44} - \sum_{j\neq 4} |b_{4j}| \geq 9.23a^2 + 7.5b^2 - 9.2c^2 \geq 2.33a^2 + 0.6b^2,$$
$$b_{55} - \sum_{j\neq 5} |b_{5j}| \geq 31.67a^2 + 13.1b^2 - 26.54c^2 \geq 4.95a^2,$$

$$b_{66} - \sum_{j\neq 6} |b_{6j}| \geq -3.31a^2 + 31.06b^2 - 13.1c^2 \geq 1.02a^2,$$

$$b_{77} - \sum_{j\neq 7} |b_{7j}| \geq 0.74a^2 + 35.05b^2 - 18.35c^2 \geq 1.17a^2,$$

$$b_{88} - \sum_{j\neq 8} |b_{8j}| \geq 4.39a^2 + 16.16b^2 - 11.35c^2 \geq 0.97a^2,$$

$$b_{99} - \sum_{j\neq 9} |b_{9j}| \geq -8.61a^2 + 17.11b^2 - 0.36c^2 \geq 2.34a^2.$$

Note $a^2 \geq 2S_Q$. So the Gerschgorin theorem implies that the minimum eigenvalue $\lambda_{\min} \geq 1.94S_Q$.

Case 2. Suppose $c^2 \geq \frac{3}{4}(a^2+b^2)$. Now we take $a^2 = b^2 = 1$, $c^2 = 2$, and perform the same transformation as in Case 1. Then we can similarly show that the minimum eigenvalue of $B$ $\lambda_{\min} \geq 0.14S_Q$.

Summarizing the above two cases and noticing (3.5.18) verify the existence of a constant $\alpha' > 0$ such that

$$I_K(u_h, \Pi_h^* u_h) \geq \frac{1}{768S_Q} \cdot 0.14S_Q Z^T (G^{-1})^T G^{-1} Z \geq \alpha' Z^T Z. \quad (3.5.19)$$

Finally, combining (3.5.6), (3.5.19) and Lemma 3.5.1 yields (3.5.17). □

### 3.5.4 Error estimates

**Theorem 3.5.2** *Let $T_h$ satisfy the assumption of Theorem 3.5.1, and let $u$ and $u_h$ be the solutions to the variational problem (3.4.2) and the cubic element difference scheme (3.5.1) respectively. Then if $u \in H^4(\Omega)$ we have the following error estimate:*

$$||u - u_h||_1 \leq Ch^3 |u|_4. \quad (3.5.20)$$

**Proof** By (3.4.2), (3.5.1) and the *a priori* estimate (3.5.17) we have

$$||u - u_h||_1 \leq ||u - \Pi_h u||_1 + \frac{1}{\alpha} \sup_{\overline{u}_h \in U_h} \frac{a(u - \Pi_h u, \Pi_h^* \overline{u}_h)|}{||\overline{u}_h||_1}, \quad (3.5.21)$$

where $\Pi_h u$ is the interpolation of $u$ onto $U_h$. By the interpolation approximation theorem,

$$||u - \Pi_h u||_1 \le Ch^3 |u|_4. \tag{3.5.22}$$

Next we deal with the second term of the right-hand side of (3.5.21). It follows from (3.5.6) and (3.5.7) that

$$a(u - \Pi_h u, \Pi_h^* \overline{u}_h) = \sum_{K \in T_h} I_K(u - \Pi_h u, \Pi_h^* \overline{u}_h), \tag{3.5.23}$$

$$\begin{aligned} & I_K(u - \Pi_h u, \Pi_h^* \overline{u}_h) \\ = & \sum_{l=i,j,k} \Big\{ [\overline{u}_h(P_l) - \overline{u}_h(Q)] \int_{\partial K^*_{P_l} \cap K} \frac{\partial(u - \Pi_h u)}{\partial y} \mathrm{d}x - \frac{\partial(u - \Pi_h u)}{\partial x} \mathrm{d}y \\ & + \frac{\partial \overline{u}_h(P_l)}{\partial x} \Big[ \int_{K^*_{P_l} \cap K} \frac{\partial(u - \Pi_h u)}{\partial x} \mathrm{d}x\mathrm{d}y \\ & + \int_{\partial K^*_{P_l} \cap K} \frac{\partial(u - \Pi_h u)}{\partial y} (x - x_l) \mathrm{d}x - \frac{\partial(u - \Pi_h u)}{\partial x} (x - x_l) \mathrm{d}y \Big] \\ & + \frac{\partial \overline{u}_h(P_l)}{\partial y} \Big[ \int_{K^*_{P_l} \cap K} \frac{\partial(u - \Pi_h u)}{\partial y} \mathrm{d}x\mathrm{d}y \\ & + \int_{\partial K^*_{P_l} \cap K} \frac{\partial(u - \Pi_h u)}{\partial y} (y - y_l) \mathrm{d}x - \frac{\partial(u - \Pi_h u)}{\partial x} (y - y_l) \mathrm{d}y \Big] \Big\}. \end{aligned} \tag{3.5.24}$$

By the definition of the discrete norm and (3.5.9) we have

$$|\overline{u}_h(P_l) - \overline{u}_h(Q)| \le C |\overline{u}_h|_{1,h,K}, \tag{3.5.25}$$

$$\Big|\frac{\partial \overline{u}_h(P_l)}{\partial x}\Big|, \Big|\frac{\partial \overline{u}_h(P_l)}{\partial y}\Big| \le Ch^{-1} |\overline{u}_h|_{1,h,K}. \tag{3.5.26}$$

Write $\phi_1 = \frac{\partial(u - \Pi_h u)}{\partial x}$, $\phi_2 = \frac{\partial(u - \Pi_h u)}{\partial y}$, $K_l = K^*_{P_l} \cap K$ and $L = \partial K^*_{P_l} \cap K$. Then, it is easy to see that

$$\begin{aligned} & \Big| \int_{K^*_{P_l} \cap K} \frac{\partial(u - \Pi_h u)}{\partial x} \mathrm{d}x\mathrm{d}y \Big| \\ \le & \; h \Big( \int_{K_l} \phi_1^2 \mathrm{d}x\mathrm{d}y \Big)^{1/2} \le h |u - \Pi_h u|_{1,K} \le Ch^4 |u|_{4,K}. \end{aligned} \tag{3.5.27}$$

Similarly

$$\Big|\int_{K^*_{P_l}\cap K}\frac{\partial(u-\Pi_h u)}{\partial y}\mathrm{d}x\mathrm{d}y\Big|\le Ch^4|u|_{4,K}. \qquad (3.5.28)$$

Also note

$$\Big|\int_{\partial K^*_{P_l}\cap K}\frac{\partial(u-\Pi_h u)}{\partial y}\mathrm{d}x-\frac{\partial(u-\Pi_h u)}{\partial x}\mathrm{d}y\Big|\le h^{1/2}\Big[\int_L\big(\phi_1^2+\phi_2^2\big)\mathrm{d}s\Big]^{1/2}, \qquad (3.5.29)$$

$$\Big|\int_{\partial K^*_{P_l}\cap K}\frac{\partial(u-\Pi_h u)}{\partial y}(x-x_l)\mathrm{d}x-\frac{\partial(u-\Pi_h u)}{\partial x}(x-x_l)\mathrm{d}y\Big|$$
$$\le h^{3/2}\Big[\int_L(\phi_1^2+\phi_2^2)\mathrm{d}s\Big]^{1/2}, \qquad (3.5.30)$$

$$\Big|\int_{\partial K^*_{P_l}\cap K}\frac{\partial(u-\Pi_h u)}{\partial y}(y-y_l)\mathrm{d}x-\frac{\partial(u-\Pi_h u)}{\partial x}(y-y_l)\mathrm{d}y\Big|$$
$$\le h^{3/2}\Big[\int_L(\phi_1^2+\phi_2^2)\mathrm{d}s\Big]^{1/2}. \qquad (3.5.31)$$

After the transformation $(x,y)\to(\lambda_j,\lambda_k)$ we have

$$\int_L\phi_i^2\mathrm{d}s\le h\int_{\hat{L}}\hat{\phi}_i^2\mathrm{d}\hat{s},\ i=1,2. \qquad (3.5.32)$$

Use the trace theorem on $\hat{K}^*_Q$ to obtain

$$\Big|\int_{\hat{L}}\hat{\phi}_i^2\mathrm{d}\hat{s}\Big|\le C\|\hat{\phi}_i\|^2_{1,\hat{K}}\le C(h^{-1}|\phi_i|_{0,K}+|\phi_i|_{1,K})^2$$
$$\le C(h^{-1}|u-\Pi_h u|_{1,K}+|u-\Pi_h u|_{2,K})^2\le Ch^4|u|^2_{4,K},\ i=1,2. \qquad (3.5.33)$$

It follows from (3.5.24)-(3.5.33) that

$$|I_K(u-\Pi_h u,\Pi_h^*\bar{u}_h)|\le Ch^3|u|_{4,K}|\bar{u}_h|_{1,h,K}. \qquad (3.5.34)$$

This together with (3.5.23) and Lemma 3.5.1 gives

$$|a(u-\Pi_h u,\Pi_h^*\bar{u}_h)|\le Ch^3|u|_4|\bar{u}_h|_1. \qquad (3.5.35)$$

Combining (3.5.21), (3.5.22) and (3.5.35) implies (3.5.20). This completes the proof. □

## 3.6 $L^2$ and Maximum Norm Estimates

Consider the boundary problem of the second order elliptic equation

$$\begin{cases} -\sum_{i,j=1}^{2} \frac{\partial}{\partial x_i}\Big(a_{ij}\frac{\partial u}{\partial x_j}\Big) + qu = f,\ x \in \Omega, & (3.6.1a) \\ u|_{\partial\Omega} = 0, & (3.6.1b) \end{cases}$$

and its generalized difference scheme: Find $u_h \in U_h$ such that

$$a(u_h, v_h) = (f, v_h),\ \forall v_h \in V_h. \tag{3.6.2}$$

Here $\Omega$ is a polygonal region; the functions $a_{ij} \in W^{1,\infty}(\Omega)$ $(i,j = 1,2)$ and $q \in L^\infty(\Omega)$ satisfy the elliptic condition; $f \in L^2(\Omega)$; $U_h$ is the linear element space related to a quasi-uniform triangulation $T_h$; and $V_h$ is the piecewise constant function space corresponding to the barycenter dual decomposition $T_h^*$. (See §3.2 for details.) In this section we discuss the error estimates in $L^2$ and maximum norms.

### 3.6.1 $L^2$ estimates

Let us introduce an auxiliary problem: Find $w \in H_0^1(\Omega)$ such that

$$a(v, w) = (g, v),\ \forall v \in H_0^1(\Omega). \tag{3.6.3}$$

Assume the problem is regular, namely for any $g \in L^2(\Omega)$ there exist a unique solution $w \in H_0^1(\Omega) \cap H^2(\Omega)$ and a constant $C$ such that

$$||w||_2 \le C||g||_0. \tag{3.6.4}$$

According to the theory of differential equations, problem (3.6.3) is regular when $\Omega$ is convex and the coefficients $a_{ij}$ and the function $q$ are sufficiently smooth.

**Theorem 3.6.1** *Let $u$ be the solution to problem (3.6.1), $u_h$ to the generalized difference scheme (3.6.2), and $u \in H_0^1(\Omega) \cap W^{3,p}(\Omega)$ $(p > 1)$. Then*

$$||u - u_h||_0 \le Ch^2||u||_{3,p}. \tag{3.6.5}$$

**Proof** We use (3.6.3) with $g = u - u_h$ to get

$$||u - u_h||_0^2 = a(u - u_h, w). \tag{3.6.6}$$

Let $\Pi_h w$ and $\Pi_h^* w$ be the interpolation projections of $w$ onto $U_h$ and $V_h$ respectively. Then, obviously the error $u - u_h$ satisfies

$$a(u - u_h, \Pi_h^* w) = 0. \tag{3.6.7}$$

This implies

$$||u - u_h||_0^2 = a(u - u_h, w - \Pi_h w) + a(u - u_h, \Pi_h w) - a(u - u_h, \Pi_h^* w). \tag{3.6.8}$$

Notice that the first two bilinear functionals of the right-hand side of (3.6.8) are in the usual sense while the last one is in the sense of generalized functions. It follows from the boundedness of the bilinear functionals, the $H^1$-estimate (3.2.36) and the interpolation estimates (3.2.8) and (3.6.4) that

$$\begin{aligned} & |a(u - u_h, w - \Pi_h w)| \\ \leq\ & C||u - u_h||_1 ||w - \Pi_h w||_1 \\ \leq\ & Ch^2 ||u||_2 ||u - u_h||_0. \end{aligned} \tag{3.6.9}$$

On the other hand, by Green's formula we have

$$\begin{aligned} & a(u - u_h, \Pi_h w) \\ =\ & \sum_K \int_K \sum_{i,j=1}^{2} a_{ij} \frac{\partial (u - u_h)}{\partial x_j} \frac{\partial \Pi_h w}{\partial x_i} \mathrm{d}x + \int_\Omega q(u - u_h) \Pi_h w \mathrm{d}x \\ =\ & \sum_K \int_K \sum_{i,j=1}^{2} [a_{ij}(x) - a_{ij}(Q)] \frac{\partial (u - u_h)}{\partial x_j} \frac{\partial \Pi_h w}{\partial x_i} \mathrm{d}x \\ & - \sum_K \int_K \sum_{i,j=1}^{2} a_{ij}(Q) \frac{\partial^2 u}{\partial x_i \partial x_j} \Pi_h w \mathrm{d}x \\ & + \sum_K \int_{\partial K} \sum_{i,j=1}^{2} a_{ij}(Q) \frac{\partial (u - u_h)}{\partial x_j} \cos(n, x_i) \Pi_h w \mathrm{d}s \\ & + \int_\Omega q(u - u_h) \Pi_h w \mathrm{d}x, \end{aligned} \tag{3.6.10}$$

$$
\begin{aligned}
&a(u-u_h,\Pi_h^* w)\\
=\;& -\sum_K\sum_{P\in \dot K}\int_{\partial K_P^*\cap K}\sum_{i,j=1}^{2} a_{ij}\frac{\partial(u-u_h)}{\partial x_j}\cos\langle n,x_i\rangle\Pi_h^* w\mathrm{d}s\\
&+\int_\Omega q(u-u_h)\Pi_h^* w\mathrm{d}x,\\
=\;& -\sum_K\sum_{P\in \dot K}\int_{\partial K_P^*\cap K}\sum_{i,j=1}^{2}[a_{ij}(x)-a_{ij}(Q)]\\
&\cdot\frac{\partial(u-u_h)}{\partial x_j}\cos\langle n,x_i\rangle\Pi_h^* w\mathrm{d}s\\
&-\sum_K\int_K\sum_{i,j=1}^{2} a_{ij}(Q)\frac{\partial^2 u}{\partial x_i\partial x_j}\Pi_h^* w\mathrm{d}x\\
&+\sum_K\int_{\partial K}\sum_{i,j=1}^{2} a_{ij}(Q)\frac{\partial(u-u_h)}{\partial x_j}\cos\langle n,x_i\rangle\Pi_h^* w\mathrm{d}s\\
&+\int_\Omega q(u-u_h)\Pi_h^* w\mathrm{d}x.
\end{aligned}
\tag{3.6.11}
$$

Here $Q$ is the barycenter of $K$, and $\dot K = K\cap\dot\Omega_h$. Hence

$$
a(u-u_h,\Pi_h w)-a(u-u_h,\Pi_h^* w)=\sum_{i=1}^{5}E_i(u-u_h,w), \tag{3.6.12}
$$

where

$$
E_1(u-u_h,w)=\sum_K\int_K\sum_{i,j=1}^{2}[a_{ij}(x)-a_{ij}(Q)]\frac{\partial(u-u_h)}{\partial x_j}\frac{\partial\Pi_h w}{\partial x_i}\mathrm{d}x,
$$

$$
\begin{aligned}
&E_2(u-u_h,w)\\
=\;& -\sum_K\sum_{P\in \dot K}\int_{\partial K_P^*\cap K}\sum_{i,j=1}^{2}[a_{ij}(x)-a_{ij}(Q)]\\
&\cdot\frac{\partial(u-u_h)}{\partial x_j}\cos\langle n,x_i\rangle\Pi_h^* w\mathrm{d}s,
\end{aligned}
$$

$$E_3(u-u_h,w) = -\sum_K \int_K \sum_{i,j=1}^2 a_{ij}(Q)\frac{\partial^2 u}{\partial x_i \partial x_j}(\Pi_h w - \Pi_h^* w)\mathrm{d}x,$$

$$E_4(u-u_h,w)$$
$$= \sum_K \int_{\partial K} \sum_{i,j=1}^2 a_{ij}(Q)\frac{\partial(u-u_h)}{\partial x_j}\cos\langle n,x_i\rangle(\Pi_h w - \Pi_h^* w)\mathrm{d}s,$$

$$E_5(u-u_h,w) = \int_Q q(u-u_h)(\Pi_h w - \Pi_h^* w)\mathrm{d}x.$$

Noticing (3.2.36) and (3.6.4) we have

$$|E_1(u-u_h,w)| \le Ch|u-u_h|_1|w|_1 \le Ch^2\|u\|_2\|u-u_h\|_0. \quad (3.6.13)$$

Using the argument in §3.2.4 (cf. the symbols therein and (3.2.44)) we have

$$\begin{aligned}
&|E_2(u-u_h,w)| \\
= \ & \Big|\sum_K \sum_{l=i,j,k} \int_{\overline{M_l Q}} \sum_{i,j=1}^2 [a_{ij}(x) - a_{ij}(Q)] \\
& \cdot \frac{\partial(u-u_h)}{\partial x_j}\cos\langle n,x_i\rangle \mathrm{d}s(w_{l+2} - w_{l+1})\Big| \\
\le \ & Ch^2 \sum_K |u|_{2,K}|w|_{1,K} \le Ch^2\|u\|_2\|u-u_h\|_0.
\end{aligned} \quad (3.6.14)$$

In the case of the barycenter decomposition, we have for any element $K$

$$\int_K (\Pi_h w - \Pi_h^* w)\mathrm{d}x = 0.$$

Write $v_j = \frac{\partial u}{\partial x_j}$ for $j = 1,2$, then $v_j \in W^{2,p}(\Omega)$. By the above equality and the fact that $\frac{\partial \Pi_h v_j}{\partial x_i}$ is a constant on each element $K$, we have

$$E_3(u-u_h,w)$$
$$= \sum_K \int_K \sum_{i,j=1}^2 a_{ij}(Q)\Big(\frac{\partial v_j}{\partial x_i} - \frac{\partial \Pi_h v_j}{\partial x_i}\Big)(\Pi_h w - \Pi_h^* w)\mathrm{d}x.$$

Hence

$$|E_3(u-u_h,w)| \le C\sum_{j=1}^{2}\|v_j-\Pi_h v_j\|_{1,p}\|\Pi_h w-\Pi_h^* w\|_{0,q},$$

where $p,q>1$, $\frac{1}{p}+\frac{1}{q}=1$. Notice the imbedding relations $W^{2,p}(\Omega)\to C(\overline{\Omega})$ and $H^2(\Omega)\to W^{1,q}(\Omega)$, and use the interpolation theorem and (3.6.4), then we have

$$|E_3(u-u_h,w)| \le Ch^2\sum_{j=1}^{2}\|v_j\|_{2,p}\|w\|_{1,q} \le Ch^2\|u\|_{3,p}\|u-u_h\|_0. \tag{3.6.15}$$

Now let us estimate $E_4$. A direct calculation shows that along any a line $L$ of an element $K$ we have

$$\int_L(\Pi_h w-\Pi_h^* w)\mathrm{d}s=0.$$

Furthermore, since $\frac{\partial u_h}{\partial x_j}\cos\langle n,x_l\rangle$ is a constant on $L$,

$$\sum_K\int_{\partial K}\sum_{l,m=1}^{2}a_{lm}(Q)\frac{\partial u_h}{\partial x_m}\cos\langle n,x_l\rangle(\Pi_h w-\Pi_h^* w)\mathrm{d}s=0.$$

So

$$E_4=\sum_K\int_{\partial K}\sum_{l,m=1}^{2}a_{lm}(Q)\frac{\partial u}{\partial x_m}\cos\langle n,x_l\rangle(\Pi_h w-\Pi_h^* w)\mathrm{d}s.$$

Because we have zero boundary condition on the outer boundary, we only have to consider the integrals on the inner boundaries. Let $K=K_Q$ (see Fig. 3.2.7), and $K_{Q'}$ is a neighbouring element sharing a common side $\overline{P_iP_j}$ with $K_Q$. Denote by $n_Q$ the unit outer normal direction of $K_Q$ along $\overline{P_iP_j}$. Obviously, the two line integrals along $\overline{P_iP_j}$ differ in the sign. Note that $\frac{\partial \Pi_h u}{\partial x_m}$ is also a constant on $\overline{P_iP_j}$. Therefore, we may write $E_4$ in the form

$$\begin{aligned}E_4=&\sum_{K_Q}\sum_{\overline{P_iP_j}}\int_{\overline{P_iP_j}}\sum_{l,m=1}^{2}(a_{lm}(Q)-a_{lm}(Q'))\Big(\frac{\partial u}{\partial x_m}-\frac{\partial \Pi_h u}{\partial x_m}\Big)\\&\cdot\cos\langle n_Q,x_l\rangle(\Pi_h w-\Pi_h^* w)\mathrm{d}s.\end{aligned}$$

It follows from the smoothness of $a_{lm}$ and the Cauchy inequality that

$$\begin{aligned}|E_4| \le\ & Ch\sum_{K_Q}\sum_{\overline{P_iP_j}}\Big(\sum_{m=1}^{2}\int_{\overline{P_iP_j}}\Big|\frac{\partial u}{\partial x_m}-\frac{\partial \Pi_h u}{\partial x_m}\Big|^2 \mathrm{d}s\Big)^{1/2}\\ &\cdot\Big(\int_{\overline{P_iP_j}}|\Pi_h w-\Pi_h^* w|^2 \mathrm{d}s\Big)^{1/2}.\end{aligned} \tag{3.6.16}$$

As in the proof to (3.2.43) we can use the interpolation property to show that

$$\begin{aligned}&\int_{\overline{P_iP_j}}|\Pi_h w-\Pi_h^* w|^2 \mathrm{d}s\\ \le\ & Ch\Big(h^{-1}|\Pi_h w-\Pi_h^* w|_{0,K_Q\cap K_{P_i}^*}+h^{-1}|\Pi_h w-\Pi_h^* w|_{0,K_Q\cap K_{P_j}^*}\\ &+|\Pi_h w-\Pi_h^* w|_{1,K_Q\cap K_{P_i}^*}+|\Pi_h w-\Pi_h^* w|_{1,K_Q\cap K_{P_j}^*}\Big)^2\\ \le\ & Ch|\Pi_h w|_{1,K_Q}^2 \le Ch\|w\|_{2,K_Q}^2 \le Ch\|u-u_h\|_{0,K_Q}^2.\end{aligned} \tag{3.6.17a}$$

Similarly

$$\int_{\overline{P_iP_j}}\Big|\frac{\partial u}{\partial x_m}-\frac{\partial \Pi_h u}{\partial x_m}\Big|^2 \mathrm{d}s \le Ch|u|_{2,K_Q}^2. \tag{3.6.17b}$$

Now we insert (3.6.17) into (3.6.16) to obtain

$$|E_4| \le Ch^2\|u\|_2\|u-u_h\|_0. \tag{3.6.18}$$

The estimate of $E_5$ is simple:

$$\begin{aligned}&|E_5(u-u_h,w)|\\ \le\ & C\|u-u_h\|_1\|\Pi_h w-\Pi_h^* w\|_0\\ \le\ & Ch^2\|u\|_2\|u-u_h\|_0.\end{aligned} \tag{3.6.19}$$

A combination of (3.6.13,14,15,18,19) yields

$$|a(u-u_h,\Pi_h w)-a(u-u_h,\Pi_h^* w)| \le Ch^2\|u\|_{3,p}\|u-u_h\|_0. \tag{3.6.20}$$

Finally, (3.6.5) results from (3.6.8), (3.6.9) and (3.6.20). This completes the proof. □

### 3.6.2 A maximum estimate and some remarks

**Theorem 3.6.2** *Under the conditions of Theorem 3.6.1 there holds the following error estimate in maximum norm*

$$||u - u_h||_{0,\infty} \le Ch||u||_{3,p} \; (p > 1). \tag{3.6.21}$$

**Proof** First we note that

$$||u - u_h||_{0,\infty} \le ||u - \Pi_h u||_{0,\infty} + ||\Pi_h u - u_h||_{0,\infty}. \tag{3.6.22}$$

For the first term on the right-hand side of (3.6.22) there exists an element $K$ such that

$$||u - \Pi_h u||_{0,\infty} = ||u - \Pi_h u||_{0,\infty,K}. \tag{3.6.23}$$

The Sobolev interpolation approximation theorem gives

$$||u - \Pi_h u||_{0,\infty,K} \le Ch|u|_{2,K} \le Ch|u|_2. \tag{3.6.24}$$

For the second term, the inverse property of the finite element implies that

$$||\Pi_h u - u_h||_{0,\infty} \le Ch^{-1}||\Pi_h u - u_h||_0. \tag{3.6.25}$$

The approximation theorem and the $L^2$ estimate (3.6.5) result in

$$||\Pi_h u - u_h||_0 \le ||\Pi_h u - u||_0 + ||u - u_h||_0 \le Ch^2||u||_{3,p}. \tag{3.6.26}$$

Combining (3.6.22)-(3.6.26) yields (3.6.21). □

**Remark 1** In Theorems 3.6.1 and 3.6.2 we obtain the error estimates of precisely the same optimal orders as those of the linear finite element method, but we require higher smoothness of the solutions. The reason behind it may be that we can not obtain the approximation order of the derivatives in the piecewise constant function spaces.

**Remark 2** The conclusions of this section also hold for the quasi-uniform rectangular mesh and the corresponding center dual decomposition. To do this we only have to slightly revise the argument by taking $\Pi_h$ as the bilinear interpolation operator on the rectangular mesh, and to note that on the rectangle $K$ we have

$$\int_K (\Pi_h w - \Pi_h^* w)\mathrm{d}x = 0.$$

**Remark 3** When the test function space $V_h$ is the piecewise linear function space, the dual argument of the finite element method is still valid, and thus we can deduce the same $L^2$ estimates as those of the finite element method. (cf. Theorem 2.5.2.)

## 3.7 Superconvergences

This section is devoted to the superconvergence of the solution to the generalized difference scheme (3.6.2) approximating the second order elliptic boundary problem (3.6.1). We take the linear element space as the trial function space $U_h$. The test function space is chosen as the piecewise constant function space related to the circumcenter dual decomposition $T_h$. Our results in this section are also valid for the case of the barycenter dual decomposition.

### 3.7.1 Weak estimate of interpolations

Now we derive an interpolation weak estimate of the bilinear form corresponding to the generalized difference scheme.

**Theorem 3.7.1** *Suppose $T_h$ is a uniform decomposition, that is, the union of any a pair of adjacent triangular elements forms a parallelogram. Also assume $u \in H_0^1(\Omega) \cap H^3(\Omega)$, $w_h \in U_h$. Then*

$$a(u - \Pi_h u, \Pi_h^* w_h)| \le Ch^2 ||u||_3 ||w_h||_1. \tag{3.7.1}$$

**Proof** Take any a triangular element $K_Q \in T_h$ with vertexes $P(x_{1l}, x_{2l})$ $(l = i, j, k)$, circumcenter $Q$ and the midpoints $M_l$ $(l = i, j, k)$ (cf. Fig. 3.7.1). Then we have (cf. (3.2.39) and (3.2.40))

$$a(u - \Pi_h u, \Pi_h^* w_h) = E_1(u, w_h) + E_2(u, w_h), \tag{3.7.2}$$

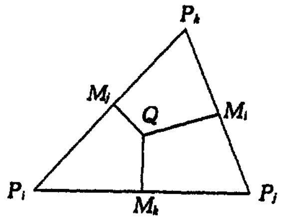


Fig. 3.7.1

where

$$
\begin{aligned}
&E_1(u, w_h) \\
&= \sum_K \sum_{l=i,j,k} [w_h(P_{l+2}) - w_h(P_{l+1})] \\
&\quad \int_{\overline{M_l Q}} \sum_{m,n=1}^{2} a_{mn} \frac{\partial(u - \Pi_h u)}{\partial x_n} \cos\langle n_l, x_m\rangle \mathrm{d}s,
\end{aligned}
\tag{3.7.3}
$$

$$
E_2(u, w_h) = \sum_K \sum_{l=i,j,k} w_h(P_l) \int_{K^*_{P_l} \cap K} q(u - \Pi_h u)\mathrm{d}x, \tag{3.7.4}
$$

where $n_l$ is the outer normal vector along the boundary $\overline{M_l Q}$ of the region $K^*_{P_{l+1}} \cap K$ and we make the convention that $i+1 = j, j+1 = k, k+1 = i$.

Employing the discrete norm (3.2.22) we have

$$
|E_2(u, w_h)| \le C \sum_K ||u - \Pi_h u||_{0,K} ||w_h||_{0,h,K} \le Ch^2 |u|_2 ||w_h||_0. \tag{3.7.5}
$$

In order to estimate $E_1(u, w_h)$, we discuss the two cases where the side $P_j P_k$ belongs to the boundary of $\Omega$ or is the common side of two adjacent elements $K_Q$ and $K_{Q'}$, respectively. In the former case, the corresponding terms in (3.7.3) vanish since $w_h|_{\partial\Omega} = 0$. So we concentrate on the latter case. In this case $K_Q \cup K_{Q'}$ is a parallelogram by the uniformity of the decomposition (cf. Fig. 3.7.2). So

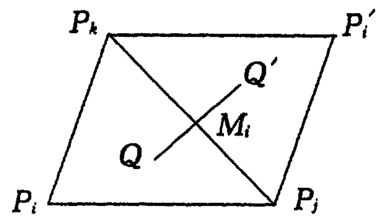



Fig. 3.7.2

we have

$$E_1(u, w_h) = \sum_{\overline{P_jP_k} \not\subset \partial\Omega} [w_h(P_k) - w_h(P_j)] \cdot \int_L \sum_{m,n=1}^{2} a_{mn} \frac{\partial(u - \Pi_h u)}{\partial x_n} \cos\langle n_l, x_m\rangle \mathrm{d}s, \tag{3.7.6}$$

where $\sum\limits_{\overline{P_jP_k} \not\subset \partial\Omega}$ means the summation over all the sides $\overline{P_jP_k}$ which do not belong to $\partial\Omega$, and $L = Q'M_l \cup M_lQ$. It is easy to see that

$$\begin{aligned}
&\Big|\int_L a_{mn} \frac{\partial(u - \Pi_h u)}{\partial x_n} \cos\langle n_l, x_m\rangle \mathrm{d}s\Big| \\
&= \Big|a_{mn}(Q) \int_L \frac{\partial(u - \Pi_h u)}{\partial x_n} \cos\langle n_l, x_m\rangle \mathrm{d}s \\
&\quad + \int_L [a_{mn}(x) - a_{mn}(Q)] \frac{\partial(u - \Pi_h u)}{\partial x_n} \cos\langle n_l, x_m\rangle \mathrm{d}s\Big| \\
&\le C \sum_{m=1}^{2} \Big|\int_L \frac{\partial(u - \Pi_h u)}{\partial x_n} \mathrm{d}x_m\Big| + Ch \int_L \Big|\frac{\partial(u - \Pi_h u)}{\partial x_n}\Big| \mathrm{d}s.
\end{aligned} \tag{3.7.7}$$

Perform the affine transformation

$$x_m = x_{mi} + (x_{mj} - x_{mi})\lambda_1 + (x_{mk} - x_{mi})\lambda_2, \; m = 1, 2.$$

Then the parallelogram $K_Q \cup K_{Q'}$ becomes a square $I = \{(\lambda_1, \lambda_2) : 0 \le \lambda_1 \le 1, 0 \le \lambda_2 \le 1\}$. Suppose the points $P_l, M_l$ and $Q$ are mapped onto $\hat{P}_l, \hat{M}_l$ and $\hat{Q}$ respectively; the segment $L$ is mapped onto $\hat{L}$ and the function $u - \Pi_h u$ onto $\hat{u} - \hat{\Pi}_h \hat{u}$. Then (cf. (3.2.6) and

(3.2.7))

$$
\begin{aligned}
&\left|\int_L \frac{\partial(u-\Pi_h u)}{\partial x_n}\mathrm{d}x_m\right| \\
&= \left|\int_{\hat L} \frac{1}{2S_K}\left[(x_{mk}-x_{mi})\frac{\partial(\hat u-\hat\Pi_h\hat u)}{\partial\lambda_1}+(x_{mi}-x_{mj})\frac{\partial(\hat u-\hat\Pi_h\hat u)}{\partial\lambda_2}\right]\right. \\
&\quad \left.\cdot[(x_{mj}-x_{mi})\mathrm{d}\lambda_1+(x_{mk}-x_{mi})\mathrm{d}\lambda_2]\right|. \\
&\le C\sum_{m,n=1}^{2}\left|\int_{\hat L}\frac{\partial(\hat u-\hat\Pi_h\hat u)}{\partial\lambda_n}\mathrm{d}\lambda_m\right|,
\end{aligned}
\tag{3.7.8}
$$

where $S_K$ is the area of $K$. Write

$$
J(\hat u)=\int_{\hat L}\frac{\partial(\hat u-\hat\Pi_h\hat u)}{\partial\lambda_n}\mathrm{d}\lambda_m. \tag{3.7.9}
$$

For any $P\in\mathcal{P}_2(I)$, we note that $\frac{\partial P}{\partial\lambda_n}$ is linear and $\frac{\partial\hat\Pi_h P}{\partial\lambda_n}$ is a constant in $\hat K_Q$ as well as in $\hat K'_Q$. Thus

$$
\int_{\hat L}\frac{\partial P}{\partial\lambda_n}\mathrm{d}\lambda_m=\frac{\partial P(\frac12,\frac12)}{\partial\lambda_n}L_m,
$$

$$
\begin{aligned}
&\int_{\hat L}\frac{\partial\hat\Pi_h P}{\partial\lambda_n}\mathrm{d}\lambda_m=\frac{\partial\hat\Pi_h P}{\partial\lambda_n}\Big|_{\hat K_Q}\cdot\frac{L_m}{2}+\frac{\partial\hat\Pi_h P}{\partial\lambda_n}\Big|_{\hat K_{Q'}}\cdot\frac{L_m}{2} \\
&=[P(1,0)-P(0,0)+P(1,1)-P(0,1)]L_m/2 \\
&=\frac{\partial P(\frac12,\frac12)}{\partial\lambda_n}L_m,
\end{aligned}
$$

where $L_m=\lambda_m(\hat Q')-\lambda_m(\hat Q)$. Hence

$$
J(P)=0,\ \forall P\in\mathcal{P}_2(I). \tag{3.7.10}
$$

This together with the trace theorem gives

$$
|J(\hat u)|=|J(\hat u-P)|\le\|\hat u-P\|_{3,I}.
$$

By the quotient space norm theorem and the relationship of the Sobolev semi-norms before and after the affine transformation we have

$$|J(\hat{u})| \le C \inf_{P\in\mathcal{P}_2(I)} ||\hat{u} - P||_{3,I} \le C|\hat{u}|_{3,I} \le Ch^2|u|_{3,K_Q\cup K_{Q'}}. \quad (3.7.11)$$

It follows from (3.7.8), (3.7.9) and (3.7.11) that

$$\Big|\int_L \frac{\partial(u-\Pi_h u)}{\partial x_n}\mathrm{d}x_m\Big| \le Ch^2|u|_{3,K_Q\cup K_{Q'}}. \quad (3.7.12)$$

On the other hand, set $\phi = \frac{\partial(u-\Pi_h u)}{\partial x_n}, L_1 = M_iQ$ and $\hat{\phi}(\lambda_1,\lambda_2) = \phi(x_1,x_2)$, then obviously

$$\int_{L_1}\Big|\frac{\partial(u-\Pi_h u)}{\partial x_n}\Big|\mathrm{d}s \le Ch\int_{\hat{L}_1}|\hat{\phi}|\mathrm{d}\hat{s} \le Ch\Big(\int_{\hat{L}_1}|\hat{\phi}|^2\mathrm{d}\hat{s}\Big)^{1/2}. \quad (3.7.13)$$

Using the trace theorem on $\hat{K}^*_{P_j}\cap\hat{K}$ implies the existence of a constant $C$ independent of $K$ such that (cf. §3.2.4)

$$\Big(\int_{\hat{L}_1}|\hat{\phi}|^2\mathrm{d}\hat{s}\Big)^{1/2} \le C||\hat{\phi}||_{1,\hat{K}}.$$

Note

$$|\hat{\phi}|_{0,\hat{K}} \le Ch^{-1}|\phi|_{0,K},\ |\hat{\phi}|_{1,\hat{K}} \le C|\phi|_{1,K}.$$

Hence

$$\int_{L_1}\Big|\frac{\partial(u-\Pi_h u)}{\partial x_n}\Big|\mathrm{d}s \le Ch(h^{-1}|\phi|_{0,K} + |\phi|_{1,K}) \le Ch|u|_{2,K}. \quad (3.7.14)$$

It follows from (3.7.7), (3.7.12) and (3.7.14) that

$$\Big|\int_L a_{mn}\frac{\partial(u-\Pi_h u)}{\partial x_n}\cos\langle n_l, x_m\rangle\mathrm{d}s\Big| \le Ch^2||u||_{3,K_Q\cup K_{Q'}}. \quad (3.7.15)$$

Noting the linearity of $w_h$ on $K_Q$ we have

$$\begin{aligned} &|w_h(P_k) - w_h(P_j)| \\ &= \Big|[x_1(P_k) - x_1(P_j)]\frac{\partial w_h}{\partial x_1} + [x_2(P_k) - x_2(P_j)]\frac{\partial w_h}{\partial x_2}\Big| \\ &\le h\Big(\Big|\frac{\partial w_h}{\partial x_1}\Big| + \Big|\frac{\partial w_h}{\partial x_2}\Big|\Big) \le C|w_h|_{1,K}. \end{aligned} \quad (3.7.16)$$

By (3.7.6), (3.7.15) and (3.7.16) we have

$$|E_1(u,w_h)| \le Ch^2||u||_3||w_h||_1. \qquad (3.7.17)$$

Finally (3.7.1) is obtained by (3.7.2), (3.7.5) and (3.7.17). □

Next, we try to relax the restriction on the decomposition. We say that a quasi-uniform triangulation $T_h$ is a $C$-uniform decomposition if for each pair of adjacent nodes $P$ and $Q$, and the side $PP'$ ($P' \neq Q$) with $P$ as its endpoint, there exists a side $QQ'$ of another triangular element with $Q$ as its endpoint such that $PP'QQ'$ forms a quasi-parallelogram, namely there exists a constant $C$ independent of $h$ such that

$$||\overline{PP'}| - |\overline{QQ'}|| \le Ch^2.$$

**Theorem 3.7.2** *Assume $T_h$ is a $C$-uniform decomposition and $u \in H_0^1(\Omega) \cap H^3(\Omega) \cap W^{2,\infty}(\Omega)$, $w_h \in U_h$. Then*

$$|a(u - \Pi_h u, \Pi_h^* w_h)| \le Ch^2(||u||_3 + ||u||_{2,\infty})||w_h||_1. \qquad (3.7.18)$$

**Proof** Since $T_h$ is a $C$-uniform decomposition, the union of any two adjacent element $K_Q$ and $K_{Q'}$ is a quasi-parallelogram. Modify $K_{Q'}$ to get $K_{Q''}$ such that $K_Q \cap K_{Q''}$ is a parallelogram. Then $|Q'Q''|$, the distance between the circumcenters of $K_{Q'}$ and $K_{Q''}$ respectively, is $O(h^2)$. By Theorem 3.7.1, it only remains to estimate the following

$$E_3(u,w_h) = \sum_{P_jP_k \not\subset \partial\Omega} [w_h(P_k) - w_h(P_j)] \cdot \int_{\overline{Q'Q''}} \sum_{m,n=1}^{2} a_{mn} \frac{\partial(u - \Pi_h u)}{\partial x_n} \cos\langle n_l, x_m\rangle ds. \qquad (3.7.19)$$

It is easy to see that

$$|E_3(u,w_h)| \le C \sum_{P_jP_k \not\subset \partial\Omega} h^3 |u|_{2,\infty}|w_h|_{1,K} \le Ch^2|u|_{2,\infty}|w_h|_1. \quad (3.7.20)$$

This implies the desired result and completes the proof. □

The next theorem extends the weak estimation from the case of $C$-uniform decomposition to the case of piecewise $C$-uniform decomposition.

**Theorem 3.7.3** *Let $T_h$ be a piecewise $C$-uniform decomposition, where the pieces are divided by several line segments which connect some of the vertexes of $\Omega$ and do not intersect each other inside the region. Assume $u \in H_0^1(\Omega) \cap W^{3,\infty}(\Omega)$, $w_h \in U_h$. Then*

$$|a(u - \Pi_h u, \Pi_h^* w_h)| \leq Ch^2 ||u||_{3,\infty} ||w_h||_1. \tag{3.7.21}$$

**Proof** By Theorem 3.7.2 it only remains to show that

$$\begin{aligned} E_4(u, w_h) = & \sum{}' [w_h(P_k) - w_h(P_j)] \\ & \cdot \int_{\overline{M_l Q}} \sum_{m,n=1}^{2} a_{mn} \frac{\partial(u - \Pi_h u)}{\partial x_n} \cos\langle n_l, x_m \rangle \mathrm{d}s, \end{aligned} \tag{3.7.22}$$

where $\sum'$ denotes the summation for the cases where $\overline{P_j P_k}$ belongs to the line segments dividing the pieces. Since the number of these line segments is finite, we only have to consider the case where $\overline{P_j P_k}$ belongs to a certain segment and $\overline{M_i Q}$'s are on the same side of the segment. So we suppose on one side of the segment $P_0 P_N$ there are $N$ elements as in Fig. 3.7.3 and correspondingly

$$\begin{aligned} E_4^{(1)}(u, w_h) = & \sum_{i=1}^{N} [w_h(P_i) - w_h(P_{i-1})] \\ & \cdot \int_{\overline{M_i Q_i}} \sum_{m,n=1}^{2} a_{mn} \frac{\partial(u - \Pi_h u)}{\partial x_n} \cos\langle \tau, x_m \rangle \mathrm{d}s \\ = & \sum_{i=1}^{N-1} w_h(P_i) \Big( \int_{\overline{M_i Q_i}} - \int_{\overline{M_{i+1} Q_{i+1}}} \Big) \\ & \cdot \sum_{m,n=1}^{2} a_{mn} \frac{\partial(u - \Pi_h u)}{\partial x_n} \cos\langle \tau, x_m \rangle \mathrm{d}s, \end{aligned} \tag{3.7.23}$$

where we have used the boundary values $w_h(P_0) = w_h(P_N) = 0$ and the notation $\tau = \overline{P_{i-1} P_i} / |\overline{P_{i-1} P_i}|$. By the argument in the proof to Theorem 3.7.2 we can assume without loss of generality that the elements are all equal to each other. So $K_{Q_i}$ will overlap $K_{Q_{i+1}}$ by displacing it along $\overline{P_0P_N}$:

$$K_{i+1} = K_i + h_1\tau \quad (h_1 = |\overline{P_{i-1}P_i}|).$$

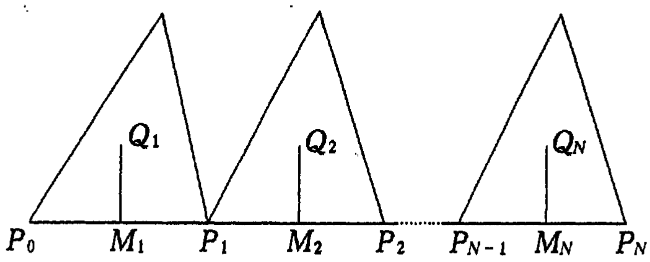



Fig. 3.7.3

Thus

$$\begin{aligned}&\Big(\int_{\overline{M_iQ_i}} - \int_{\overline{M_{i+1}Q_{i+1}}}\Big) a_{mn}\frac{\partial(u-\Pi_h u)}{\partial x_n}\cos\langle\tau, x_m\rangle \mathrm{d}s\\ &= -\int_{\overline{M_iQ_i}}\Big[a_{mn}(x+h_1\tau)\frac{\partial(u-\Pi_h u)}{\partial x_n}(x+h_1\tau)\\ &\quad -a_{mn}(x)\frac{\partial(u-\Pi_h u)}{\partial x_n}(x)\Big]\cos\langle\tau, x_m\rangle \mathrm{d}s.\end{aligned} \tag{3.7.24}$$

Thus the fact $a_{mn} \in W^{1,\infty}(\Omega)$ implies that

$$\begin{aligned}&\Big|\Big(\int_{\overline{M_iQ_i}} - \int_{\overline{M_{i+1}Q_{i+1}}}\Big) a_{mn}\frac{\partial(u-\Pi_h u)}{\partial x_n}\cos\langle\tau, x_m\rangle \mathrm{d}s\Big|\\ &\le C\int_{\overline{M_iQ_i}}\Big|\frac{\partial(u-\Pi_h u)}{\partial x_n}(x+h_1\tau) - \frac{\partial(u-\Pi_h u)}{\partial x_n}(x)\Big|\mathrm{d}s\\ &\quad +Ch\int_{\overline{M_iQ_i}}\Big|\frac{\partial(u-\Pi_h u)}{\partial x_n}(x)\Big|\mathrm{d}s\\ &\le Ch^3\|u\|_{3,\infty}.\end{aligned} \tag{3.7.25}$$

It results from (3.7.23) and (3.7.25) that

$$|E_4^{(1)}(u - w_h)| \le Ch^3 \sum_{i=1}^{N-1} |w_h(P_i)| \cdot ||u||_{3,\infty} \le Ch^2 ||u||_{3,\infty} ||w_h||_0.$$

This leads to (3.7.21) and completes the proof. □

### 3.7.2 Superconvergence estimates

**Theorem 3.7.4** *Let $u$ be the generalized solution to the second order elliptic boundary value problem (3.6.1) and $u_h \in U_h$ the solution to the generalized difference scheme (3.6.2). Then under the assumptions of Theorems 3.7.1, 3.7.2 and 3.7.3, respectively, we have the following estimates*

$$||u_h - \Pi_h u||_1 \le Ch^2 ||u||_3, \tag{3.7.26}$$

$$||u_h - \Pi_h u||_1 \le Ch^2 (||u||_3 + ||u||_{2,\infty}), \tag{3.7.27}$$

$$||u_h - \Pi_h u||_1 \le Ch^2 ||u||_{3,\infty}. \tag{3.7.28}$$

*So*

$$\Big(\frac{1}{r} \sum_{P_0 \in M_h} |\overline{\nabla}(u - u_h)(P_0)|^2\Big)^{1/2} = O(h^2). \tag{3.7.29}$$

*Here $M_h$ stands for the set of the optimal stress points of the $U_h$ interpolation (cf. §2.6), $\overline{\nabla}$ the average of the gradient over the elements containing $P_0$, and $r$ the number of points in $M_h$. Therefore, the generalized difference method has the same optimal stress points as the finite element method.*

**Proof** The conclusion follows directly from Theorems 3.7.1-3 of this section, and Theorem 2.6.1. □

## Bibliography and Comments

The papers [A-30,62] (cf. [B-57]) extend the generalized difference method to the boundary value problems of second order elliptic partial differential equations on planar regions, discuss the generalized difference methods on triangular and quadrilateral meshes respectively, and propose the basic idea of the method. Error estimates are derived for the cases where the trial function spaces are chosen as piecewise linear and bilinear function spaces, and the test function spaces as piecewise constant function spaces respectively (cf. §3.2 and §3.3). These results further support the following opinion: The convergence order of the generalized difference method is determined mainly by the trial function space, while the test function space influences only the coefficients in the error estimate. These papers have drawn people's attention to the generalized difference method for multidimensional problems (cf. the related references in the end of this book).

The paper [A-55] constructs, using the hierarchical meshes, several generalized difference schemes (including a five-point quadratic scheme, a nine-point bilinear scheme, and a nine-point bi-quadratic scheme etc.) for second order elliptic equations. A numerical analysis is carried out there for the nine-point biquadratic scheme. Recently, [B-62] discusses the generalized difference schemes on arbitrary quadrilateral grids and presents the optimal order convergence estimates. The paper [B-58] applies the generalized difference method to a nonlinear elliptic Dirichlet problem and gives an error estimate (cf. §4.5). In paper [B-55], a $W^{1,p}$-estimate and an $L^2$-estimate of the generalized difference method for second order elliptic equations are studied. Some high order element generalized difference methods on triangular meshes are discussed in [A-41] and [B-14], including the schemes based on the Lagrangian quadratic element and the Hermitian cubic element respectively. The same optimal order error estimates as those of the finite element method are obtained (cf. §3.4 and §3.5). The paper [B-15] investigates the superconvergence of the generalized difference method and shows that the linear element generalized difference method has the same optimal stress points as those of the linear finite element method (cf. §3.7). An $L^2$-estimate with an optimal order is proved for the linear element generalized difference method in [A-10] (cf. §3.6). It can be seen from the discussion of this chapter that generalized difference methods enjoy the same $H^1$-estimates as those of finite element methods, save that we require a little bit higher smoothness of the solution for the $L^2$-estimate to hold for the linear element scheme.

Finite difference and finite element methods are the two most effective and popular numerical methods for partial differential equations. The finite element method has several remarkable advantages such as that its decompositions is flexible to effectively approximate irregular regions; that it is easy to deal with the boundary conditions as well as the intersection of different media; that it may use high order elements to get a better accuracy without involving too many nodes; that the approximate solution may converge to the generalized solution but not necessarily the classical solution; and that one can use the functional analysis and the Sobolev space theory to provide a systematical numerical analysis. On the other hand, the finite difference method also possesses some useful advantages such as that the construction of the schemes is simple, that the discretization near the node is local and intuitive; and that the computational effort is much less for the same accuracy. But the classical difference method does not share these advantages mentioned above, of the finite element method.

A lot of research has been devoted to the reformation of the classical difference method. A popular approach is to start from the integral conservative forms of boundary value problems and to use numerical integrations to construct conservative difference schemes on irregular networks. From various applied or theoretical points of view, these difference schemes are given different names such as the finite difference method on irregular networks, the box integration method, the balance method, the finite control volume method, the finite volume method, the discretization operator method, and the multi-element balancing method (cf. the references in the end of this book). Most of these methods have mechanical or physical backgrounds and reflect the conservation or the balance of the mechanical system or other physical systems on the element. From the numerical analysis point of view, these methods are basically regarded as the integral interpolation methods. These difference methods constructed through the integration interpolation over irregular networks possess many advantages of the finite element methods and are effective methods for the numerical computation of partial differential equations. But it is still difficult to construct through this approach the difference schemes with high accuracy.

The generalized difference methods proposed in this chapter essentially reform the integration interpolation methods by absorbing more ideas and tricks of the finite element methods. The first step is to write the boundary value problem into a generalized variational form (the generalized integral conservative form). The second step is to choose the finite element spaces as the trial function spaces (the idea of the finite element method) and the common terms of the local Taylor expansion as the test function spaces (the idea of the difference method). By doing so we keep to the utmost the simplicity of the difference method, while we can construct the difference schemes with high convergence order like the finite element methods and we can employ more results and tricks of the finite element methods in the error analysis.

We end this section by some problems for further investigation.

**Problem 1.** Further investigate the error estimates in $L^2$ and maximum norms of the generalized difference methods (the high order element schemes).

**Problem 2.** Further investigate the superconvergence theory of the generalized difference methods, including the superconvergence of the displacement and the optimal stress point theorem for high order element difference schemes.

**Problem 3.** Apply the extrapolation method to generalized difference methods and build up the corresponding theory.

**Problem 4.** Establish the general error estimates in Sobolev spaces for higher order element generalized difference schemes.

# Chapter 4

# FOURTH ORDER AND NONLINEAR ELLIPTIC EQUATIONS

In this chapter we first consider the mixed and the nonconforming generalized difference methods for fourth order elliptic equations, taking a biharmonic equation as an example. Then we discuss the generalized difference method for a class of nonlinear elliptic equations.

## 4.1 Mixed Generalized Difference Methods Based on Ciarlet-Raviart Variational Principle

Consider the following Dirichlet problem of the biharmonic equation:

$$\begin{cases} \Delta^2 u \equiv \dfrac{\partial^4 u}{\partial x_1^4} + 2\dfrac{\partial^4 u}{\partial x_1^2 \partial x_2^2} + \dfrac{\partial^4 u}{\partial x_2^4} = f, \ (x_1, x_2) \in \Omega, & (4.1.1a) \\ u = \dfrac{\partial u}{\partial n} = 0, \ (x_1, x_2) \in \partial\Omega, & (4.1.1b) \end{cases}$$

where $\Omega$ is a convex polygon region on the plane with boundary $\partial\Omega$, $n$ the unit outer normal vector of $\partial\Omega$, and $f \in L^2(\Omega)$. This kind of boundary problem occupies an important position in, e.g., elastic mechanics and water kinetics. It is well-known that if $f \in H^{-1}(\Omega)$, then problem (4.1.1) is regular, i.e., there exists a unique solution $u \in H_0^2(\Omega) \cap H^3(\Omega)$ and a constant $C$ such that

$$\|u\|_3 \le C\|f\|_{-1}, \ \forall f \in H^{-1}(\Omega). \tag{4.1.2}$$

Introduce a new unknown function $v = -\Delta u$ to reduce problem (4.1.1) into a second order equation:

$$\begin{cases} -\Delta u = v, & (x_1, x_2) \in \Omega, & (4.1.3a) \\ -\Delta v = f, & (x_1, x_2) \in \Omega, & (4.1.3b) \\ u = \dfrac{\partial u}{\partial n} = 0, & (x_1, x_2) \in \partial\Omega. & (4.1.3c) \end{cases}$$

Multiply (4.1.3a) and (4.1.3b) by $\psi \in H^1(\Omega)$ and $\phi \in H_0^1(\Omega)$ respectively, integrate them on $\Omega$, and use Green's formula and the boundary condition (4.1.3c) to obtain a corresponding variational form: Find $(u, v) \in H_0^1(\Omega) \times H^1(\Omega)$ such that

$$\begin{cases} a(u, \psi) = (v, \psi), & \forall \psi \in H^1(\Omega), & (4.1.4a) \\ a(v, \phi) = (f, \phi), & \forall \phi \in H_0^1(\Omega), & (4.1.4b) \end{cases}$$

where

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v dx,$$

$$(f, \phi) = \int_\Omega f\phi dx.$$

We can use the regularity condition of (4.1.1) to show the equivalence (the Ciarlet-Raviart variational principle [B-18]) of (4.1.1) and (4.1.4), that is, if $u$ is the solution to (4.1.1) and $v = -\Delta u$ then $(u, v)$ is the solution to (4.1.4); and if $(u, v)$ is the solution to (4.1.4) then $u$ is the solution to (4.1.1) and $v = -\Delta u$.

### 4.1.1 Mixed generalized difference equations

As in §3.2, let $T_h$ be a quasi-uniform grid and $T_h^*$ the corresponding barycenter dual grid. Suppose $U_h$ is the piecewise linear function space with respect to $T_h$:

$$U_h = \{u_h \in C(\overline{\Omega}) : u_h|_K \in \mathcal{P}_1, \ \forall K \in T_h\},$$

and $V_h \in L^2(\Omega)$ is the piecewise constant function space corresponding to $T_h^*$:

$$V_h = \{v_h : \ v_h \text{ is a constant on the interior of each } K^* \in T_h^*\}.$$

Set

$$U_{0h} = \{u_h \in U_h : \ u_h(P_0) = 0, \ \forall P_0 \in \overline{\Omega}_h \cap \partial\Omega\},$$
$$V_{0h} = \{v_h \in V_h : \ v_h(P_0) = 0, \ \forall P_0 \in \overline{\Omega}_h \cap \partial\Omega\},$$

where $\bar{\Omega}_h$ stands for the set of the nodes of $T_h$.

Multiply (4.1.3a) and (4.1.3b) by $\psi_h \in V_h$ and $\phi_h \in V_{0h}$ respectively, and integrate them on $\Omega$ to obtain

$$(-\Delta u, \psi_h) = (v, \psi_h),$$

$$(-\Delta v, \phi_h) = (f, \phi_h).$$

Applying Green's formula to each dual element and noting the boundary condition (4.1.2c) yield

$$(-\Delta u, \psi_h) = -\sum_{P_0 \in \overline{\Omega}_h} \int_{K_{P_0}^*} \Delta u \cdot \psi_h \mathrm{d}x \tag{4.1.5a}$$

$$= -\sum_{P_0 \in \overline{\Omega}_h} \psi_h(P_0) \int_{\partial K_{P_0}^*} \frac{\partial u}{\partial n} \mathrm{d}s = \sum_{K \in T_h} I_K(u, \psi_h),$$

$$(-\Delta v, \phi_h) = -\sum_{P_0 \in \overline{\Omega}_h} \int_{K_{P_0}^*} \Delta v \cdot \phi_h \mathrm{d}x \tag{4.1.5b}$$

$$= -\sum_{P_0 \in \overline{\Omega}_h} \phi_h(P_0) \int_{\partial K_{P_0}^*} \frac{\partial v}{\partial n} \mathrm{d}s = \sum_{K \in T_h} I_K(v, \phi_h),$$

where $K = \Delta P_i P_j P_k$ (cf. Fig. 4.1.1), and

$$I_K(u, \psi_h) = \sum_{l=i,j,k} [\psi_h(P_{l+2} - \psi_h(P_{l+1})] \int_{\overline{M_l Q}} \frac{\partial u}{\partial n_l} \mathrm{d}s, \tag{4.1.6a}$$

$$I_K(v, \phi_h) = \sum_{l=i,j,k} [\phi_h(P_{l+2} - \phi_h(P_{l+1})] \int_{\overline{M_l Q}} \frac{\partial v}{\partial n_l} \mathrm{d}s, \tag{4.1.6b}$$

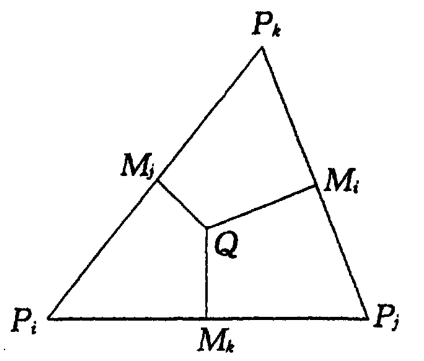



**Fig. 4.1.1**

where $n_l$ is the unit outer normal direction of $K^*_{P_{l+1}} \cap K$ along $\overline{M_l Q}$, and we make the convention that $i+1=j, j+1=k, k+1=i$.

(4.1.5a) and (4.1.5b) can be respectively rewritten as

$$a(u,\psi_h) = \int_\Omega \nabla u \cdot \nabla \psi_h \mathrm{d}x,$$

$$a(v,\phi_h) = \int_\Omega \nabla v \cdot \nabla \phi_h \mathrm{d}x.$$

Here the right-hand sides are in the sense of generalized functions (cf. Chapter 3). Therefore, we have the following variational problem related to (4.1.1): Find $(u,v) \in H_0^1(\Omega) \times H^1(\Omega)$ such that

$$\begin{cases} a(u,\psi) = (v,\psi), \quad \forall \psi \in S \equiv \bigcup_h V_h, & (4.1.7a) \\ a(v,\phi) = (f,\phi), \quad \forall \phi \in S_0 \equiv \bigcup_h V_{0h}. & (4.1.7b) \end{cases}$$

We interpret (4.1.7) as a generalization of the variational form (4.1.1) in the sense of generalized functions. Note that the density of $S$ and $S_0$ in $L^2(\Omega)$ implies the equivalence of (4.1.7) and (4.1.1).

Based on the above variational form, we define a mixed generalized difference scheme approximating (4.1.1): Find $(u_h, v_h) \in U_{0h} \times U_h$ such that

$$\begin{cases} a(u_h, \psi_h) = (v_h, \psi_h), & \forall \psi_h \in V_h, \quad (4.1.8a) \\ a(v_h, \phi_h) = (f, \phi_h), & \forall \phi_h \in V_{0h}. \quad (4.1.8b) \end{cases}$$

Obviously (4.1.8) is a linear system with order $\dim U_{0h} + \dim U_h$.

Let $\Pi_h^*$ be the interpolation projector from $U_h$ to $V_h$:

$$\Pi_h^* w_h = \sum_{P_0 \in \overline{\Omega}_h} w_h(P_0)\chi_{P_0}, \; w_h \in U_h,$$

where $\chi_{P_0}$ is the characteristic function of the set $K_{P_0}^*$. Then (4.1.8) is equivalent to

$$\begin{cases} a(u_h, \Pi_h^*\psi_h) = (v_h, \Pi_h^*\psi_h), & \forall \psi_h \in U_h, \\ a(v_h, \Pi_h^*\phi_h) = (f, \Pi_h^*\phi_h), & \forall \phi_h \in U_{0h}. \end{cases}$$

**Lemma 4.1.1** *The bilinear form $a(\cdot, \Pi_h^*\cdot)$ is symmetric and positive definite:*

$$a(u_h, \Pi_h^* w_h) = a(w_h, \Pi_h^* u_h), \;\; \forall u_h, w_h \in U_h, \quad (4.1.9)$$

$$a(u_h, \Pi_h^* u_h) = |u_h|_1^2, \; \forall u_h \in U_h. \quad (4.1.10)$$

**Proof** Since $u_h$ and $w_h$ are piecewise linear functions, $\frac{\partial u_h}{\partial x}$ and $\frac{\partial u_h}{\partial y}$ are respectively constants in each element $K$. Thus $I_K(\cdot,\cdot)$ can be expressed as (cf. §3.2)

$$\begin{aligned} &I_K(u_h, \Pi_h^* w_h) \\ &= \sum_{P \in \dot{K}} w_h(P)\Big(\int_{\partial K_P^* \cap K} -\frac{\partial u_h}{\partial x_1}\mathrm{d}x_2 + \frac{\partial u_h}{\partial x_2}\mathrm{d}x_1\Big) \\ &= \Big[\frac{\partial u_h}{\partial x_1}(x_2(M_k) - x_2(M_j)) + \frac{\partial u_h}{\partial x_2}(x_1(M_j) - x_1(M_k))\Big] w_h(P_i) \\ &\quad + \Big[\frac{\partial u_h}{\partial x_1}(x_2(M_i) - x_2(M_k)) + \frac{\partial u_h}{\partial x_2}(x_1(M_k) - x_1(M_i))\Big] w_h(P_j) \\ &\quad + \Big[\frac{\partial u_h}{\partial x_1}(x_2(M_j) - x_2(M_i)) + \frac{\partial u_h}{\partial x_2}(x_1(M_i) - x_1(M_j))\Big] w_h(P_k) \\ &= \Big(\frac{\partial u_h}{\partial x_1}\frac{\partial w_h}{\partial x_1} + \frac{\partial u_h}{\partial x_2}\frac{\partial w_h}{\partial x_2}\Big) S_K, \end{aligned} \quad (4.1.11)$$

and

$$a(u_h, \Pi_h^* w_h) = \int_\Omega \nabla u_h \cdot \nabla w_h \mathrm{d}x. \tag{4.1.12}$$

So (4.1.9) and (4.1.10) hold. This completes the proof. □

**Lemma 4.1.2** *There hold the following statements:*

(i) $(u_h, \Pi_h^* \bar{u}_h) = (\bar{u}_h, \Pi_h^* u_h),\ \forall u_h, \bar{u}_h \in U_h.$ (4.1.13)

(ii) *Write* $|||u_h|||_0 = (u_h, \Pi_h^* u_h)^{1/2}$. *Then* $|||\cdot|||_0$, $\|\cdot\|_{0,h}$ *and* $\|\cdot\|_0$ *are all equivalent on* $U_h$*: There exist positive constant* $c_1$ *and* $c_2$ *independent of* $U_h$ *such that*

$$c_1 \|u_h\|_0 \le |||u_h|||_0 \le c_2 \|u_h\|_0, \quad \forall u_h \in U_h. \tag{4.1.14}$$

**Proof** First we have

$$(u_h, \Pi_h^* \bar{u}_h) = \sum_{K \in T_h} \sum_{l=i,j,k} \bar{u}_h(P_l) \int_{K_{P_l}^* \cap K} u_h \mathrm{d}x.$$

By the linearity of $u_h$ we have

$$\begin{aligned} \int_{K_{P_i}^* \cap K} u_h \mathrm{d}x = & \frac{1}{3}[u_h(P_i) + u_h(M_j) + u_h(Q)] \cdot \frac{S_K}{6} \\ & + \frac{1}{3}[u_h(P_i) + u_h(M_k) + u_h(Q)] \cdot \frac{S_K}{6} \\ = & \frac{S_K}{108}[22u_h(P_i) + 7u_h(P_j) + 7u_h(P_k)]. \end{aligned}$$

This gives

$$\begin{aligned} & (u_h, \Pi_h^* \bar{u}_h) \\ = & \sum_{K \in T_h} \frac{S_K}{108} [\bar{u}_h(P_i), \bar{u}_h(P_j), \bar{u}_h(P_k)] \begin{bmatrix} 22 & 7 & 7 \\ 7 & 22 & 7 \\ 7 & 7 & 22 \end{bmatrix} \begin{bmatrix} u_h(P_i) \\ u_h(P_j) \\ u_h(P_k) \end{bmatrix}. \end{aligned}$$

This together with Lemma 3.2.1 leads to the desired result and completes the proof. □

**Theorem 4.1.1** *The mixed generalized difference scheme (4.1.8) possesses a unique solution.*

**Proof** We only have to show that the homogeneous equation

$$\begin{cases} a(u_h, \Pi_h^* \psi_h) = (v_h, \Pi_h^* \psi_h), & \forall \psi_h \in U_h \qquad (4.1.15a) \\ a(v_h, \Pi_h^* \phi_h) = 0, & \forall \phi_h \in U_{0h} \qquad (4.1.15b) \end{cases}$$

admits solely the trivial solution. In fact, if we set $\psi_h = v_h$, $\phi_h = u_h$ in (4.1.15a) and (4.1.15b) respectively, take their subtraction, and use (4.1.9), then we have

$$(v_h, \Pi_h^* v_h) = 0.$$

This together with (4.1.14) implies $v_h = 0$. Thus (4.1.15a) becomes

$$a(u_h, \Pi_h^* \psi_h) = 0, \quad \forall \psi_h \in U_h.$$

So setting $\psi_h = u_h$ and using (4.1.10) yields $u_h = 0$. This completes the proof. □

### 4.1.2 Error estimates

Let $\Pi_h u$ be the interpolation projection of $u \in H_0^1(\Omega)$ into the linear element space $U_{0h}$, and $P_h v$ the elliptic projection of $v \in H^1(\Omega)$ into $U_h$ in the following sense:

$$a(P_h v, \Pi_h^* \psi_h) = a(v, \Pi_h^* \psi_h), \ \forall \psi_h \in U_h, \qquad (4.1.16a)$$

$$\int_\Omega P_h v \mathrm{d}x = \int_\Omega v \mathrm{d}x. \qquad (4.1.16b)$$

For these projections we have

$$|u - \Pi_h u|_m \le C h^{2-m} |u|_2, \ m = 0, 1, \qquad (4.1.17)$$

$$\|v - P_h v\|_1 \le C h |v|_2. \qquad (4.1.18)$$

Here (4.1.18) can be obtained as in §3.2.

**Lemma 4.1.3** *Suppose that $T_h$ is a quasi-uniform grid and that $u \in H_0^2(\Omega) \cap W^{3,\infty}(\Omega)$, then*

$$|a(u - \Pi_h u, \Pi_h^* \psi_h)| \le Ch^2 \|u\|_{3,\infty} |\psi_h|_1, \quad \forall \psi_h \in U_h. \tag{4.1.19}$$

**Proof** Note

$$\begin{aligned}
& a(u - \Pi_h u, \Pi_h^* \psi_h) \\
= & \sum_{K \in T_h} \sum_{l=i,j,k} [\psi_h(P_{l+2}) - \psi_h(P_{l+1})] \int_{\overline{M_l Q}} \frac{\partial (u - \Pi_h u)}{\partial \tau_l} \mathrm{d}s \\
= & \sum{}' [\psi_h(P_{l+2}) - \psi_h(P_{l+1})] \int_{\overline{M_l Q}} \frac{\partial (u - \Pi_h u)}{\partial \tau_l} \mathrm{d}s \\
& + \sum{}'' [\psi_h(P_{l+2}) - \psi_h(P_{l+1})] \int_{\overline{M_l Q}} \frac{\partial (u - \Pi_h u)}{\partial \tau_l} \mathrm{d}s,
\end{aligned}$$

where $\sum'$ and $\sum''$ denote the summations for $\overline{P_{l+1}P_{l+2}}$ not belonging to and belonging to the boundary $\partial\Omega$, respectively. In the former case we have (cf. Theorem 3.7.2)

$$\begin{aligned}
& \Big| \sum{}' [\psi_h(P_{l+2}) - \psi_h(P_{l+1})] \int_{\overline{M_l Q}} \frac{\partial (u - \Pi_h u)}{\partial \tau_l} \mathrm{d}s \Big| \\
\le & Ch^2 (\|u\|_3 + \|u\|_{2,\infty}) |\psi_h|_1.
\end{aligned}$$

In the latter case, $\overline{P_{l+1}P_{l+2}}$ belongs to the boundary $\partial\Omega$. Now we expand $\frac{\partial(u - \Pi_h u)}{\partial \tau_l}$ at $M_l$ and use the boundary condition to obtain

$$\Big| \frac{\partial (u - \Pi_h u)}{\partial \tau_l} \Big| \le Ch^2 \|u\|_{3,\infty}, \ x \in \overline{M_l Q}.$$

So we have

$$\Big| \sum{}'' [\psi_h(P_{l+2}) - \psi_h(P_{l+1})] \int_{\overline{M_l Q}} \frac{\partial (u - \Pi_h u)}{\partial \tau_l} \mathrm{d}s \Big| \le Ch^2 \|u\|_{3,\infty} |\psi_h|_1. \tag{4.1.22}$$

Finally, (4.1.19) follows from (4.1.20)-(4.1.22). □

**Theorem 4.1.2** *Let $T_h$ be a quasi-uniform grid, $(u,v) \in H_0^2(\Omega) \times H^2(\Omega)$ the solution to (4.1.3), and $(u_h, v_h) \in U_{0h} \times U_h$ the solution to the mixed generalized difference scheme (4.1.8). Then*

$$\|u - u_h\|_1 + \|v - v_h\|_0 \le Ch(\|u\|_{3,\infty} + |v|_2). \tag{4.1.23}$$

**Proof** It follows from (4.1.14,7,8,19,15) that

$$\begin{aligned}
&\|v_h - P_h v\|_0^2 \le Ca(v_h - P_h v, \Pi_h^*(v_h - P_h v)) \\
&= C[a(u_h - \Pi_h u, \Pi_h^*(v_h - P_h v)) + a(\Pi_h u - u, \Pi_h^*(v_h - P_h v)) \\
&\quad +(v - P_h v, \Pi_h^*(v_h - P_h v))] \\
&\le C[a(v_h - P_h v, \Pi_h^*(u_h - \Pi_h u)) + h\|u\|_{3,\infty}|v_h - P_h v|_1 \\
&\quad +h|v|_2\|v_h - P_h v\|_0.
\end{aligned}$$

It follows from (4.1.7b), (4.1.8b) and (4.1.16a) that

$$a(v_h - P_h v, \Pi_h^*(u_h - \Pi_h u)) = 0.$$

By the above two estimates and the inverse property of the finite element we have

$$\|v_h - P_h v\|_0 \le Ch(\|u\|_{3,\infty} + |v|_2).$$

This together with (4.1.18) implies

$$\|v - v_h\|_0 \le Ch(\|u\|_{3,\infty} + |v|_2). \tag{4.1.24}$$

It follows from (4.1.10,19,7a,8a,18,23) that

$$\begin{aligned}
&|\Pi_h u - u_h|_1^2 \\
&= a(\Pi_h u - u,, \Pi_h^*(\Pi_h u - u_h)) + a(u - u_h, \Pi_h^*(\Pi_h u - u_h)) \\
&= a(\Pi_h u - u,, \Pi_h^*(\Pi_h u - u_h)) + (v - v_h, \Pi_h^*(\Pi_h u - u_h)) \\
&\le Ch^2\|u\|_{3,\infty}|\Pi_h u - u_h|_1 + Ch(\|u\|_{3,\infty} + |v|_2)\|\Pi_h u - u_h\|_0.
\end{aligned}$$

Hence

$$|\Pi_h u - u_h|_1 \le Ch(\|u\|_{3,\infty} + |v|_2). \tag{4.1.25}$$

A combination of (4.1.24), (4.1.25) and (4.1.17) leads to (4.1.23). This completes the proof. □

## 4.2 Mixed Generalized Difference Methods Based on Hermann-Miyoshi Variational Principle

Again consider the biharmonic equation (4.1.1). Now let us introduce a new unknown function $v_{ij} = \frac{\partial^2 u}{\partial x_i \partial x_j}$ $(i, j = 1, 2)$ so as to rewrite (4.1.1) into the following system of second order equations:

$$\begin{cases} \dfrac{\partial^2 u}{\partial x_i \partial x_j} = v_{ij}, & (x_1, x_2) \in \Omega,\ i, j = 1, 2, \quad (4.2.1a) \\ \displaystyle\sum_{i,j=1}^{2} \dfrac{\partial^2 v_{ij}}{\partial x_i \partial x_j} = f, & (x_1, x_2) \in \Omega, \quad (4.2.1b) \\ u = \dfrac{\partial u}{\partial n} = 0, & (x_1, x_2) \in \partial\Omega. \quad (4.2.1c) \end{cases}$$

Write

$$U = H^1(\Omega),\ U_0 = H_0^1(\Omega),$$

$$\tilde{U} = \{v = (v_{ij}),\ 1 \le i, j \le 2 : v_{12} = v_{21}, v_{ij} \in U\}.$$

Use $\psi \in \tilde{U}$ and $\phi \in U_0$ to multiply (4.2.1a) and (4.2.1b) respectively, integrate them on $\Omega$ and employ the Green's formula and the boundary condition (4.2.1c), then we have the variational form corresponding to (4.2.1): Find $(u, v) \in U_0 \times \tilde{U}$ such that

$$a(u, \psi) = -\langle v, \psi \rangle,\ \forall \psi \in \tilde{U}, \tag{4.2.2a}$$

$$a(\phi, v) = -(f, \phi),\ \forall \phi \in U_0, \tag{4.2.2b}$$

where

$$a(u, v) = \sum_{i,j=1}^{2} \int_\Omega \frac{\partial u}{\partial x_i} \frac{\partial v_{ij}}{\partial x_j} \mathrm{d}x,\ \forall u \in U, v \in \tilde{U}, \tag{4.2.3}$$

$$\langle v, \psi \rangle = \sum_{i,j=1}^{2} \int_\Omega v_{ij} \psi_{ij} \mathrm{d}x,\ \forall v, \psi \in \tilde{U}, \tag{4.2.4}$$

$$(f, \phi) = \int_\Omega f \phi \mathrm{d}x,\ \forall f, \phi \in U. \tag{4.2.5}$$

We can also use the regularity of (4.1.1) to show the equivalence of (4.2.2) and (4.1.1) [cf. [B-30]: If $u$ is the solution to (4.1.1) and $v = (v_{ij})$ with $v_{ij} = \frac{\partial^2 u}{\partial x_i \partial x_j}$ $(i, j = 1, 2)$, then $(u, v)$ solves (4.2.2). On the other hand, if $(u, v)$ is the solution to (4.2.2), then $u$ is the solution to (4.1.1) and $v = (v_{ij}) = \left(\frac{\partial^2 u}{\partial x_i \partial x_j}\right)$ $(i, j = 1, 2)$.

### 4.2.1 Mixed generalized difference equations

Let $U_h, V_h, U_{0h}, V_{0h}$ be defined as in the last section and

$$\tilde{U}_h = \{u_h = (u_h^{ij}), 1 \leq i, j \leq 2 : u_h^{12} = u_h^{21}, u_h^{ij} \in U_h\},$$

$$\tilde{V}_h = \{v_h = (v_h^{ij}), 1 \leq i, j \leq 2 : v_h^{12} = v_h^{21}, v_h^{ij} \in V_h\}.$$

Based on the variational form (4.2.2), the mixed generalized difference scheme for (4.1.1) is defined as: Find $(u_h, v_h) \in U_{0h} \times \tilde{U}_h$ such that

$$a(u_h, \psi_h) = -\langle v_h, \psi_h \rangle, \ \forall \psi_h \in \tilde{V}_h, \tag{4.2.6a}$$

$$a(\phi_h, v_h) = -(f, \phi_h), \ \forall \phi_h \in V_{0h}, \tag{4.2.6b}$$

where $a(\cdot, \cdot)$ is interpreted in the sense of generalized functions (cf. Chapter 3), that is,

$$\begin{aligned} a(u_h, \psi_h) &= \sum_{i,j=1}^{2} \int_\Omega \frac{\partial u_h}{\partial x_i} \frac{\partial \psi_h^{ij}}{\partial x_j} \mathrm{d}x \\ = \sum_{K_P^* \in T_h^*} \int_{\partial K_P^*} \Big( &-\frac{\partial u_h}{\partial x_1} \psi_h^{11}(P) \mathrm{d}x_2 + \frac{\partial u_h}{\partial x_1} \psi_h^{12}(P) \mathrm{d}x_1 \\ &-\frac{\partial u_h}{\partial x_2} \psi_h^{21}(P) \mathrm{d}x_2 + \frac{\partial u_h}{\partial x_2} \psi_h^{22}(P) \mathrm{d}x_1 \Big), \\ &u_h \in U_{0h}, \psi_h \in \tilde{V}_h, \end{aligned} \tag{4.2.7}$$

$$
\begin{aligned}
a(\phi_h, v_h) &= \sum_{i,j=1}^{2} \int_\Omega \frac{\partial \phi_h}{\partial x_i}\frac{\partial v_h^{ij}}{\partial x_j}\mathrm{d}x \\
= \sum_{K_P^* \in T_h^*} \phi_h(P) \int_{\partial K_P^*} &-\Big(\frac{\partial v_h^{11}}{\partial x_1} + \frac{\partial v_h^{12}}{\partial x_2}\Big)\mathrm{d}x_2 \\
&+\Big(\frac{\partial v_h^{21}}{\partial x_1} + \frac{\partial v_h^{22}}{\partial x_2}\Big)\mathrm{d}x_1, \\
&\phi_h \in V_{0h}, v_h \in \tilde{U}_h.
\end{aligned} \tag{4.2.8}
$$

Scheme (4.2.6) can also be deduced as in §4.1.

If we take $\psi_h$ as the basis function of $\tilde{V}_h$, i.e., for each $P \in \overline{\Omega}_h$, we take $\psi_{ij}$ $(1 \le i, j \le 2)$ as the characteristic function of $K_P^*$, then (4.2.6a) becomes

$$\int_{\partial K_P^*} \frac{\partial u_h}{\partial x_1}\mathrm{d}x_2 = \int_{K_P^*} v_h^{11}\mathrm{d}x, \tag{4.2.9a}$$

$$\int_{\partial K_P^*} \frac{\partial u_h}{\partial x_1}\mathrm{d}x_1 = -\int_{K_P^*} v_h^{12}\mathrm{d}x, \tag{4.2.9b}$$

$$\int_{\partial K_P^*} \frac{\partial u_h}{\partial x_2}\mathrm{d}x_2 = \int_{K_P^*} v_h^{21}\mathrm{d}x, \tag{4.2.9c}$$

$$\int_{\partial K_P^*} \frac{\partial u_h}{\partial x_2}\mathrm{d}x_1 = -\int_{K_P^*} v_h^{22}\mathrm{d}x. \tag{4.2.9d}$$

If $\phi_h$ is taken as the basis function of $V_{0h}$, namely, the characteristic function of $K_P^*$ for each $P \in \dot{\Omega}_h$, then (4.2.6b) becomes

$$\int_{\partial K_P^*} -\Big(\frac{\partial v_h^{11}}{\partial x_1} + \frac{\partial v_h^{12}}{\partial x_2}\Big)\mathrm{d}x_2 + \Big(\frac{\partial v_h^{21}}{\partial x_1} + \frac{\partial v_h^{22}}{\partial x_2}\Big)\mathrm{d}x_1 = -\int_{K_P^*} f\mathrm{d}x. \tag{4.2.10}$$

Therefore, Scheme (4.2.6) becomes a system of linear algebraic equations (4.2.9) and (4.2.10) with order $(\dim U_{0h} + 3\dim U_h)$.

### 4.2.2 Numerical experiments

Consider the numerical solution to the biharmonic boundary value problem:

$$\begin{cases} \Delta^2 u(x,y) = f(x,y), & (x,y) \in \Omega = (0,1) \times (0,1), \quad (4.2.11a) \\ u = \dfrac{\partial u}{\partial n} = 0, & (x,y) \in \partial\Omega, \quad (4.2.11b) \end{cases}$$

where

$$\begin{aligned} f(x,y) = \ & 8[3x^2(1-x)^2 + 3y^2(1-y)^2 \\ & +(6x^2-6x+1)(6y^2-6y+1)]. \end{aligned}$$

The true solution to this problem is

$$u(x,y) = x^2y^2(1-x)^2(1-y)^2.$$

### Thirteen point finite difference method

Place a square grid of $\Omega$ with a mesh step size $h = 1/n$, the nodes $P_{ij} = (x_i, y_j) = (i/n, j/n)$, $i, j = 0, 1, 2, \cdots, n$, and the mesh function of the nodes $u_{ij} = u_h(x_i, y_j)$.

The well-known thirteen point difference scheme is (cf. [c-6]):

$$\begin{aligned} & h^{-4}[20u_{ij} - 8(u_{i,j-1} + u_{i,j+1} + u_{i-1,j} + u_{i+1,j}) \\ & \quad +2(u_{i-1,j-1} + u_{i-1,j+1} + u_{i+1,j-1} + u_{i+1,j+1}) \\ & \quad +(u_{i,j-2} + u_{i,j+2} + u_{i-2,j} + u_{i+2,j})] = f(x_i, y_j), \\ & \qquad\qquad 1 \le i, j \le n-1. \end{aligned} \tag{4.2.12}$$

Here when $i = 1, n-1$ or $j = 1, n-1$, we need to define the values on the virtual nodes $P_{ij} = (x_i, y_j)$ (for either $i = -1, n+1$ or $j = -1, n+1$) outside of $\Omega$. For instance, for $P_{-1,j} = (x_{-1}, y_j)$ $(0 \le j \le n)$ on the left side of $\Omega$, we may employ

$$\frac{u(x_{-1}, y_j) - u(x_1, y_j)}{2h} \doteq \frac{\partial u}{\partial n}(x_0, y_j), \ 0 \le j \le n$$

to define

$$u_{-1,j} = u_{1,j} + 2h\frac{\partial u}{\partial n}(x_0, y_j) = u_{1,j}, \ 0 \le j \le n.$$

**Mixed generalized difference methods**

On the square mesh with mesh step size $h = 1/n$, we use the diagonals of every small square, parallel to the line $y = x$, to obtain a right angle triangulation together with a corresponding circumcenter dual grid.

Let $u_{ij} = u_h(x_i, y_j)$, $v_{ij}^{kl} = v_h^{kl}(x_i, y_j)$ $(i, j = 0, 1, \cdots, n; k, l = 1, 2)$ denote the mesh function defined on the nodes $P_{ij} = (x_i, y_j)$ $(i, j = 0, 1, \cdot, n)$. The generalized difference equation can be easily deduced by a computation of (4.2.9) and (4.2.10). The equation related to the interior point $P_{ij}$ reads:

$$\begin{aligned} &\frac{h^2}{24}(14v_{ij}^{11} + 2v_{i-1,j}^{11} + 2v_{i+1,j}^{11} + 2v_{i,j-1}^{11} \\ &+2v_{i,j+1}^{11} + v_{i-1,j-1}^{11} + v_{i+1,j+1}^{11}) \\ = \ &u_{i-1,j} - 2u_{ij} + u_{i+1,j}, \end{aligned} \tag{4.2.9a$'$}$$

$$\begin{aligned} &\frac{h^2}{24}(14v_{ij}^{12} + 2v_{i-1,j}^{12} + 2v_{i+1,j}^{12} + 2v_{i,j-1}^{12} \\ &+2v_{i,j+1}^{12} + v_{i-1,j-1}^{12} + v_{i+1,j+1}^{12}) \\ = \ &\frac{1}{2}(2u_{i,j} + u_{i-1,j-1} + u_{i+1,j+1} \\ &-u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1}), \end{aligned} \tag{4.2.9b$'$}$$

$$\begin{aligned} &\frac{h^2}{24}(14v_{ij}^{21} + 2v_{i-1,j}^{21} + 2v_{i+1,j}^{21} + 2v_{i,j-1}^{21} \\ &+2v_{i,j+1}^{21} + v_{i-1,j-1}^{21} + v_{i+1,j+1}^{21}) \\ = \ &\frac{1}{2}(2u_{i,j} + u_{i-1,j-1} + u_{i+1,j+1} \\ &-u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1}), \end{aligned} \tag{4.2.9c$'$}$$

$$\begin{aligned} &\frac{h^2}{24}(14v_{ij}^{22} + 2v_{i-1,j}^{22} + 2v_{i+1,j}^{22} + 2v_{i,j-1}^{22} \\ &+2v_{i,j+1}^{22} + v_{i-1,j-1}^{22} + v_{i+1,j+1}^{22}) \\ = \ &u_{i,j-1} - 2u_{ij} + u_{i,j+1}, \end{aligned} \tag{4.2.9d$'$}$$

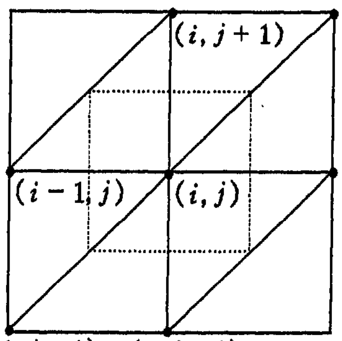



**Fig. 4.2.1**

$$\begin{aligned}
&2v_{ij}^{11} - v_{i-1,j}^{11} - v_{i+1,j}^{11} + 2v_{ij}^{22} - v_{i,j-1}^{22} - v_{i,j+1}^{22} - 2v_{ij}^{12} \\
&-v_{i-1,j-1}^{12} - v_{i+1,j+1}^{12} + v_{i-1,j}^{12} + v_{i+1,j}^{12} + v_{i,j-1}^{12} + v_{i,j+1}^{12} \\
= &-\int_{K_{P_{ij}}^*} f \mathrm{d}x \doteq -f(x_i, y_j)h^2.
\end{aligned} \qquad (4.2.10)'$$

(4.2.9b)′ and (4.2.9c)′ are identical since $v_{ij}^{12} = v_{ij}^{21}$ $(0 \le i, j \le n)$. So we should take only one of them. To save the space, we do not present here the difference equations with respect to the boundary nodes. We remark that the discrete equations obviously result in a seven point scheme (cf. Fig. 4.2.1).

## Numerical results

We use respectively the thirteen point finite difference scheme (FDM) and the mixed generalized difference scheme (GDM) mentioned above to approximate (4.2.11). For $n = 10$, the numerical results and the true solution (TS) are listed in Table 4.2.1 for comparison. This numerical experiment is done in [A-31].

The numerical experiments indicate that the mixed generalized difference method needs less computation time than the corresponding mixed finite element method, while it enjoys a better accuracy as well as a more flexible decomposition than the thirteen point finite difference method.

**Table 4.2.1. Numerical Results**

| $(x, y)$ | FDM $u_h$ | GDM $u_h$ | TS $u$ |
|---|---|---|---|
| $(x_1, y_1)$ | 0.00009030 | 0.00006841 | 0.00006561 |
| $(x_2, y_1)$ | 0.00026062 | 0.00021000 | 0.00020736 |
| $(x_2, y_2)$ | 0.00075917 | 0.00065948 | 0.00065536 |
| $(x_3, y_1)$ | 0.00043654 | 0.00036458 | 0.00035721 |
| $(x_3, y_2)$ | 0.00127511 | 0.00114762 | 0.00112896 |
| $(x_3, y_3)$ | 0.00214373 | 0.00198859 | 0.00194481 |
| $(x_4, y_1)$ | 0.00056410 | 0.00048280 | 0.00046656 |
| $(x_4, y_2)$ | 0.00164913 | 0.00151576 | 0.00147456 |
| $(x_4, y_3)$ | 0.00277345 | 0.00261473 | 0.00254016 |
| $(x_4, y_4)$ | 0.00358862 | 0.00342425 | 0.00331776 |
| $(x_5, y_1)$ | 0.00061032 | 0.00053178 | 0.00050625 |
| $(x_5, y_2)$ | 0.00178460 | 0.00166290 | 0.00160000 |
| $(x_5, y_3)$ | 0.00300153 | 0.00285651 | 0.00275625 |
| $(x_5, y_4)$ | 0.00388386 | 0.00372713 | 0.00360000 |
| $(x_5, y_5)$ | 0.00420342 | 0.00404309 | 0.00390625 |

## 4.3 Nonconforming Generalized Difference Method Based on Zienkiewicz Elements

### 4.3.1 Variational principle

Consider the Dirichlet problem of the biharmonic operator:

$$\begin{cases} \Delta^2 u = f, & (x_1, x_2) \in \Omega, & (4.3.1a) \\ u = \dfrac{\partial u}{\partial n} = 0, & (x_1, x_2) \in \partial\Omega, & (4.3.1b) \end{cases}$$

where $\Omega$ is a bounded plane region with a Lipschitz continuous boundary $\partial\Omega$, and $\frac{\partial}{\partial n}$ the derivative operator along the outer normal direction, $f \in L^2(\Omega)$.

As mentioned earlier, the generalized difference method is a kind of difference method based on a variational principle over an irregular network. A basic idea of it is to choose the test function space to be as flexible and simple as possible (usually the piecewise constant or piecewise linear function spaces) so as to reduce the computational effort, while keep the approximation order of the trial function space, ending up with a scheme enjoying both the simplicity of the finite difference method and the accuracy of the finite element method. It is obvious that the usual variational form (where the set of freedoms of the test functions involves second order derivatives) of (4.3.1) requires the test function space to contain piecewise quadratic polynomials, increasing greatly the complexity of the computation. In order to construct a simpler difference scheme we seek another form of the variational principle.

Let $\sigma$ be a decomposition of $\Omega$, dividing $\overline{\Omega}$ into a sum of finite number of closed subsets $K$ which possess Lipschitz continuous boundaries, have nonempty interiors and share no common inner point:

$$\overline{\Omega} = \bigcup_{K\in\sigma} K,$$

$$\mathrm{int}K_1 \cap \mathrm{int}K_2 = \emptyset, \ \forall K_1, K_2 \in \sigma, \ K_1 \neq K_2.$$

Here int$K$ denotes the interior of K. The family of functions

$$S_\sigma(\Omega) = \{v \in L^2(\Omega) : \ v|_{\mathrm{int}K} \in \mathcal{P}_1, \ \forall K \in \sigma\}$$

is called the family of piecewise linear functions related to the decomposition $\sigma$, and the family of functions

$$S(\Omega) = \bigcup_{\sigma} S_\sigma(\Omega)$$

the family of piecewise linear functions on $\Omega$.

Take any $v \in S_\sigma(\Omega)$ to multiply (4.3.1a), integrate it on $\Omega$, and use the following Green's formulas on each $K \in \sigma$ (cf. (1.2.10))

$$\int_K \Delta^2 u \cdot v\mathrm{d}x = \int_K \Delta u \Delta v \mathrm{d}x + \int_{\partial K} \Big(\frac{\partial \Delta u}{\partial n} v - \Delta u \frac{\partial v}{\partial n}\Big)\mathrm{d}s, \quad (4.3.2)$$

$$\begin{aligned}
&\int_K\Big(\frac{\partial^2 u}{\partial x_1^2}\frac{\partial^2 v}{\partial x_2^2}+\frac{\partial^2 u}{\partial x_2^2}\frac{\partial^2 v}{\partial x_1^2}-2\frac{\partial^2 u}{\partial x_1\partial x_2}\frac{\partial^2 v}{\partial x_1\partial x_2}\Big)\mathrm{d}x \\
= &\int_{\partial K}\Big(\frac{\partial^2 u}{\partial \tau^2}\frac{\partial v}{\partial n}-\frac{\partial^2 u}{\partial n\partial\tau}\frac{\partial v}{\partial\tau}\Big)\mathrm{d}s,
\end{aligned} \tag{4.3.3}$$

then we have the following variational form of (4.3.1): Find $u \in H_0^2(\Omega)\cap H^4(\Omega)$ such that

$$a_\sigma(u,v)=(f,v),\ \forall v\in S_\sigma(\Omega), \sigma\in\{\sigma\}, \tag{4.3.4}$$

where

$$\begin{aligned}
a_\sigma(u,v) &= \int_\Omega v\Delta^2 u\mathrm{d}x \\
&= \sum_{K\in\sigma}\int_{\partial K}\Big[\Big(\frac{\partial\Delta u}{\partial n}v-\Delta u\frac{\partial v}{\partial n}\Big)+\mu\Big(\frac{\partial^2 u}{\partial\tau^2}\frac{\partial v}{\partial n}-\frac{\partial^2 u}{\partial n\partial\tau}\frac{\partial v}{\partial \tau}\Big)\Big]\mathrm{d}s,
\end{aligned} \tag{4.3.5}$$

where $\frac{\partial}{\partial n}$ and $\frac{\partial}{\partial\tau}$ denote respectively the derivatives along the outer normal and the tangent directions, the value of $v$ on $\partial K$ is interpreted as the continuous extension of the values of $v|_{\mathrm{int}K}$ to $\partial K$, and $\mu$ is a constant. The density of $S(\Omega)$ in $L^2(\Omega)$ implies that $u$ solves (4.3.1) as long as $u$ satisfies (4.3.4). This observation gives the following variational principle.

**Theorem 4.3.1** *If $u\in H_0^2(\Omega)\cap H^4(\Omega)$ is a solution to (4.3.1), then $u$ solves (4.3.4), and vice versa.*

### 4.3.2 Generalized difference schemes based on Zienkiewicz elements

Let $\Omega$ be a polygon region, $T_h$ a quasi-uniform triangulation of $\Omega$ ($h$ stands for the largest element diameter), and $T_h^*$ a dual grid by connecting the circumcenters of adjacent elements.

Furthermore, for simplicity we assume $T_h$ divide $\overline{\Omega}$ into a sum of finite number of right triangles, and the two right sides of each triangle are parallel to the coordinate axes respectively. Now the dual grid $T_h$ can be regarded as a result of using the perpendicular bisectors parallel to the right sides of the right triangles to divide

$\overline{\Omega}$. Each vertex $P$ of the triangle element (called a node of $T_h$) is surrounded by a small polygon (in particular, a small rectangle for inner nodes) called a dual element and denoted by $K_P^*$. (cf. Fig. 4.3.1, where the shaded part is $K_P^*$).

Choose the trial function space as the finite element space with respect to $T_h$, the Zienkiewicz triangulation (cf. [p.68, B-17]), of which the function $u_h$ satisfies, at each boundary node $P_0$, $u_h(P_0) = \frac{\partial u_h(P_0)}{\partial x_1} = \frac{\partial u_h(P_0)}{\partial x_2} = 0$.

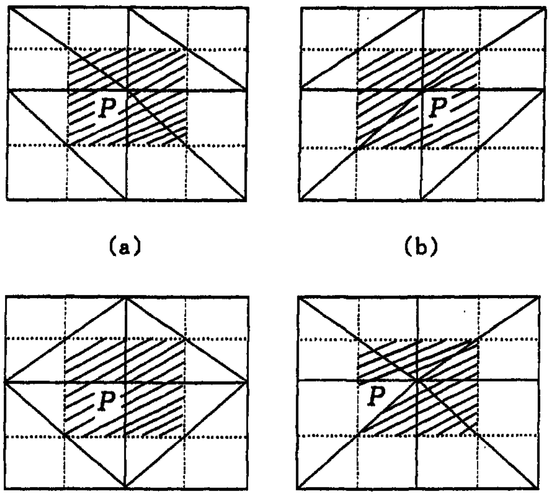



Fig. 4.3.1

The test function space $V_h$ is chosen as the piecewise linear function space corresponding to $T_h^*$. A function $v_h \in V_h$ satisfies on the boundary nodes $v_h(P_0) = \frac{\partial v_h(P_0)}{\partial x_1} = \frac{\partial v_h(P_0)}{\partial x_2} = 0$. The basis functions with respect to an inner node $P_0$ are

$$\psi_{P_0}^{(0)}(P) = \begin{cases} 1, & P \in K_{P_0}^*, \\ 0, & P \notin K_{P_0}^*, \end{cases}$$

$$\psi_{P_0}^{(1)}(P) = \begin{cases} x_1 - x_1(P_0), & P \in K_{P_0}^*, \\ 0, & P \notin K_{P_0}^*, \end{cases}$$

$$\psi_{P_0}^{(2)}(P) = \begin{cases} x_2 - x_2(P_0), & P \in K_{P_0}^*, \\ 0, & P \notin K_{P_0}^*. \end{cases}$$

The generalized difference scheme approximating (4.3.1) then becomes: Find $u_h \in U_h$ such that

$$a_h(u_h, v_h) = (f, v_h), \ \forall v_h \in V_h, \tag{4.3.6}$$

where

$$a_h(u_h, v_h) = \sum_{K \in T_h} I_K(u_h, v_h), \tag{4.3.7}$$

$$\begin{aligned} I_K(u_h, v_h) = \sum_{P \in \dot{K}} \int_{\partial K_P^* \cap K} \Big( & \frac{\partial \Delta u_h}{\partial n} v_h - \Delta u_h \frac{\partial v_h}{\partial n} \\ & + \frac{\partial^2 u_h}{\partial \tau^2} \frac{\partial v_h}{\partial n} - \frac{\partial^2 u_h}{\partial n \partial \tau} \frac{\partial v_h}{\partial \tau} \Big) \mathrm{d}s, \end{aligned} \tag{4.3.8}$$

where $\dot{K}$ denotes the set of vertexes of the element $K$.

The trial functions are chosen as Zienkiewicz elements (cf. [p.68, B-17]), so $U_h \not\subset H^2(\Omega)$ and we only have $U_h \in H^1(\Omega)$. The test function $V_h$ is in $L^2(\Omega)$ but not $H^1(\Omega)$. Therefore Scheme (4.3.6) is nonconforming.

The computations of $I_K(u_h, \psi_{P_0}^{(l)})$ $(l = 0, 1, 2)$ give the element matrices, and their summation gives the overall matrix, i.e., the coefficient matrix of the linear algebraic system of the approximation problem. Here the computation of the element matrices is simpler than the corresponding nonconforming finite element methods, since the computation here merely involves some line integrals, of which the integral paths are parallel to the coordinate axes; the basis functions $\psi_{P_0}^{(l)}$'s are extremely simple; many terms in $I_K$ are zero; and the nonzero terms are easy to compute.

We do not require the triangulation to satisfy the condition "the three sides of the triangle are parallel to three given directions." Various kinds of grids as illustrated in Fig. 4.3.1 are feasible. The approximation equation related to an inner node $P_0$ leads to seven point

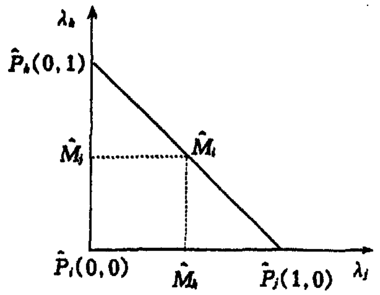



**Fig. 4.3.2**

generalized difference schemes in the cases of (a) and (b) of Fig. 4.3.1; to a five point scheme in case (c); and to a nine point scheme in case (d).

### 4.3.3 Error analyses

Take any triangular element $K \in T_h$ with vertexes $P_l$ $(l = i, j, k)$. Let $P_i$ denote the right vertex, $M_l$ the midpoint of the side opposite to $P_l$, $S_K$ the area of $K$, and $\lambda_K = |\overline{P_iP_j}|^2/|\overline{P_iP_k}|^2$. Perform an affine transformation

$$\lambda_i = \frac{1}{2S_K}\begin{vmatrix} 1 & x_1 & x_2 \\ 1 & x_1(P_j) & x_2(P_j) \\ 1 & x_1(P_k) & x_2(P_k) \end{vmatrix},$$

$$\lambda_j = \frac{1}{2S_K}\begin{vmatrix} 1 & x_1 & x_2 \\ 1 & x_1(P_k) & x_2(P_k) \\ 1 & x_1(P_i) & x_2(P_i) \end{vmatrix},$$

$$\lambda_k = \frac{1}{2S_K}\begin{vmatrix} 1 & x_1 & x_2 \\ 1 & x_1(P_i) & x_2(P_i) \\ 1 & x_1(P_j) & x_2(P_j) \end{vmatrix}.$$

Then $K$ is mapped onto a reference element $\hat{K}$ on $(\lambda_j, \lambda_k)$ plane (Fig. 4.3.2), and accordingly, $P_l$ and $M_l$ are mapped into $\hat{P}_l$ and $\hat{M}_l$ $(l = i, j, k)$, respectively. For any $u_h \in U_h$, we have on $K$ that

$$u_h = \sum_{l=i,j,k} (-2\lambda_l^3 + 3\lambda_l^2 + 2\lambda_i\lambda_j\lambda_k)u_h(P_l) + \sum_{l=j,k} \Big[(\lambda_l^3 - \lambda_l^2 - \lambda_i\lambda_j\lambda_k)\frac{\partial u_h(P_l)}{\partial \lambda_l} + \sum_{\substack{m=i,j,k \\ m\neq l}} (\lambda_m^2\lambda_l + \frac{1}{2}\lambda_i\lambda_j\lambda_k)\frac{\partial u_h(P_m)}{\partial \lambda_l}\Big]. \tag{4.3.9}$$

Define the interpolation operator $\Pi_h^*$ from $U_h$ to $V_h$ as follows

$$\Pi_h^* w_h = \sum_{P_0 \in \dot{\Omega}_h} \Big[w_h(P_0)\psi_{P_0}^{(0)} + \frac{\partial w_h(P_0)}{\partial x_1}\psi_{P_0}^{(1)} + \frac{\partial w_h(P_0)}{\partial x_2}\psi_{P_0}^{(2)}\Big], \quad \forall w_h \in U_h, \tag{4.3.10}$$

where $\dot{\Omega}_h$ denotes the set of all the inner nodes of $T_h$.

A direct computation leads to

$$I_K(u_h, \Pi_h^* w_h) = \frac{1}{16S_K}\delta_K(w_h)^T A_K \delta_K(u_h), \tag{4.3.11}$$

where

$$\begin{aligned}\delta_K(v) = \Big[&\frac{\partial v(P_i)}{\partial \lambda_j} + v(P_i) - v(P_j), \frac{\partial v(P_j)}{\partial \lambda_j} + v(P_i) - v(P_j), \\ &\frac{\partial v(P_k)}{\partial \lambda_j} + v(P_i) - v(P_j), \frac{\partial v(P_i)}{\partial \lambda_k} + v(P_i) - v(P_k), \\ &\frac{\partial v(P_j)}{\partial \lambda_k} + v(P_i) - v(P_k), \frac{\partial v(P_k)}{\partial \lambda_k} + v(P_i) - v(P_k)\Big]^T,\end{aligned}$$

$$A_K^T = \begin{bmatrix} 6+13\lambda & 2+11\lambda & -4 & 6+6\lambda-\frac{1}{\lambda} & -6\lambda & 6+\frac{1}{\lambda} \\ 4+6\lambda & 4+18\lambda & 0 & 4+6\lambda-\frac{2}{\lambda} & -6\lambda & 4+\frac{2}{\lambda} \\ -2+\lambda & 2-\lambda & 4 & -2-\frac{1}{\lambda} & 0 & -2+\frac{1}{\lambda} \\ 6-\lambda+\frac{6}{\lambda} & 6+\lambda & -\frac{6}{\lambda} & 6+\frac{13}{\lambda} & -4 & 2+\frac{11}{\lambda} \\ -2-\lambda & -2+\lambda & 0 & -2+\frac{1}{\lambda} & 4 & 2-\frac{1}{\lambda} \\ 4-2\lambda+\frac{6}{\lambda} & 4+2\lambda & -\frac{6}{\lambda} & 4+\frac{6}{\lambda} & 0 & 4+\frac{18}{\lambda} \end{bmatrix},$$

where we have written $\lambda = \lambda_K$ for short.

**Theorem 4.3.2** *Define*

$$\|u_h\|_h = \Big( \sum_{K\in T_h} \frac{1}{S_K}\delta_K(u_h)^T\delta_K(u_h)\Big)^{1/2}, \quad \forall u_h \in U_h. \qquad (4.3.12)$$

*Suppose the triangulation $T_h$ is quasi-uniform: There exists a constant $\lambda_0 > 0$ such that $\lambda_0 \le \lambda_K \le \lambda_0^{-1}$. Then $\|\cdot\|_h$ is equivalent to the following $|\cdot|_{2,h}$ norm*

$$|u_h|_{2,h} = \Big( \sum_{K\in T_h} |u_h|_{2,K}^2\Big)^{1/2}, \; \forall u_h \in U_h, \qquad (4.3.13)$$

*which means the existence of positive constants $c_1$ and $c_2$ independent of $U_h$ such that*

$$c_1\|u\|_h \le |u_h|_{2,h} \le c_2\|u_h\|_h, \quad \forall u_h \in U_h. \qquad (4.3.14)$$

**Proof** We only have to show the existence of constants $c_1'$, $c_2' > 0$ independent of $U_h$ and $K$ such that

$$\frac{c_1'}{S_K}\delta_K(u_h)^T\delta_K(u_h) \le |u_h|_{2,K}^2 \le \frac{c_2'}{S_K}\delta_K(u_h)^T\delta_K(u_h), \qquad (4.3.15)$$

$$\forall u_h \in U_h, K \in T_h.$$

Express $\frac{\partial^2 u_h}{\partial\lambda_j^2}$, $\frac{\partial^2 u_h}{\partial\lambda_j\lambda_k}$, and $\frac{\partial^2 u_h}{\partial\lambda_k^2}$ as multiplications of vectors and matrices, e.g.,

$$\frac{\partial^2 u_h}{\partial\lambda_j^2} = (\lambda_j, 1, \lambda_k)\begin{bmatrix} 6 & 6 & 0 & 0 & 0 & 0 \\ -4 & -2 & 0 & 0 & 0 & 0 \\ 3 & 2 & -1 & 1 & 1 & 2 \end{bmatrix}\delta_K(u_h).$$

A simple calculation gives

$$\begin{aligned} |u_h|_{2,K}^2 &= \int_{\hat K}\Big[\Big(\frac{1}{|\overline{P_iP_j}|^2}\frac{\partial^2 u_h}{\partial\lambda_j^2}\Big)^2 + \Big(\frac{1}{|\overline{P_iP_j}|\cdot|\overline{P_iP_k}|}\frac{\partial^2 u_h}{\partial\lambda_j\lambda_k}\Big)^2 \\ &\quad + \Big(\frac{1}{|\overline{P_iP_j}|^2}\frac{\partial^2 u_h}{\partial\lambda_k^2}\Big)^2\Big]2S_K\mathrm{d}\lambda_j\mathrm{d}\lambda_k \\ &= \frac{1}{S_K}\delta_K(u_h)^T D_K\delta_K(u_h), \end{aligned}$$

where

$$D_K = \frac{1}{24} G^T \mathrm{diag}(\lambda_K D_0, D_0, \lambda_K^{-1} D_0) G,$$

$$G^T = \begin{bmatrix} 6 & -4 & 3 & 3 & -3/2 & 1 & 1 & 0 & 0 \\ 6 & -2 & 2 & 2 & 1 & 2 & 2 & 0 & 0 \\ 0 & 0 & -1 & -1 & 1/2 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & -3/2 & 3 & 3 & -4 & 6 \\ 0 & 0 & 1 & 1 & 1/2 & -1 & -1 & 0 & 0 \\ 0 & 0 & 2 & 2 & -1 & 2 & 2 & -2 & 6 \end{bmatrix},$$

$$D_0 = \begin{bmatrix} 2 & 4 & 1 \\ 4 & 12 & 4 \\ 1 & 4 & 2 \end{bmatrix}.$$

It is easy to verify that $D_0$ is a positive definite matrix, and that the column vectors of $G$ are linearly independent. Now (4.3.15) holds since $\lambda_0 \le \lambda_K \le \lambda_0^{-1}$. This completes the proof. □

**Theorem 4.3.3** *Suppose the triangulation $T_h$ satisfies $\frac{2}{3} \le \lambda_K \le \frac{3}{2}$ $(K \in T_h)$, then the bilinear form $a_h(\cdot, \Pi_h^* \cdot)$ is uniformly positive definite, i.e., there is a constant $\alpha > 0$ independent of the subspace $U_h$ such that*

$$a_h(u_h, \Pi_h^* u_h) \ge \alpha |u_h|_{2,h}^2, \ \forall u_h \in U_h. \tag{4.3.16}$$

**Proof** Replace $w_h$ in (4.3.11) by $u_h$ and write it into a symmetric form:

$$I_K(u_h, \Pi_h^* u_h) = \frac{1}{32 S_K} \delta_K(u_h)^T B_K \delta(u_h), \tag{4.3.17}$$

where $B_K = A_K + A_K^T$ is symmetric. We might as well assume $\frac{2}{3} \le \lambda_K^{-1} \le 1 \le \lambda_K \le \frac{3}{2}$. Let us set $\lambda_K = 1$ in $B_K$ to obtain a matrix $\hat{B}_K$, written in a block form:

$$\hat{B}_K = B_K|_{\lambda_K=1} = \begin{bmatrix} \hat{B}_{11} & \hat{B}_{12} \\ \hat{B}_{21} & \hat{B}_{22} \end{bmatrix},$$

where each subblock is a $3 \times 3$ matrix. Write

$$G_1 = \begin{bmatrix} I & -\hat{B}_{11}^{-1}\hat{B}_{12} \\ 0 & I \end{bmatrix}, \ G_2 = \mathrm{diag}\Big(\frac{2}{3}, \frac{1}{2}, 1, 1, 1, \frac{7}{8}\Big),$$

and define the symmetric matrix

$$\tilde{B}_K = G_2^T G_1^T B_K G_1 G_2 = [\tilde{b}_{ij}]_{6\times 6},$$

where

$$\begin{aligned}
&\tilde{b}_{11} = 11.56\lambda + 5.33, && \tilde{b}_{12} = 5.67\lambda + 2, \\
&\tilde{b}_{13} = 0.67\lambda - 4, && \tilde{b}_{14} = -4.11\lambda + 0.78 + 3.33\lambda^{-1}, \\
&\tilde{b}_{15} = -0.11\lambda + 0.11, && \tilde{b}_{16} = -5.25\lambda + 1.17 + 4.08\lambda^{-1}, \\
&\tilde{b}_{22} = 9\lambda + 2, && \tilde{b}_{23} = -0.5\lambda + 1, \\
&\tilde{b}_{24} = -3.17\lambda + 4.17 - \lambda^{-1}, && \tilde{b}_{25} = 0.02\lambda - 0.02, \\
&\tilde{b}_{26} = -3.92\lambda + 3.05 + 0.87\lambda^{-1}, && \tilde{b}_{33} = 8, \\
&\tilde{b}_{34} = -0.17\lambda + 7.17 - 7\lambda^{-1}, && \tilde{b}_{35} = 0.20\lambda - 0.20, \\
&\tilde{b}_{36} = 0.03\lambda + 4.34 - 4.37\lambda^{-1}, && \tilde{b}_{44} = 0.13\lambda + 8.93 + 10.22\lambda^{-1}, \\
&\tilde{b}_{45} = 1.65\lambda - 3.37 + 1.11\lambda^{-1}, && \tilde{b}_{46} = 2.20\lambda + 2.66 + 3.11\lambda^{-1} \\
&\tilde{b}_{55} = -1.82\lambda + 7.47, && \tilde{b}_{56} = -0.20\lambda + 3.63, \\
&\tilde{b}_{66} = 2.51\lambda + 4.07 + 19.06\lambda^{-1}, &&
\end{aligned}$$

where we simplify $\lambda_K$ as $\lambda$.

Notice

$$|-0.17\lambda + 7.17 - 7\lambda^{-1}| \le 0.17\lambda + 6.49 - 6.66\lambda^{-1},$$

$$|1.65\lambda - 3.37 + 1.11\lambda^{-1}| \le 0.57\lambda - 1.07 + 1.11\lambda^{-1}.$$

Thus, under the given conditions it is easy to check that

$$\begin{aligned}
\tilde{b}_{11} - \sum_{j\neq 1} |\tilde{b}_{1j}| &\geq -2.91\lambda + 1.39 + 7.41\lambda^{-1} \geq 1.96,\\
\tilde{b}_{22} - \sum_{j\neq 2} |\tilde{b}_{2j}| &\geq -2.64\lambda + 6.24 - 0.13\lambda^{-1} \geq 1.23,\\
\tilde{b}_{33} - \sum_{j\neq 3} |\tilde{b}_{3j}| &\geq 0.77\lambda - 7.63 + 11.03\lambda^{-1} \geq 0.75,\\
\tilde{b}_{44} - \sum_{j\neq 4} |\tilde{b}_{4j}| &\geq -10.09\lambda + 5.8 + 14.99\lambda^{-1} \geq 0.65,\\
\tilde{b}_{55} - \sum_{j\neq 5} |\tilde{b}_{5j}| &\geq -2.52\lambda + 5.24 - 1.11\lambda^{-1} \geq 0.72,\\
\tilde{b}_{66} - \sum_{j\neq 6} |\tilde{b}_{6j}| &\geq -8.69\lambda - 2.34 + 25.27\lambda^{-1} \geq 1.46.
\end{aligned}$$

Hence the Gerschgorin theorem guarantees that the minimum eigenvalue of $\tilde{B}_K$ is not less than 0.65. So by (4.3.17) there exists a constant $\bar{\alpha} > 0$ such that

$$\begin{aligned}
& I_K(u_h, \Pi_h^* u_h)\\
\geq\ & \frac{0.65}{32S_K}\delta_K(u_h)^T(G_2^{-1}G_1^{-1})^T(G_2^{-1}G_1^{-1})\delta_K(u_h)\\
\geq\ & \frac{\bar{\alpha}}{S_K}\delta_K(u_h)^T\delta_K(u_h), \quad \forall u_h \in U_h.
\end{aligned}$$

Now the desired result follows from (4.3.7) and Theorem 4.3.2. This completes the proof. □

The existence and the uniqueness of the solution to the generalized difference scheme (4.3.6) results from Theorem 4.3.3.

We pause to point out that a more careful estimation can relax the restriction on the value of $\lambda_K$. Our purpose to introduce the last two terms in (4.3.8) is to obtain the uniform ellipticity of the discrete problem. If we multiply these two terms by a parameter $\mu$, then for $\frac{1}{2} \leq \mu \leq \frac{3}{2}$ we have the uniform ellipticity for $\lambda_K$ in a certain range.

**Lemma 4.3.1** *Let $\Lambda_h$ be a piecewise linear interpolation operator with respect to $T_h$:*

$$\forall K \in T_h,\ \Lambda_h w(P_0) = w(P_0)\ \ (P_0 \in \dot{K});\ \Lambda_h w|_K \in \mathcal{P}_1(K).$$

*Then for any $w \in U_h$ and on every $K \in T_h$ we have*

$$|\Pi_h^* w_h - \Lambda_h w_h| \le Ch|w_h|_{2,K}, \tag{4.3.18}$$

$$\Big|\frac{\partial}{\partial x_i}(\Pi_h^* w_h - \Lambda_h w_h)\Big| \le C|w_h|_{2,K},\ i = 1, 2. \tag{4.3.19}$$

*Here $C$ denotes a constant independent of the subspace $U_h$ and the element $K$.*

**Proof** Let $\overline{P_iP_j}$ and $\overline{P_iP_k}$ be the two sides of the triangle $K = \triangle P_iP_jP_k$, parallel respectively to $x_1$ and $x_2$ axes. Then any linear function $\Lambda_h w_h$ on $K_{P_0}^* \cap K$ can be expressed as

$$\begin{aligned}\Lambda_h w_h = \ & w_h(P_l) + \frac{w_h(P_j) - w_h(P_i)}{x_1(P_j) - x_1(P_i)}(x_1 - x_1(P_l)) \\ & + \frac{w_h(P_k) - w_h(P_i)}{x_2(P_k) - x_2(P_i)}(x_2 - x_2(P_l)).\end{aligned} \tag{4.3.20}$$

Hence it follows from (4.3.10) and (4.3.18) that on $K_{P_l}^* \cap K$

$$\begin{aligned}& \Pi_h^* w_h - \Lambda_h w_h \\ = \ & \Big(\frac{\partial w_h(P_l)}{\partial \lambda_j} + w_h(P_i) - w_h(P_j)\Big)\frac{x_1 - x_1(P_l)}{x_1(P_j) - x_1(P_i)} \\ & + \Big(\frac{\partial w_h(P_l)}{\partial \lambda_k} + w_h(P_i) - w_h(P_k)\Big)\frac{x_2 - x_2(P_l)}{x_2(P_k) - x_2(P_i)}.\end{aligned}$$

This together with (4.3.15) gives

$$|\Pi_h^* w_h - \Lambda_h w_h| \le 2(\delta_K(w_h)^T\delta_K(w_h))^{1/2} \le Ch|w_h|_{2,K},$$

$$\begin{aligned}& \Big|\frac{\partial}{\partial x_1}(\Pi_h^* w_h - \Lambda_h w_h)\Big| \\ \le \ & \frac{1}{|x_1(P_j) - x_1(P_i)|}(\delta_K(w_h)^T\delta_K(w_h))^{1/2} \le C|w_h|_{2,K}.\end{aligned}$$

Similarly we can show (4.3.19) for $i = 2$. This completes the proof.□

In the sequel, we always use $u_h$ to denote the solution to the nonconforming generalized difference scheme (4.3.6).

**Theorem 4.3.4** *If the grid $T_h$ satisfies $\frac{2}{3} \le \lambda_K \le \frac{3}{2}$ for every $K \in T_h$, and the weak solution $u \in H^3(\Omega) \cap H_0^2(\Omega)$, then we have a constant $C$ independent of the subspace $U_h$ such that*

$$|u - u_h|_{2,h} \le Ch(|u|_3 + h|f|_0). \tag{4.3.21}$$

**Proof** Let $\Pi_h u$ be the $U_h$-interpolation of $u$, and $w_h = u_h - \Pi_h u$. Then by the uniform ellipticity (4.3.16) and the generalized difference scheme (4.3.6) we have

$$\begin{aligned} &\alpha|u_h - \Pi_h u|_{2,h}^2 \le a_h(u_h - \Pi_h u, \Pi_h^*(u_h - \Pi_h u)) \\ = \ &(f, \Pi_h^* w_h - \Lambda_h w_h) + (f, \Lambda_h w_h) - a_h(\Pi_h u, \Pi_h^* w_h). \end{aligned} \tag{4.3.22}$$

Note $\Delta^2 u \in H^{-1}(\Omega)$ since $u \in H^3(\Omega)$. In terms of Green's formula we have for any $v \in C_0^\infty(\Omega)$

$$(f, v) = (\Delta^2 u, v) = -\int_\Omega \nabla\Delta u \cdot \nabla v \mathrm{d}x.$$

Due to the density of $C_0^\infty(\Omega)$ in $H_0^1(\Omega)$, the above equality is also valid for $v = \Lambda_h w_h \in H_0^1(\Omega)$:

$$(f, \Lambda_h w_h) = -\int_\Omega \nabla\Delta u \cdot \nabla\Lambda_h w_h \mathrm{d}x. \tag{4.3.23}$$

Use Green's formula on each $K^* \in T_h^*$ and note $\Pi_h^* w_h|_{K^*} \in \mathcal{P}_1(K^*)$ to obtain

$$\begin{aligned} &\sum_{K^* \in T_h^*} \int_{K^*} \nabla\Delta u \cdot \nabla\Pi_h^* w_h \mathrm{d}x \\ = &\sum_{K^* \in T_h^*} \Big(-\int_{K^*} \Delta u \cdot \Delta\Pi_h^* w_h \mathrm{d}x + \int_{\partial K^*} \Delta u \frac{\partial \Pi_h^* w_h}{\partial n} \mathrm{d}s\Big) \\ = &\sum_{K^* \in T_h^*} \int_{\partial K^*} \Delta u \frac{\partial \Pi_h^* w_h}{\partial n} \mathrm{d}s. \end{aligned} \tag{4.3.24}$$

By (4.3.7) and (4.3.8) we have

$$\begin{aligned} & a_h(\Pi_h u, \Pi_h^* w_h) \\ = & \sum_{K\in T_h}\sum_{P\in \dot{K}}\int_{\partial K_P^*\cap K}\Big(\frac{\partial \Delta \Pi_h u}{\partial n}\Pi_h^* w_h - \Delta\Pi_h u\frac{\partial \Pi_h^* w_h}{\partial n} \\ & +\frac{\partial^2 \Pi_h u}{\partial \tau^2}\frac{\partial \Pi_h^* w_h}{\partial n} - \frac{\partial^2 \Pi_h u}{\partial n\partial \tau}\frac{\partial \Pi_h^* w_h}{\partial \tau}\Big)\mathrm{d}s. \end{aligned} \tag{4.3.25}$$

Apply Green's formula (4.3.3) on each $K^*\cap K (\neq \emptyset)$, then we have

$$\begin{aligned} & \int_{\partial(K^*\cap K)}\Big(-\frac{\partial^2 u}{\partial \tau^2}\frac{\partial v}{\partial n} + \frac{\partial^2 u}{\partial n\partial \tau}\frac{\partial v}{\partial \tau}\Big)\mathrm{d}s \\ = & \int_{K^*\cap K}\Big(2\frac{\partial^2 u}{\partial x_1\partial x_2}\frac{\partial^2 v}{\partial x_1 \partial x_2} - \frac{\partial^2 u}{\partial x_1^2}\frac{\partial^2 v}{\partial x_2^2} - \frac{\partial^2 u}{\partial x_2^2}\frac{\partial^2 v}{\partial x_1^2}\Big)\mathrm{d}x. \end{aligned}$$

The right-hand side vanishes if we take $v = \Pi_h^* w_h$. So

$$\begin{aligned} & \sum_{K^*\in T_h^*}\int_{\partial K^*}\Big(\frac{\partial^2 u}{\partial \tau^2}\frac{\partial \Pi_h^* w_h}{\partial n} - \frac{\partial^2 u}{\partial n\partial \tau}\frac{\partial \Pi_h^* w_h}{\partial \tau}\Big)\mathrm{d}s \\ = & \sum_{K^*\in T_h^*}\sum_{K\cap K^*\neq\emptyset}\int_{\partial K\cap K^*}\Big(-\frac{\partial^2 u}{\partial \tau^2}\frac{\partial \Pi_h^* w_h}{\partial n} + \frac{\partial^2 u}{\partial n\partial \tau}\frac{\partial \Pi_h^* w_h}{\partial \tau}\Big)\mathrm{d}s = 0. \end{aligned} \tag{4.3.26}$$

The last equality holds since here the contribution of a line integral is always null no matter whether $\partial K$ is a common side of two adjacent elements or it is on $\partial\Omega$, thanks to the continuity of the integral functions and the boundary condition.

A combination of (4.3.22)-(4.3.26) leads to

$$\begin{aligned} & \alpha|u_h - \Pi_h u|_{2,h}^2 \\ \leq & (f, \Pi_h^* w_h - \Lambda_h w_h) + \sum_{i=1}^{4}\sum_{K\in T_h} E_i^K(u, w_h), \end{aligned} \tag{4.3.27}$$

where

$$E_1^K(u, w_h) = \sum_{P\in\dot{K}}\int_{K_P^*\cap K}\nabla\Delta u\cdot\nabla(\Pi_h^* w_h - \Lambda_h w_h)\mathrm{d}x,$$

$$E_2^K(u,w_h)=\sum_{P\in\dot{K}}\int_{\partial K_P^*\cap K}-\frac{\partial\Delta\Pi_h u}{\partial n}\Pi_h^* w_h \mathrm{d}s,$$

$$E_3^K(u,w_h)=\sum_{P\in\dot{K}}\int_{\partial K_P^*\cap K}\Big[-\Delta(u-\Pi_h u)+\frac{\partial^2(u-\Pi_h u)}{\partial\tau^2}\Big]\frac{\partial\Pi_h^* w_h}{\partial n}\mathrm{d}s,$$

$$E_4^K(u,w_h)=\sum_{P\in\dot{K}}\int_{\partial K_P^*\cap K}-\frac{\partial^2(u-\Pi_h u)}{\partial n\partial\tau}\frac{\partial\Pi_h^* w_h}{\partial\tau}\mathrm{d}s.$$

Now we estimate in turn these terms. It follows from the Cauchy inequality, Lemma 4.3.1, (4.3.18) and (4.3.19) that

$$\begin{aligned}&|(f,\Pi_h^* w_h-\Lambda w_h)|\\ \le\ & |f|_0\Big(\sum_{K\in T_h}\int_K|\Pi_h^* w_h-\Lambda_h w_h|^2\mathrm{d}s\Big)^{1/2}\\ \le\ & Ch^2|f|_0|w_h|_{2,h},\end{aligned}\tag{4.3.28}$$

$$|E_1^K(u,w_h)|\le Ch|u|_{3,K}|w_h|_{2,K}.\tag{4.3.29}$$

It is apparent that

$$E_2^K(u,w_h)=\sum_{P\in\dot{K}}\int_{\partial K_P^*\cap K}-\frac{\partial\Delta\Pi_h u}{\partial n}(\Pi_h^* w_h-\Lambda_h w_h)\mathrm{d}s.$$

Write $\phi=\frac{\partial\Delta\Pi_h u}{\partial n}$ and $L_P=\partial K_P^*\cap K$, and note (4.3.18), then we have

$$\begin{aligned}&|E_2^K(u,w_h)|\\ \le\ & \sum_{P\in\dot{K}}\Big(\int_{L_P}|\phi|^2\mathrm{d}s\Big)^{1/2}\Big(\int_{L_P}|\Pi_h^* w_h-\Lambda_h w_h|^2\mathrm{d}s\Big)^{1/2}\\ \le\ & Ch^{3/2}|w_h|_{2,K}\sum_{P\in\dot{K}}\Big(\int_{L_P}|\phi|^2\mathrm{d}s\Big)^{1/2}.\end{aligned}\tag{4.3.30}$$

By the inequality (3.2.43) we have

$$\begin{aligned}\Big(\int_{L_P}|\phi|^2\mathrm{d}s\Big)^{1/2}&\le Ch^{1/2}(h^{-1}|\phi|_{0,K}+|\phi|_{1,K})\\ &\le Ch^{-1/2}|\Pi_h u|_{3,K}\le Ch^{-1/2}|u|_{3,K}.\end{aligned}\tag{4.3.31}$$

It follows from (4.3.30) and (4.3.31) that

$$|E_2^K(u,w_h)| \le Ch|u|_{3,K}|w_h|_{2,K}. \tag{4.3.32}$$

Similarly we note

$$\begin{aligned} E_3^K(u,w_h) = & \sum_{P\in \dot{K}} \int_{\partial K_P^* \cap K} \Big[-\Delta(u-\Pi_h u) + \frac{\partial^2(u-\Pi_h u)}{\partial \tau^2}\Big] \\ & \cdot \frac{\partial(\Pi_h^* w_h - \Lambda_h w_h)}{\partial n} \mathrm{d}s. \end{aligned}$$

Set $\phi = -\Delta(u-\Pi_h u) + \frac{\partial^2(u-\Pi_h u)}{\partial\tau^2}$, note (4.3.19) and imitate (4.3.31) to obtain

$$\begin{aligned} & |E_3^K(u,w_h)| \\ \le & \sum_{P\in\dot{K}} \Big(\int_{L_P} |\phi|^2 \mathrm{d}s\Big)^{1/2} \Big(\int_{L_P} \Big|\frac{\partial(\Pi_h^* w_h - \Lambda_h w_h)}{\partial n}\Big|^2 \mathrm{d}s\Big)^{1/2} \\ \le & \; Ch^{1/2}(h^{-1}|\phi|_{0,K} + |\phi|_{1,K}) \cdot h^{1/2}|w_h|_{2,K} \\ \le & \; Ch(h^{-1}|u-\Pi_h u|_{2,K} + |u-\Pi_h u|_{3,K})|w_h|_{2,K} \\ \le & \; Ch|u|_{3,K}|w_h|_{2,K}. \end{aligned} \tag{4.3.33}$$

Similarly one can show that

$$|E_4^K(u,w_h)| \le Ch|u|_{3,K}|w_h|_{2,K}. \tag{4.3.34}$$

Combining (4.3.27)-(4.3.29) and (4.3.32)-(4.3.34) yields

$$\alpha|u_h - \Pi_h u|_{2,h}^2 \le Ch(|u|_3 + h|f|_0)|u_h - \Pi_h u|_{2,h},$$

$$|u_h - \Pi_h u|_{2,h} \le Ch(|u|_3 + h|f|_0).$$

This together with a standard estimate for Zienkiewicz elements

$$|u - \Pi_h u|_{2,h} \le Ch|u|_3$$

implies the error estimate (4.3.21) and completes the proof. □

### 4.3.4 Numerical experiment

The above nonconforming generalized difference scheme (4.3.6) is used to approximate the following biharmonic equation:

$$\begin{cases} \Delta^2 u = f, & (x,y) \in \Omega, \qquad (4.3.35a) \\ u = \dfrac{\partial u}{\partial n} = 0, & (x,y) \in \partial\Omega, \qquad (4.3.35b) \end{cases}$$

where $\Omega = (0,\pi) \times (0,\pi)$ and $f = 16\cos 2x \cos 2y - 4\cos 2x - 4\cos 2y$.

Place over $\Omega$ a uniform square grid with a side size $h = \pi/16$, and then further decompose it into a right angle triangulation, as illustrated in Fig. 4.3.3. Derive a discrete system of equations from Scheme (4.3.6), and solve it by the Seidel iteration method. Table 4.3.1 provides a comparison of the numerical solution and the true solution $u = \sin^2 x \cdot \sin^2 y$. We observe that the maximum relative error of the function values is about 0.003, and that the first partial derivative is also satisfactorily approximated.

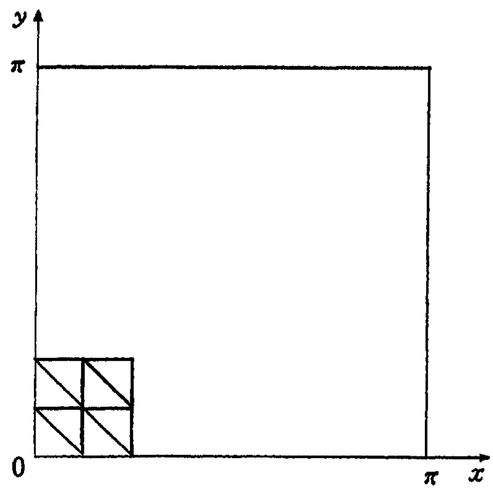



Fig. 4.3.3

**Table 4.3.1 Numerical results** $(x_i = i\pi/8,\ y_j = j\pi/8)$

| $(x,y)$ | $u_h$ | $u_h - u$ | $\frac{\partial u_h}{\partial x}$ | $\frac{\partial u_h}{\partial x} - \frac{\partial u}{\partial x}$ | $\frac{\partial u_h}{\partial y}$ | $\frac{\partial u_h}{\partial y} - \frac{\partial u}{\partial y}$ |
|---|---|---|---|---|---|---|
| $(x_1,y_1)$ | 0.02144 | -.00000 | 0.10642 | .00286 | .10642 | .00286 |
| $(x_1,y_2)$ | 0.07328 | .00006 | 0.35783 | .00427 | .14882 | .00238 |
| $(x_1,y_3)$ | 0.12495 | -.00005 | 0.60713 | .00357 | .10481 | .00126 |
| $(x_1,y_4)$ | 0.14628 | -.00017 | 0.70843 | .00132 | .00019 | .00019 |
| $(x_2,y_2)$ | 0.25070 | .00070 | 0.50403 | .00403 | .50400 | .00400 |
| $(x_2,y_3)$ | 0.42780 | .00102 | 0.85890 | .00535 | .35532 | .00177 |
| $(x_2,y_4)$ | 0.50081 | .00081 | 1.00482 | .00482 | -.00060 | -.00060 |
| $(x_3,y_3)$ | 0.73075 | .00220 | 0.60687 | .00332 | .60677 | .00322 |
| $(x_3,y_4)$ | 0.85593 | .00238 | 0.71117 | .00406 | -.00076 | -.00076 |
| $(x_4,y_4)$ | 1.00308 | .00308 | -0.00010 | -.00010 | -.00027 | -.00027 |

## 4.4 Nonconforming Generalized Difference Methods Based on Adini Elements

### 4.4.1 Generalized difference scheme

As in the last section, we again consider the Dirichlet problem of the biharmonic equation:

$$\begin{cases} \Delta^2 u = f, & (x_1, x_2) \in \Omega, \quad (4.4.1\text{a}) \\ u = \dfrac{\partial u}{\partial n} = 0, & (x_1, x_2) \in \partial\Omega, \quad (4.4.1\text{b}) \end{cases}$$

Assume that each side of the polygon region $\overline{\Omega}$ is parallel to a coordinate axis. So we can divide $\overline{\Omega}$ to obtain a grid $T_h$ consisting of Adini rectangular elements (cf. [p.364, B-17]). Let $h$ be the maximum diameter of the elements, let the vertexes of the rectangles be the nodes, let $\overline{\Omega}_h$ be the set of all the nodes, and let $\dot{\Omega}_h = \overline{\Omega}_h \backslash \partial\Omega$. The trial function space is chosen as the finite element space with respect to the Adini rectangle, i.e., the incomplete bi-cubic, Hermite type, polynomial space. Any function $u_h \in U_h$ satisfies $u_h(P_0) = \frac{\partial u_h(P_0)}{\partial x_1} = \frac{\partial u_h(P_0)}{\partial x_2} = 0$ at every boundary node $P_0$.

In each rectangular element, connect the midpoints of every two opposite sides. Then we re-divide $\overline{\Omega}$ into a sum of some other small rectangles or polygons. Each node $P_0$ of $T_h$ has a surrounding small rectangle (or possibly a small polygon if $P_0$ is a boundary node), called a dual element and denoted by $K_{P_0}^*$ (cf. Fig. 4.4.1). The entire dual elements constitute a dual grid $T_h^*$. The test function space is taken as the piecewise linear function space, which has three basis functions for each interior node $P_0$ of $T_h$:

$$\psi_{P_0}^{(0)}(P) = \begin{cases} 1, & P \in K_{P_0}^*, \\ 0, & P \notin K_{P_0}^*, \end{cases}$$

$$\psi_{P_0}^{(1)}(P) = \begin{cases} x_1 - x_1(P_0), & P \in K_{P_0}^*, \\ 0, & P \notin K_{P_0}^*, \end{cases}$$

$$\psi_{P_0}^{(2)}(P) = \begin{cases} x_2 - x_2(P_0), & P \in K_{P_0}^*, \\ 0, & P \notin K_{P_0}^*. \end{cases}$$

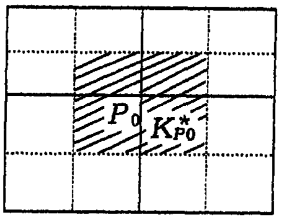



Fig. 4.4.1

Any function $v_h \in V_h$ also satisfies $v_h(P_0) = \frac{\partial v_h(P_0)}{\partial x_1} = \frac{\partial v_h(P_0)}{\partial x_2} = 0$ at any boundary node $P_0$.

Based on the variational form (4.3.4), the generalized difference scheme for (4.4.1) is: Find $u_h \in U_h$ such that

$$a_h(u_h, v_h) = (f, v_h), \quad \forall v_h \in V_h, \tag{4.4.2}$$

where

$$a_h(u_h, v_h) = \sum_{K \in T_h} I_K(u_h, v_h), \tag{4.4.3}$$

$$I_K(u_h, v_h) = \sum_{P \in \dot{K}} \int_{\partial K_P^* \cap K} \Big(\frac{\partial \Delta u_h}{\partial n} v_h - \Delta u_h \frac{\partial v_h}{\partial n} + \frac{\partial^2 u_h}{\partial \tau^2}\frac{\partial v_h}{\partial n} - \frac{\partial^2 u_h}{\partial n \partial \tau}\frac{\partial v_h}{\partial \tau}\Big) ds, \tag{4.4.4}$$

where $\dot{K}$ denotes the set of all the vertexes of the element $K$.

Obviously $U_h \subset H^1(\Omega)$, $V_h \in L^2(\Omega)$. But $U_h \not\subset H^2(\Omega)$, $V_h \not\subset H^1(\Omega)$. Thus (4.4.2) is a nonconforming scheme. We see from the supports of the basis functions $\psi_{P_0}^{(l)}$ $(l = 0, 1, 2)$ that the resulting discrete equation is a nine point generalized difference scheme.

We can successively compute the discrete equation

$$a_h(u_h, \psi_{P_0}^{(l)}) = (f, \psi_{P_0}^{(l)}), \ l = 0, 1, 2, \ P_0 \in \dot{\Omega}_h$$

for every node $P_0 \in \dot{\Omega}_h$; or we can first compute $I_K(u_h, \psi_{P_0}^{(l)})$ $(l = 0, 1, 2)$ to get the element matrices, and then pile them up to obtain an algebraic system of the discrete problem. Observing that $\psi_{P_0}^{(l)}$ is very simple and that $I_K$ only involves line integrals with the integral paths being parallel to the coordinate axes, the computation here of the element matrices is simpler and more economical compared with the corresponding nonconforming finite element method.

### 4.4.2 Error estimate

Take any $K \in T_h$ with vertexes $P_m$ $(m = i, j, k, l)$, midpoints $M_{ij}$, $M_{jk}$, $M_{kl}$, $M_{li}$ of the sides, the barycenter $Q$ (cf. Fig. 4.4.2(a)), and the area $S_K$. Set $\Delta x_1 = |\overline{P_iP_j}|$, $\Delta x_2 = |\overline{P_iP_l}|$, and $\lambda_K = (\Delta x_2/\Delta x_1)^2$. Then the mapping

$$\begin{aligned} \xi &= (x_1 - x_1(P_i))/\Delta x_1, \\ \eta &= (x_2 - x_2(P_i))/\Delta x_2 \end{aligned} \tag{4.4.5}$$

maps the rectangle $K$ onto a unit square $\hat{K} = [0,1] \times [0,1]$, and the nodes $P_i, M_i, Q, \cdots$ into $\hat{P}_i, \hat{M}_i, \hat{Q}, \cdots$. (cf. Fig. 4.4.2(b).)

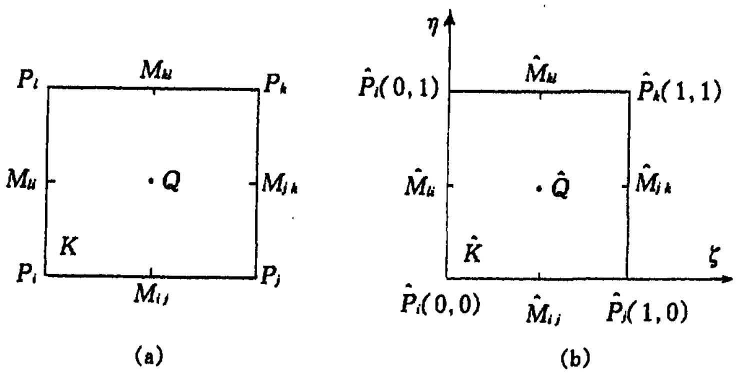



Fig. 4.4.2

Introduce on $U_h$ a discrete norm

$$\|u_h\|_h = \Big(\sum_{K \in T_h} \frac{1}{S_K} \delta_K(u_h)^T \delta_K(u_h)\Big)^{1/2}, \quad u_h \in U_h, \tag{4.4.6}$$

where

$$\begin{aligned}\delta_K(v) = \ & \Big[v_i - v_j + v_k - v_l, \Big(\frac{\partial v}{\partial \xi}\Big)_i + v_i - v_j, \Big(\frac{\partial v}{\partial \xi}\Big)_j + v_i - v_j, \\ & \Big(\frac{\partial v}{\partial \xi}\Big)_k + v_l - v_k, \Big(\frac{\partial v}{\partial \xi}\Big)_l + v_l - v_k, \Big(\frac{\partial v}{\partial \eta}\Big)_i + v_i - v_l, \\ & \Big(\frac{\partial v}{\partial \eta}\Big)_j + v_j - v_k, \Big(\frac{\partial v}{\partial \eta}\Big)_k + v_j - v_k, \Big(\frac{\partial v}{\partial \eta}\Big)_l + v_i - v_l\Big]^T.\end{aligned}$$

Here we write for short $v_i = v_h(P_i)$ etc..

**Theorem 4.4.1** *Suppose the grid $T_h$ is quasi-uniform: There exists a $\lambda_0 > 0$ such that $\lambda_0 \le \lambda_K \le \lambda_0^{-1}$, $\forall K \in T_h$. Then the norm $\|\cdot\|_h$ is equivalent to the norm $|\cdot|_{2,h}$ defined as follows*

$$|u_h|_{2,h} = \Big(\sum_{K\in T_h} |u_h|_{2,K}^2\Big)^{1/2},\ u_h \in U_h,$$

*namely, there exist constants $C_1, C_2 > 0$ independent of $U_h$ such that*

$$C_1\|u_h\|_h \le |u_h|_{2,h} \le C_2\|u_h\|_h,\ \ \forall u_h \in U_h. \tag{4.4.7}$$

**Proof** We only have to show the existence of constants $C_1', C_2' > 0$ satisfying

$$\begin{aligned}\frac{C_1'}{S_K}\delta_K(u_h)^T\delta_K(u_h) \le |u_h|_{2,K}^2 \le \frac{C_2'}{S_K}\delta_K(u_h)^T\delta_K(u_h),\\ \forall u_h \in U_h,\ K \in T_h.\end{aligned} \tag{4.4.8}$$

The definition of the Adini element on $K$ is as follows:

$$\begin{aligned}u_h = \ & (-2\xi^3\eta + 2\xi^3 + 3\xi^2\eta - 3\xi^2 - 2\xi\eta^3 \\ & +3\xi\eta^2 - \xi\eta + 2\eta^3 - 3\eta^2 + 1)(u_h)_i \\ & -(-2\xi^3\eta + 2\xi^3 + 3\xi^2\eta - 3\xi^2 - 2\xi\eta^3 + 3\xi\eta^2 - \xi\eta)(u_h)_j \\ & +(-2\xi^3\eta + 3\xi^2\eta - 2\xi\eta^3 + 3\xi\eta^2 - \xi\eta)(u_h)_k\end{aligned}$$

$$
\begin{aligned}
&-(-2\xi^3\eta+3\xi^2\eta-2\xi\eta^3+3\xi\eta^2-\xi\eta+2\eta^3-3\eta^2)(u_h)_l\\
&+(-\xi^3\eta+\xi^3+2\xi^2\eta-2\xi^2-\xi\eta+\xi)\Big(\frac{\partial u_h}{\partial\xi}\Big)_i\\
&+(-\xi^3\eta+\xi^3+\xi^2\eta-\xi^2)\Big(\frac{\partial u_h}{\partial\xi}\Big)_j+(\xi^3\eta-\xi^2\eta)\Big(\frac{\partial u_h}{\partial\xi}\Big)_k\\
&+(\xi^3\eta-2\xi^2\eta+\xi\eta)\Big(\frac{\partial u_h}{\partial\xi}\Big)_l \qquad (4.4.9)\\
&+(-\xi\eta^3+2\xi\eta^2-\xi\eta+\eta^3-2\eta^2+\eta)\Big(\frac{\partial u_h}{\partial\eta}\Big)_i\\
&+(\xi\eta^3-2\xi\eta^2+\xi\eta)\Big(\frac{\partial u_h}{\partial\eta}\Big)_j\\
&+(\xi\eta^3-\xi\eta^2)\Big(\frac{\partial u_h}{\partial\eta}\Big)_k+(-\xi\eta^3+\xi\eta^2+\eta^3-\eta^2)\Big(\frac{\partial u_h}{\partial\eta}\Big)_l.
\end{aligned}
$$

Let us express $\frac{\partial^2 u_h}{\partial x_1^2}$, $\frac{\partial^2 u_h}{\partial x_1\partial x_2}$ and $\frac{\partial^2 u_h}{\partial x_2^2}$ as multiplications of vectors and matrices:

$$\frac{\partial^2 u_h}{\partial x_1^2}=\frac{1}{\Delta x_1^2}\frac{\partial^2 u_h}{\partial\xi^2}=\frac{1}{\Delta x_1^2}(\xi\eta,\xi,\eta,1)G_1\delta_K(u_h),$$

$$\frac{\partial^2 u_h}{\partial x_1\partial x_2}=\frac{1}{\Delta x_1\Delta x_2}\frac{\partial^2 u_h}{\partial\xi\partial\eta}=\frac{1}{\Delta x_1\Delta x_2}(\xi^2,\eta^2,\xi,\eta,1)G_2\delta_K(u_h),$$

$$\frac{\partial^2 u_h}{\partial x_2^2}=\frac{1}{\Delta x_2^2}\frac{\partial^2 u_h}{\partial\eta^2}=\frac{1}{\Delta x_2^2}(\xi\eta,\xi,\eta,1)G_3\delta_K(u_h),$$

where

$$G_1=\begin{bmatrix}0&-6&-6&6&6&0&0&0&0\\0&6&6&0&0&0&0&0&0\\0&4&2&-2&-4&0&0&0&0\\0&-4&-2&0&0&0&0&0&0\end{bmatrix},$$

$$G_2=\begin{bmatrix}0&-3&-3&3&3&0&0&0&0\\0&0&0&0&0&-3&3&3&-3\\0&4&2&-2&-4&0&0&0&0\\0&0&0&0&0&4&-4&-2&2\\1&-1&0&0&1&-1&1&0&0\end{bmatrix},$$

$$G_3 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & -6 & 6 & 6 & -6 \\ 0 & 0 & 0 & 0 & 0 & 4 & -4 & -2 & 2 \\ 0 & 0 & 0 & 0 & 0 & 6 & 0 & 0 & 6 \\ 0 & 0 & 0 & 0 & 0 & -4 & 0 & 0 & -2 \end{bmatrix}.$$

A direct computation leads to

$$\begin{aligned} |u_h|^2_{2,K} = & \int_{\hat{K}} \Big[ \frac{1}{\Delta x_1^4} \Big( \frac{\partial^2 u_h}{\partial \xi^2} \Big)^2 + \frac{1}{\Delta x_1^2 \Delta x_2^2} \Big( \frac{\partial^2 u_h}{\partial \xi \partial \eta} \Big)^2 \\ & + \frac{1}{\Delta x_2^4} \Big( \frac{\partial^2 u_h}{\partial \eta^2} \Big)^2 \Big] \Delta x_1 \Delta x_2 \mathrm{d}\xi \mathrm{d}\eta \\ = & \frac{1}{S_K} \delta_K(u_h)^T G^T \mathrm{diag}(\lambda_K D_0, D_1, \lambda_K^{-1} D_0) G \delta_K(u_h), \end{aligned} \tag{4.4.10}$$

where

$$D_0 = \int_{\hat{K}} (\xi\eta, \xi, \eta, 1)(\xi\eta, \xi, \eta, 1)^T \mathrm{d}\xi \mathrm{d}\eta,$$

$$D_1 = \int_{\hat{K}} (\xi^2, \eta^2, \xi, , \eta, 1)(\xi^2, \eta^2, \xi, \eta, 1)^T \mathrm{d}\xi \mathrm{d}\eta,$$

$$G^T = \left[ G_1^T G_2^T G_3^T \right].$$

One can easily verify that $D_0$ and $D_1$ are positive definite matrices, and that the column vectors of $G$ are linearly independent. (4.4.8) finally follows by use of the quasi-uniform condition of the decomposition. This completes the proof. □

Define an interpolation operator $\Pi_h^* : U_h \to V_h$ as follows:

$$\Pi_h^* w_h = \sum_{P_0 \in \dot{\Omega}_h} \Big[ w_h(P_0) \psi_{P_0}^{(0)} + \frac{\partial w_h(P_0)}{\partial x_1} \psi_{P_0}^{(1)} + \frac{\partial w_h(P_0)}{\partial x_2} \psi_{P_0}^{(2)} \Big]. \tag{4.4.11}$$

**Theorem 4.4.2** *Suppose the grid $T_h$ satisfies the condition $\frac{2}{3} \le \lambda_K \le \frac{3}{2}$ for all $K \in T_h$. Then the bilinear form $a_h(\cdot, \Pi_h^* \cdot)$ is uniformly positive definite: There exists a constant $\alpha > 0$ independent of $U_h$ such that*

$$a_h(u_h, \Pi_h^* u_h) \ge \alpha |u_h|^2_{2,h}, \quad \forall u_h \in U_h. \tag{4.4.12}$$

In the sequel we use $u$ and $u_h$ to denote the weak solution of (4.4.1) and the solution of the nonconforming generalized difference scheme (4.4.2), respectively.

**Theorem 4.4.3** *Assume that* $\frac{2}{3} \leq \lambda_K \leq \frac{3}{2}$ *for all* $K \in T_h$, *and that the weak solution* $u \in H^3(\Omega) \cap H_0^2(\Omega)$, *then there exists a constant* $C$ *independent of* $U_h$ *such that*

$$|u - u_h|_{2,h} \leq Ch(|u|_3 + h|f|_0). \tag{4.4.13}$$

The proofs to Theorems 4.4.2 and 4.4.3 are omitted to save the space, cf. [A-8].

### 4.4.3 Numerical example

The nonconforming generalized difference method (4.4.2) is used to approximate the following biharmonic equation:

$$\begin{cases} \Delta^2 u = f, & (x_1, x_2) \in \Omega, \\ u = \dfrac{\partial u}{\partial n} = 0, & (x_1, x_2) \in \partial\Omega, \end{cases}$$

where $\Omega = (0, \pi) \times (0, \pi)$ and

$$f(x, y) = 16 \cos 2x \cos 2y - 4 \cos 2x - 4 \cos 2y.$$

**Table 4.4.1 Numerical results** $(x_i = i\pi/8,\ y_j = j\pi/8)$

| $(x, y)$ | $u_h$ | $u_h - u$ | $\frac{\partial u_h}{\partial x}$ | $\frac{\partial u_h}{\partial x} - \frac{\partial u}{\partial x}$ | $\frac{\partial u_h}{\partial y}$ | $\frac{\partial u_h}{\partial y} - \frac{\partial u}{\partial y}$ |
|---|---|---|---|---|---|---|
| $(x_1, y_1)$ | 0.02135 | -0.00010 | 0.10306 | -0.00049 | 0.10306 | -0.00069 |
| $(x_1, y_2)$ | 0.07289 | -0.00033 | 0.35232 | -0.00123 | 0.14589 | -0.00056 |
| $(x_1, y_3)$ | 0.12448 | -0.00051 | 0.60188 | 0.00168 | 0.10326 | -0.00079 |
| $(x_1, y_4)$ | 0.14587 | -0.00058 | 0.70529 | -0.00181 | -0.00010 | -0.00010 |
| $(x_2, y_2)$ | 0.24924 | -0.00076 | 0.49923 | -0.00077 | 0.49921 | -0.00079 |
| $(x_2, y_3)$ | 0.42583 | -0.00095 | 0.85314 | -0.00041 | 0.35334 | -0.00022 |
| $(x_2, y_4)$ | 0.49901 | -0.00099 | 0.99983 | -0.00017 | -0.00040 | -0.00040 |
| $(x_3, y_3)$ | 0.72764 | -0.00091 | 0.60391 | 0.00036 | 0.60387 | 0.00032 |
| $(x_3, y_4)$ | 0.85272 | -0.00083 | 0.70774 | 0.00063 | -0.00006 | -0.00006 |
| $(x_4, y_4)$ | 0.99932 | -0.00068 | -0.00002 | -0.00002 | -0.00006 | -0.00006 |

We place a square grid over $\Omega$ with $h = \pi/16$, and use (4.4.2). The resulting generalized difference solution is compared with the true solution $u = \sin^2 x \sin^2 y$ in Table 4.4.1. We observe that $u_h$, $\frac{\partial u_h}{\partial x}$ and $\frac{\partial u_h}{\partial y}$ are all good approximations. We note that this method behaves better than the Ziekiewicz generalized difference method, cf. the numerical example in §4.3.4.

## 4.5 Second Order Nonlinear Elliptic Equations

In this section we are concerned with the following Dirichlet problem of the second order nonlinear elliptic equation:

$$\begin{cases} -\nabla(a(x,y,u)\nabla u) = f(x,y), & (x,y) \in \Omega, \quad (4.5.1a) \\ u(x,y) = 0, & (x,y) \in \partial\Omega, \quad (4.5.1b) \end{cases}$$

where $\Omega$ is a plane bounded region with a sufficiently smooth boundary $\partial\Omega$; $a(x,y,u)$ is a twice continuously differentiable mapping from $\overline{\Omega} \times R$ to $[\alpha_1, \alpha_2]$ $(0 < \alpha_1 < \alpha_2)$; and all the second order partial derivatives of $a(x,y,u)$ are bounded on $\overline{\Omega} \times R$. By the Schauder theory and [B-25], if $f \in C^\alpha(\overline{\Omega})$ for some integer $\alpha > 0$, then (4.5.1) has a unique weak solution $u$, and $u \in C^{2+\alpha}(\overline{\Omega})$. Set

$$\begin{aligned} A(w;u,v) &= (a(x,y,w)\nabla u, \nabla v) \\ &= \int_\Omega a(x,y,w)\nabla u \cdot \nabla v dxdy. \end{aligned} \tag{4.5.2}$$

Then, a weak form of (4.5.1) is: Find $u \in H_0^1(\Omega)$ such that

$$A(u;u,v) = (f,v), \quad \forall v \in H_0^1(\Omega). \tag{4.5.3}$$

### 4.5.1 Generalized difference scheme

Write for short $a(x,y,w) = a(w)$ and suppose $\Omega$ is a convex polygonal region. As in §3.2 we place a triangulation $T_h$ and its dual grid $T_h^*$ over $\overline{\Omega}$. Let $h$ be the maximum diameter of all the triangular elements of $T_h$. Also assume $T_h$ and $T_h^*$ are quasi-uniform, so that there exist constants $c_1, c_2 > 0$ such that

$$c_1 h^2 \le S_K \le c_2 h^2, \quad \forall K \in T_h,$$

$$c_1 h^2 \le S_{K^*} \le c_2 h^2, \quad \forall K^* \in T_h^*,$$

where $S_K$ and $S_{K^*}$ stand for the areas of $K$ and $K^*$ respectively.

We choose the trial function space as the standard finite element space with respect to $T_h$. Corresponding to the freedom

$$l_{P_0}^{i,j} : \; u \to \frac{\partial^{i+j} u(P_0)}{\partial x^i \partial y^j}$$

for each interpolation node $P_0 = (x_0, y_0)$ of $U_h$, we take the $i, j$ term

$$\psi_{P_0}^{(i,j)}(P) = \begin{cases} (x - x_0)^i (y - y_0)^j / (i!j!), & P = (x, y) \in K_{P_0}^* \\ 0, & P \notin K_{P_0}^* \end{cases}$$

of the local Taylor expansion as the basis function of the test function space $V_h$. In particular, if $U_h$ is a linear element space (piecewise linear polynomial space) with the following freedom related to the node $P_0$

$$l_{P_0}^{(0,0)} : = u \to u(P_0),$$

then $V_h$ is a piecewise constant function space, of which the basis function for $P_0$ is a characteristic function of $K_{P_0}^*$:

$$\psi_{P_0}^{(0,0)}(P) = \begin{cases} 1, & P = (x, y) \in K_{P_0}^*, \\ 0, & P \notin K_{P_0}^*. \end{cases}$$

Now the generalized difference scheme for (5.5.1) is: Find $u_h \in U_h$ such that

$$A(u_h; u_h, v_h) = (f, v_h), \quad \forall v_h \in V_h, \tag{4.5.4}$$

where

$$\begin{aligned} & A(u_h; u_h, v_h) \\ = & \int_\Omega a(u_h) \nabla u_h \cdot \nabla v_h \mathrm{d}x\mathrm{d}y \\ = & \int_\Omega a(u_h) \Big( \frac{\partial u_h}{\partial x} \frac{\partial v_h}{\partial x} + \frac{\partial u_h}{\partial y} \frac{\partial v_h}{\partial y} \Big) \mathrm{d}x\mathrm{d}y. \end{aligned} \tag{4.5.5}$$

Here $\frac{\partial u_h}{\partial y}$ and $\frac{\partial v_h}{\partial y}$ etc. should be interpreted in the sense of generalized functions. Suppose $\psi_h$ is a basis function of $V_h$, whose support is a dual element $K^*_{P_0}$, then

$$\begin{aligned}&\int_\Omega a(u_h)\nabla u_h\cdot\nabla v_h \mathrm{d}x\mathrm{d}y\\ =\ &\int_{K^*_{P_0}} a(u_h)\nabla u_h\cdot\nabla v_h \mathrm{d}x\mathrm{d}y - \int_{\partial K^*_{P_0}} a(u_h)\frac{\partial u_h}{\partial n}\mathrm{d}s,\end{aligned} \tag{4.5.5$'$}$$

where $n$ is the unit outer normal vector along $\partial K^*_{P_0}$.

In particular when $U_h$ is chosen as the linear element space corresponding to $T_h$ and $V_h$ as the piecewise constant function space, then we have the following linear element difference scheme:

$$A(u_h; u_h, \psi^{(0,0)}_{P_0}) = \int_{K^*_{P_0}} f\mathrm{d}x\mathrm{d}y,\ \ \forall P_0\in\dot{\Omega}_h, \tag{4.5.6}$$

where

$$\begin{aligned}&A(u_h; u_h, \psi^{(0,0)}_{P_0})\\ =\ &-\int_{K^*_{P_0}} a(u_h)\frac{\partial u_h}{\partial n}\mathrm{d}s\\ =\ &-\int_{\partial K^*_{P_0}} a(u_h)\frac{\partial u_h}{\partial x}\mathrm{d}y + \int_{\partial K^*_{P_0}} a(u_h)\frac{\partial u_h}{\partial y}\mathrm{d}x.\end{aligned} \tag{4.5.7}$$

Various kinds of numerical methods can be used to compute the line integrals in (4.5.7). For instance, as shown in Fig. 4.5.1, one can write the integral on $\partial K^*_{P_0}$ as a sum of integrals on the fold line segments $\overline{Q_1M_2Q_2}, \cdots, \overline{Q_6M_1Q_1}$, then employ, e.g., the following quadrature formula:

$$\begin{aligned}&-\int_{\overline{Q_1M_2Q_2}} a(u_h)\frac{\partial u_h}{\partial n}\mathrm{d}s\\ \doteq\ &-(|\overline{Q_1M_2}| + |\overline{M_2Q_2}|)(a(u_h))_{M_2}(u_h(P_2)-u_h(P_0))/|\overline{P_0P_2}|,\end{aligned}$$

where $u_h(M_2) = (u_h(P_2)+u_h(P_0))/2$. If $T^*_h$ is a circumcenter dual grid, then (4.5.6) is identical with a finite difference equation derived by an integral conservation law. (See, e.g., [C-6,7].)

Another way is to write the integral on $\partial K_{P_0}^*$ as a sum of integrals on the fold line segments $\overline{M_1Q_1M_2}, \cdots, \overline{M_6Q_6M_1}$, and to use the quadrature formula:

$$\int_{\overline{M_1Q_1M_2}} a(u_h)\frac{\partial u_h}{\partial x}\mathrm{d}y \doteq (y_{M_2} - y_{M_1})(a(u_h))_{Q_1}\frac{\partial u_h(Q_1)}{\partial x},$$

$$\int_{\overline{M_1Q_1M_2}} a(u_h)\frac{\partial u_h}{\partial y}\mathrm{d}x \doteq (x_{M_2} - x_{M_1})(a(u_h))_{Q_1}\frac{\partial u_h(Q_1)}{\partial y}.$$

This leads to another sort of difference equation.

Let $\Pi_h^* : U_h \to V_h$ be an interpolation operator:

$$\Pi_h^*\bar{u}_h = \sum_{P\in\dot{\Omega}_h} \bar{u}(P_0)\psi_{P_0}^{(0,0)}(P), \quad \forall \bar{u}_h \in U_h,$$

then we can rewrite the generalized difference scheme (4.5.4) into an equivalent form:

$$A(u_h; u_h, \Pi_h^*\bar{u}_h) = (f, \Pi_h^*\bar{u}_h), \quad \forall \bar{u}_h \in U_h, \tag{4.5.8}$$

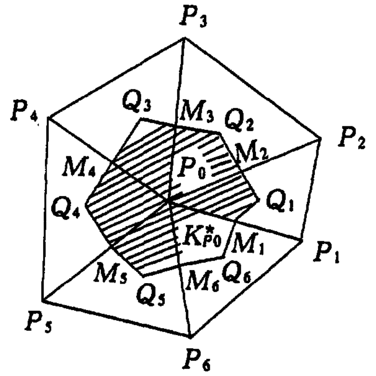



Fig. 4.5.1

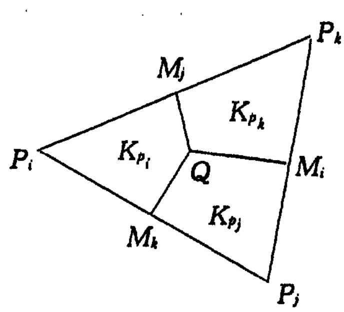



Fig. 4.5.2

### 4.5.2 Error estimate

Next we analyze the linear element scheme. Write $A(w; u_h, \Pi_h^* \bar{u}_h)$ into the following form:

$$A(w; u_h, \Pi_h^* \bar{u}_h) = \sum_{K \in T_h} I_K(w; u_h, \Pi_h^* \bar{u}_h), \tag{4.5.9}$$

where $K = \triangle P_i P_j P_k$ (cf. Fig. 4.5.2), and

$$\begin{aligned} & I_K(w; u_h, \Pi_h^* \bar{u}_h) \\ = \ & (\bar{u}_h(P_j) - \bar{u}_h(P_i)) \int_{\overline{M_k Q}} a(w) \frac{\partial u_h}{\partial n} \mathrm{d}s \\ & + (\bar{u}_h(P_k) - \bar{u}_h(P_j)) \int_{\overline{M_i Q}} a(w) \frac{\partial u_h}{\partial n} \mathrm{d}s \\ & + (\bar{u}_h(P_i) - \bar{u}_h(P_k)) \int_{\overline{M_j Q}} a(w) \frac{\partial u_h}{\partial n} \mathrm{d}s \\ = \ & \sum_{l=i,j,k} |\overline{P_{l+1} P_{l+2}}| \int_{\overline{M_l Q}} a(w) \frac{\partial u_h}{\partial n_l} \frac{\partial \bar{u}_h}{\partial \tau_l} \mathrm{d}s, \end{aligned} \tag{4.5.10}$$

where $n_l$ denotes the unit outer normal vector of $K_{P_{l+2}}$ along $\overline{M_l Q}$, and $\tau_l = \overline{P_{l+1} P_{l+2}} / |\overline{P_{l+1} P_{l+2}}|$. ($l = i, j, k$; $i+1 = j, j+1 = k, k+1 = i$.) Now we are ready to deduce the boundedness and the positive definiteness of $A(w; u_h, \Pi_h^* \bar{u}_h)$.

**Theorem 4.5.1** *For the bilinear form $A(w; \cdot, \Pi_h^* \cdot)$, there exists a constant $C$ independent of $U_h$ such that*

$$|A(w; u_h, \Pi_h^* \bar{u}_h)| \leq C \|u_h\|_1 \|\bar{u}_h\|_1, \quad \forall u_h, \bar{u}_h \in U_h, \tag{4.5.11}$$

$$|A(w; u, \Pi_h^* \bar{u}_h)| \leq C \|u\|_{1,\infty} \|\bar{u}_h\|_1, \quad \forall u \in W^{1,\infty}(\Omega), \bar{u}_h \in U_h. \tag{4.5.12}$$

*Moreover, if $T_h^*$ is a circumcenter dual grid, then we have a constant $\beta > 0$ independent of $U_h$ satisfying*

$$A(w; u_h, \Pi_h^* u_h) \geq \beta \|u_h\|_1^2, \quad \forall u_h \in U_h. \tag{4.5.13}$$

*This also implies the existence of a solution to the linear element difference scheme (4.5.6).*

**Proof** (4.5.11) follows from (4.5.9), (4.5.10) and the equivalent norms defined in §3.2 :

$$\begin{aligned}
&|A(w; u_h, \Pi_h^* \bar{u}_h)| \\
\leq\ & C \sum_{K \in T_h} h^2 \Big( \Big| \frac{\partial u_h(Q)}{\partial x} \Big| + \Big| \frac{\partial u_h(Q)}{\partial y} \Big| \Big) \Big( \Big| \frac{\partial \bar{u}_h(Q)}{\partial x} \Big| + \Big| \frac{\partial \bar{u}_h(Q)}{\partial y} \Big| \Big) \\
\leq\ & C \|u_h\|_1 \|\bar{u}_h\|_1.
\end{aligned}$$

(4.5.12) can be similarly proved. In the case of the circumcenter dual grid, we have $n_l = \tau_l$. Hence by use of (4.5.10) we have

$$\begin{aligned}
&I_K(w; u_h, \Pi_h^* u_h) \\
=\ & \sum_{l=i,j,k} |\overline{P_{l+1} P_{l+2}}| \int_{\overline{M_l Q}} a(w) \Big( \frac{\partial u_h}{\partial \tau_l} \Big)^2 ds \\
\geq\ & \beta' S_K \Big( \Big( \frac{\partial u_h(Q)}{\partial x} \Big)^2 + \Big( \frac{\partial u_h(Q)}{\partial y} \Big)^2 \Big). \quad (\beta' > 0.)
\end{aligned}$$

This gives (4.5.13).

To show the solvability of (4.5.6) we define $T: U_h \to U_h$ by

$$A(w_h; Tw_h, \Pi_h^* \bar{u}_h) = (f, \Pi_h^* \bar{u}_h), \quad \forall \bar{u}_h \in U_h. \tag{4.5.14}$$

By virtue of (4.5.11) and (4.5.13), we have the existence and the uniqueness of the solution $Tw_h$ as well as the estimate

$$\|Tw_h\|_1 \leq \|f\|_0 / \beta.$$

Thus the mapping $T$ maps the ball $\{w_h \in U_h : \|w_h\|_1 \leq \|f\|_0/\beta\}$ to itself. Also note that $T$ is obviously a continuous mapping. Therefore the Brouwer fixed point theorem guarantees the existence of the solution $u_h \in U_h$ and the estimate

$$\|u_h\|_1 \leq \|f\|_0 / \beta.$$

This completes the proof. □

**Theorem 4.5.2** *Let $u$ be the solution of the weak form (4.5.3) of the problem (4.5.1), and $u_h \in U_h$ the solution of the linear element generalized difference scheme (4.5.6). If $u \in W^{2,\infty}(\Omega)$, then there holds the following error estimate:*

$$\|u - u_h\|_1 \leq Ch.$$

The proof to this theorem is omitted. For the details, see [B-58].

## Bibliography and Comments

This chapter provides some main results on the mixed and the nonconforming generalized difference methods for the boundary value problem of the fourth order elliptic equation. As regards the mixed method, [A-51] gives a mixed generalized difference method for biharmonic equations, based on a Ciarlet-Raviart mixed variational principle. Another mixed generalized difference method is presented in [A-31] based on a Hermann-Miyoshi mixed variational principle. But there are some errors in the proofs of the error estimations in both the two papers. A refined error analysis is provided in §4.1 (Theorem 4.1.2). In order to construct nonconforming generalized difference methods for biharmonic equations, a corresponding variational principle is discussed and is used to give a nonconforming generalized difference scheme with Zienkiewicz elements [A-9]. Another nonconforming generalized difference method is proposed in [A-8], based on Adini elements. The error analysis and numerical experiments are carries out in these two papers. Theoretical analysis indicates that these nonconforming methods enjoy the same error estimate as the corresponding finite element methods. Both the mixed and the nonconforming generalized difference methods require less computational time than the finite element methods, and have better accuracy than the usual finite difference methods. We also note that it is easy for them to deal with complex boundaries and various of boundary conditions.

With regard to nonlinear elliptic equations, [B-58] constructed a generalized difference scheme for a second order nonlinear Dirichlet problem, and presented corresponding error estimates.

**Problem** The cubic generalized difference method has a stronger non-conformity since the piecewise cubic element space $U_h$ generally is not contained in $H^2$ when the dimension $n \geq 2$. This brings new difficulties for the construction of the difference scheme and for the error estimation. But if we adopt the Hsich-Clough-Tocher triangular elements (cf. [A-27,2] and [p.340, B-17]), then we indeed have $U_h \in H^2$. Try to use such a $U_h$ (and a certain corresponding $V_h$) to construct generalized difference schemes and to deduce the error estimates.

# Chapter 5

# PARABOLIC EQUATIONS

We present in this chapter, for second order parabolic differential equations, semi- and fully-discrete generalized difference methods and one of their varieties – a mass concentration method. The construction as well as the theoretical analysis of these schemes are discussed. A nonlinear parabolic equation is considered in the last section.

## 5.1 Semi-discrete Generalized Difference Schemes

### 5.1.1 Problem and schemes

Consider the parabolic differential equation:

$$\begin{cases} u_t + Au = f(x,t), & x \in \Omega,\ 0 < t \le T, & \text{(5.1.1a)} \\ u = 0, & x \in \partial\Omega,\ 0 < t \le T, & \text{(5.1.1b)} \\ u = u_0(x), & x \in \Omega,\ t = 0, & \text{(5.1.1c)} \end{cases}$$

where $\Omega$ is a bounded region in $R^n$, with a Lipschitz continuous boundary; $u_t = \frac{\partial u}{\partial t}$; and $A$ a second order elliptic differential operator:

$$Au \equiv -\sum_{i,j}\frac{\partial}{\partial x_i}\Big(a_{ij}\frac{\partial u}{\partial x_j}\Big)+\sum_j b_j\frac{\partial u}{\partial x_j}+cu,$$

where $a_{ij}(=a_{ji})$, $b_j$ and $c$ are sufficiently smooth functions of $x$. We assume there is a constant $\alpha_0 > 0$ satisfying

$$\begin{aligned} a(u,u) &= \int_\Omega\Big(\sum_{ij}a_{ij}\frac{\partial u}{\partial x_i}\frac{\partial u}{\partial x_j}+\sum_j b_j\frac{\partial u}{\partial x_j}u+cu^2\Big)\mathrm{d}x \\ &\geq \alpha_0\|u\|_1^2, \quad \forall u \in H_0^1(\Omega). \end{aligned} \tag{5.1.2}$$

The variational problem related to (5.1.1) is: Find $u = u(\cdot,t) \in H_0^1(\Omega)$ $(0 \leq t \leq T)$ such that

$$\begin{cases} (u_t,v)+a(u,v)=(f,v), & \forall v \in H_0^1(\Omega),\ t>0, \quad (5.1.3a) \\ u(x,0)=u_0(x), & x\in\Omega, \quad (5.1.3b) \end{cases}$$

where $(\cdot,\cdot)$ denotes the inner product of $L^2(\Omega)$, and

$$a(u,v)=\int_\Omega\Big(\sum_{ij}a_{ij}\frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_i}+\sum_j b_j\frac{\partial u}{\partial x_j}v+cuv\Big)\mathrm{d}x. \tag{5.1.4}$$

The solution to (5.1.3) is called the generalized solution of (5.1.1).

As in Chapters 2 and 3, we place a quasi-uniform grid and a corresponding dual grid on $\overline{\Omega}$, and construct a trial function space $U_h \in H_0^1(\Omega)$ and a test function space $V_h \in L^2(\Omega)$. Then the semi-discrete generalized difference scheme for (5.1.1) is: Find $u_h = u_h(\cdot,t) \in U_h$ $(0 \leq t \leq T)$ such that

$$\begin{cases} (u_{h,t},v_h)+a(u_h,v_h)=(f,v_h), & \forall v_h \in V_h, t>0, \quad (5.1.5a) \\ u_h(x,0)=u_{0h}(x), & x\in\Omega, \quad (5.1.5b) \end{cases}$$

where $a(u_h,v_h)$ is a bilinear form obtained by applying piecewise Green's formula to $(Au,v)$; or by interpreting the right-hand side of (5.1.4) *in the sense of generalized functions, namely*, the *integrals* are computed in terms of a $\delta$-function method on the boundaries of neighbouring dual elements. $u_{0h}$ is a certain approximation of $u_0$ on $U_h$. A commonly used method is to choose $u_{0h}$ as an interpolation projection of $u_0$ in $U_h$. Another way is to replace (5.1.5b) by

$$(u_h(\cdot,0),v_h)=(u_0,v_h), \quad \forall v_h \in V_h.$$

Let $\{\phi_j(x) : j = 1, 2, \cdots, n\}$ and $\{\psi_j(x) : j = 1, 2, \cdots, n\}$ be the bases of $U_h$ and $V_h$ respectively. Then (5.1.5) can be expressed as: Find a solution of the form

$$u_h = \sum_{j=1}^{n} \mu_j(t)\phi_j(x),$$

such that its coefficients $\mu_1(t), \mu_2(t), \cdots, \mu_n(t)$ satisfy

$$\begin{cases} \sum_{j=1}^{n}\Big[\dfrac{\mathrm{d}\mu_j(t)}{\mathrm{d}t}(\phi_j, \psi_i) + \mu_j(t)a(\phi_j, \phi_i)\Big] = (f, \psi_i), \ t > 0, & \\ \qquad i = 1, 2, \cdots, n, & (5.1.5\text{a})' \\ \mu_j(0) = \alpha_j, \ \ j = 1, 2, \cdots, n, & (5.1.5\text{b})' \end{cases}$$

where $\alpha_j$'s are the coefficients in $u_{0h} = \sum_{j=1}^{n} \alpha_j\phi_j$.

Let us introduce the following matrix and vector notations:

$$M = [m_{ij}] = [(\phi_j, \psi_i)], \ K = [k_{ij}] = [a(\phi_j, \psi_i)],$$

$$\mathbf{u} = [\mu_1(t), \cdots, \mu_n(t)]^T, \ F = [(f, \psi_1), \cdots, (f, \psi_n)]^T,$$

$$\alpha = [\alpha_1, \cdots, \alpha_n]^T.$$

Then we can rewrite (5.1.5)$'$ as

$$\begin{cases} M\dfrac{\mathrm{d}\mathbf{u}}{\mathrm{d}t} + K\mathbf{u} = F, & (5.1.5\text{a})'' \\ \mathbf{u}(0) = \alpha. & (5.1.5\text{b})'' \end{cases}$$

As in the finite element method, we call $M$ a mass matrix, and $K$ a stiff matrix. $M$ is clearly nonsingular. The ordinary differential equation theory tells us that this semi-discrete generalized difference scheme has a unique solution for any $f \in L^2(\Omega)$.

We are mainly concerned in this chapter with two-dimensional problems. So we always assume that $\Omega$ is a planar polygonal region, and that $A$ is a second order elliptic differential operator:

$$Au \equiv -\Big[\frac{\partial}{\partial x}\Big(a_{11}\frac{\partial u}{\partial x} + a_{12}\frac{\partial u}{\partial y}\Big) + \frac{\partial}{\partial y}\Big(a_{21}\frac{\partial u}{\partial x} + a_{22}\frac{\partial u}{\partial y}\Big)\Big] + qu.$$

We also assume $a_{ij}(x,y)$'s $(i,j = 1,2)$ and $q(x,y)$ are sufficiently smooth and positive definite: There exists a constant $r > 0$ such that

$$\sum_{i,j=1}^{2} a_{ij}(x,y)\xi_i\xi_j \geq r\sum_{i=1}^{2}\xi_i^2,\ q(x,y) \geq 0,$$

$$\forall(\xi_i,\xi_j) \in \mathrm{R}^2,\ (x,y) \in \overline{\Omega}.$$

In this and the next two sections, we always assume that $U_h$ is a piecewise linear function space corresponding to a grid $T_h$ of $\overline{\Omega}$, and that $V_h$ is a piecewise constant function space with respect to the barycenter dual grid $T_h^*$. (cf. §3.2 for details.)

### 5.1.2 Some lemmas

First let us restate some results of the last two chapters.

**Lemma 5.1.1** *Set*

$$\begin{aligned} \|u_h\|_{0,h} &= \|\Pi_h^* u_h\|_0 = \Big\{\sum_{K_{P_0}^* \in T_h^*} u_h^2(P_0)S_{P_0}^*\Big\}^{1/2} \\ &= \Big\{\frac{1}{3}\sum_{K_Q\in T_h}[u_h^2(P_i)+u_h^2(P_j)+u_h^2(P_k)]S_Q\Big\}^{1/2}, \end{aligned} \tag{5.1.6}$$

$$|u_h|_{1,h} = \Big\{\sum_{K_Q\in T_h}\Big[\Big(\frac{\partial u_h(Q)}{\partial x}\Big)^2 + \Big(\frac{\partial u_h(Q)}{\partial y}\Big)^2\Big]S_Q\Big\}^{1/2}, \tag{5.1.7}$$

$$\|u_h\|_{1,h} = \{\|u\|_{0,h}^2 + |u_h|_{1,h}^2\}^{1/2}. \tag{5.1.8}$$

*Then on $U_h$ the following pairs of norms are equivalent respectively: $|\cdot|_{1,h}$ and $|\cdot|_1$; $\|\cdot\|_{0,h}$ and $\|\cdot\|_0$; and $\|\cdot\|_{1,h}$ and $\|\cdot\|_1$.*

**Lemma 5.1.2** *The bilinear form $a(u_h,\Pi_h^*\bar{u}_h)$ can be expressed as*

$$a(u_h,\Pi_h^*\bar{u}_h) = a_h(u_h,\Pi_h^*\bar{u}_h) + b_h(u_h,\Pi_h^*\bar{u}_h), \tag{5.1.9}$$

*where the leading term*

$$
\begin{aligned}
& a_h(u_h, \Pi_h^* \bar{u}_h) \\
= & \sum_{K_Q \in T_h} \Big\{ \Big[ a_{11}(Q) \frac{\partial u_h(Q)}{\partial x} + a_{12}(Q) \frac{\partial u_h(Q)}{\partial y} \Big] \frac{\partial \bar{u}_h(Q)}{\partial x} \\
& + \Big[ a_{21}(Q) \frac{\partial u_h(Q)}{\partial x} + a_{22}(Q) \frac{\partial u_h(Q)}{\partial y} \Big] \frac{\partial \bar{u}_h(Q)}{\partial y} \Big\} S_Q \\
& + \sum_{K_{P_0}^* \in T_h^*} q(P_0) u_h(P_0) \bar{u}_h(P_0) S_{P_0}^*
\end{aligned}
\tag{5.1.10}
$$

*is symmetric and positive definite ($c_1$ and $c_2$ are positive constants):*

$$a_h(u_h, \Pi_h^* \bar{u}_h) = a_h(\bar{u}_h, \Pi_h^* u_h), \quad \forall \bar{u}_h, u_h \in U_h, \tag{5.1.11}$$

$$c_1 \|u_h\|_1^2 \le a_h(u_h, \Pi_h^* u_h) \le c_1 \|u_h\|_1^2, \quad \forall u_h \in U_h, \tag{5.1.12}$$

*and the remainder $b_h(u_h, \Pi_h^* \bar{u}_h)$ satisfies*

$$|b_h(u_h, \Pi_h^* \bar{u}_h)| \le Ch \|u_h\|_1 \|\bar{u}_h\|_1, \quad \forall \bar{u}_h, u_h \in U_h. \tag{5.1.13}$$

*If we define $|||u_h|||_1 = [a_h(u_h, \Pi_h^* u_h)]^{1/2}$, then $|||\cdot|||_1$ and $\|\cdot\|_1$ are equivalent on $U_h$ (cf. (5.1.12)). We also have (see (5.1.9), (5.1.11) and (5.1.13))*

$$
\begin{aligned}
& |a(u_h, \Pi_h^* \bar{u}_h) - a(\bar{u}_h, \Pi_h^* u_h)| \\
\le \quad & Ch \|u_h\|_1 \|\bar{u}_h\|_1, \quad \forall \bar{u}_h, u_h \in U_h.
\end{aligned}
\tag{5.1.14}
$$

**Lemma 5.1.3** *There exist positive constants $h_0, \alpha$ and $M$ such that when $0 < h \le h_0$*

$$a(u_h, \Pi_h^* u_h) \ge \alpha \|u_h\|_1^2, \quad \forall u_h \in U_h, \tag{5.1.15}$$

$$|a(u_h, \Pi_h^* \bar{u}_h)| \le M \|u_h\|_1 \|\bar{u}_h\|_1, \quad \forall \bar{u}_h, u_h \in U_h. \tag{5.1.16}$$

**Lemma 5.1.4** *Let $u \in H_0^1(\Omega)$ be the solution to the variational problem*

$$a(u, v) = (f, v), \quad \forall v \in H_0^1(\Omega)$$

*and* $u_h \in U_h$ *to the generalized difference scheme*

$$a(u_h, v_h) = (f, v_h), \quad \forall v_h \in V_h,$$

*then we have*

$$\|u - u_h\|_1 \le Ch|u|_2, \tag{5.1.17}$$

$$\|u - u_h\|_0 \le Ch^2\|u\|_{3,p}. \quad (p > 1) \tag{5.1.18}$$

**Lemma 5.1.5** *There hold the following statements:*

(i) $(u_h, \Pi_h^* \bar{u}_h) = (\bar{u}_h, \Pi_h^* u_h), \quad \forall \bar{u}_h, u_h \in U_h.$ (5.1.19)

(ii) *Set* $|||u_h|||_0 = (u_h, \Pi_h^* u_h)^{1/2}$. *Then* $|||\cdot|||_0$ *is equivalent to* $\|\cdot\|_0$ *on* $U_h$*, that is, there exist positive constants* $c_3$ *and* $c_4$ *such that*

$$c_3\|u_h\|_0 \le |||u_h|||_0 \le c_4\|u_h\|_0, \quad \forall u_h \in U_h. \tag{5.1.20}$$

The above results can be found in §3.2 and §4.1.

Let us introduce an elliptic projection operator

$$P_h : \; H^2(\Omega) \cap H_0^1(\Omega) \to U_h,$$

defined by the following generalized difference scheme:

$$a(P_h u, v_h) = a(u, v_h), \quad \forall v_h \in V_h. \tag{5.1.21}$$

By Lemma 5.1.3 we see that $P_h u$ is uniquely defined by (5.1.21) for any $u \in H^2(\Omega) \cap H_0^1(\Omega)$. We call $P_h u$ the elliptic projection of $u$ (with respect to the generalized difference scheme). By Lemma 5.1.4 we have the following estimate.

**Lemma 5.1.6** *Let* $P_h u$ *be the elliptic projection of* $u$ *defined by (5.1.21), then*

$$\|u - P_h u\|_1 \le Ch|u|_2, \tag{5.1.22}$$

$$\|u - P_h u\|_0 \le Ch^2\|u\|_{3,p}. \quad (p > 1) \tag{5.1.23}$$

### 5.1.3 $L^2$-error estimate

**Theorem 5.1.1** *Let $u$ and $u_h$ be the solutions to the problem (5.1.3) and the semi-discrete generalized difference scheme (5.1.5), respectively. Then we have*

$$\begin{aligned}&\|u-u_0\|_0\\ \le\ & C\Big\{\|u_0-u_{0h}\|_0+h^2\Big[\|u_0\|_{3,p}+\int_0^t\|u_\tau\|_{3,p}\mathrm{d}\tau\Big]\Big\}.\ (p>1)\end{aligned} \tag{5.1.24}$$

**Proof** Write

$$\rho=u-P_hu,\ e=P_hu-u_h, \tag{5.1.25}$$

where $P_h$ is the elliptic projection operator. Then we have

$$u-u_h=\rho+e. \tag{5.1.26}$$

It follows from (5.1.23) that

$$\begin{aligned}\|\rho\|_0\ &\le Ch^2\|u\|_{3,p}=Ch^2\|u_0+\int_0^t u_\tau\mathrm{d}\tau\|_{3,p}\\ &\le Ch^2\Big[\|u_0\|_{3,p}+\int_0^t\|u_\tau\|_{3,p}\mathrm{d}\tau\Big].\end{aligned} \tag{5.1.27}$$

We turn to estimate $e$. Since $u$ and $u_h$ satisfy (5.1.3) and (5.1.5) respectively, we have

$$(u_t-u_{h,t},v_h)+a(u-u_h,v_h)=0,\ \ \forall v_h\in V_h. \tag{5.1.28}$$

This together with (5.1.21) gives

$$(e_t,v_h)+a(e,v_h)=-(\rho_t,v_h),\ \ \forall v_h\in V_h. \tag{5.1.29}$$

Choosing $v_h=\Pi_h^*e$ and using (5.1.19) and (5.1.15) yield

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|||e|||_0^2\le\|\rho_t\|_0\|\Pi_h^*e\|_0.$$

So it follows from Lemmas 5.1.1 and 5.1.5 that

$$\frac{\mathrm{d}}{\mathrm{d}t}|||e|||_0\le C\|\rho_t\|_0.$$

Integrate it with respect to $t$ and note the equivalence of the norms $|||\cdot|||_0$ and $\|\cdot\|_0$, then we have

$$\|e\|_0 \le C\Big[\|e(0)\|_0 + \int_0^t \|\rho_\tau\|_0 \mathrm{d}\tau\Big]. \tag{5.1.30}$$

By virtue of Lemma 5.1.6 we have

$$\begin{aligned}\|e(0)\|_0 &\le \|P_h u_0 - u_0\|_0 + \|u_0 - u_{0h}\|_0 \\ &\le Ch^2\|u_0\|_{3,p} + \|u_0 - u_{0h}\|_0,\end{aligned} \tag{5.1.31}$$

$$\|\rho_\tau\|_0 = \|u_\tau - P_h u_\tau\|_0 \le Ch^2\|u_\tau\|_{3,p}. \tag{5.1.32}$$

A combination of (5.1.27) and (5.1.30)-(5.1.32) leads to (5.1.24). This completes the proof. □

### 5.1.4 $H^1$-error estimate

**Theorem 5.1.2** *Let $u$ and $u_h$ be the solutions to the problem (5.1.3) and the semi-discrete generalized difference scheme (5.1.5), respectively. Then we have*

$$\begin{aligned}&\|u - u_h\|_1 \\ \le\ & C\Big\{\|u_0 - u_{0h}\|_1 + h\Big[\|u_0\|_2 + \int_0^t \|u_\tau\|_2 \mathrm{d}\tau + \Big(\int_0^t \|u_\tau\|_2^2 \mathrm{d}\tau\Big)^{1/2}\Big]\Big\}.\end{aligned} \tag{5.1.33}$$

**Proof** Take $v_h = \Pi_h^* e_t$ in (5.1.29) to get

$$|||e_t|||_0^2 + a(e, \Pi_h^* e_t) = -(\rho_t, \Pi_h^* e_t),$$

$$\begin{aligned}&|||e_t|||_0^2 + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t} a(e, \Pi_h^* e) \\ =\ & -(\rho_t, \Pi_h^* e_t) + \frac{1}{2}[a(e_t, \Pi_h^* e) - a(e, \Pi_h^* e_t)].\end{aligned} \tag{5.1.34}$$

It follows from (5.1.14) and the inverse property of the finite element space that

$$\begin{aligned}&|a(e_t,\Pi_h^* e)-a(e,\Pi_h^* e_t)|\le Ch\|e_t\|_1\|e\|_1\\ \le\;& C\|e_t\|_0\|e\|_1\le \||e_t\||_0^2+C'\|e\|_1^2.\end{aligned}$$

Thus

$$\frac{\mathrm{d}}{\mathrm{d}t}a(e,\Pi_h^* e)\le C(\|\rho_t\|_0^2+\|e\|_1^2).$$

Integrate on $t$ and use Lemma 5.1.3 to obtain

$$\begin{aligned}\alpha\|e\|_1^2 &\le a(e,\Pi_h^* e)\\ &\le a(e(0),\Pi_h^* e(0))+C\int_0^t(\|\rho_\tau\|_0^2+\|e\|_1^2)\mathrm{d}\tau\\ &\le M\|e(0)\|_1^2+C\int_0^t(\|\rho_\tau\|_0^2+\|e\|_1^2)\mathrm{d}\tau.\end{aligned}$$

Note

$$\begin{aligned}\|e(0)\|_1 &\le \|P_h u_0-u_0\|_1+\|u_0-u_{0h}\|_1\\ &\le Ch\|u_0\|_2+\|u_0-u_{0h}\|_1.\end{aligned}$$

Hence

$$\|e\|_1^2\le C\Big\{\|u_0-u_{0h}\|_1^2+h^2\|u_0\|_2^2+\int_0^t(\|\rho_\tau\|_0^2+\|e\|_1^2)\mathrm{d}\tau\Big\}.$$

By the well-known Gronwall inequality we have

$$\|e\|_1^2\le C\Big\{\|u_0-u_{0h}\|_1^2+h^2\|u_0\|_2^2+\int_0^t\|\rho_\tau\|_0^2\mathrm{d}\tau\Big\}. \tag{5.1.35}$$

By virtue of Lemma 5.1.6 we have

$$\begin{aligned}\|\rho\|_1 &=\|u-P_h u\|_1\le Ch\|u\|_2\\ &\le Ch(\|u_0\|_2+\int_0^t\|u_\tau\|_2\mathrm{d}\tau),\end{aligned}\tag{5.1.36}$$

$$\|\rho_\tau\|_0=\|u_\tau-P_h u_\tau\|_0\le Ch\|u_\tau\|_2. \tag{5.1.37}$$

Inserting (5.1.37) into (5.1.35) yields

$$\|e\|_1 \leq \; C\Big\{\|u_0 - u_{0h}\|_1 + h\|u_0\|_2 + h\Big(\int_0^t \|u_\tau\|_2^2 \mathrm{d}\tau\Big)^{1/2}\Big\}. \tag{5.1.38}$$

A combination of (5.1.36) and (5.1.38) leads to (5.1.33). This completes the proof. □

## 5.2 Fully-discrete Generalized Difference Schemes

### 5.2.1 Fully-discrete schemes

In the last section the semi-discrete schemes are obtained by discretizing the space variable. In order to finally get numerical solutions we need to further discretize the time variable to obtain fully-discrete schemes. To this end, there are two methods most in use: the implicit Euler's scheme (backward differencing) and the Crank-Nicolson scheme (central differencing).

Let $\tau$ denote the time step size, and $t_n = n\tau$ $(n = 0, 1, \cdots)$, $u_h^n = u_h(t_n)$. At time $t = t_n$, if we use the backward difference quotient

$$\bar{\partial}_t u_h^n = (u_h^n - u_h^{n-1})/\tau$$

to approximate the differential quotient $u_{h,t}$ in the semi-discrete scheme, then we obtain a fully-discrete scheme: Find $u_h^n \in U_h$ $(n = 1, 2, \cdots)$ such that

$$\begin{cases} (\bar{\partial}_t u_h^n, v_h) + a(u_h^n, v_h) = (f(t_n), v_h), \quad \forall v_h \in V_h, & \text{(5.2.1a)} \\ \quad n = 1, 2, \cdots, \\ u_h^0 = u_{0h}. & \text{(5.2.1b)} \end{cases}$$

Or we can equivalently write it as

$$\begin{cases} (u_h^n, v_h) + \tau a(u_h^n, v_h) = (u_h^{n-1} + \tau f(t_n), v_h), \quad \forall v_h \in V_h, \\ \quad n = 1, 2, \cdots, \\ u_h^0 = u_{0h}. \end{cases}$$

This scheme is referred to as a backward Euler generalized difference scheme.

By (5.1.15)

$$\begin{aligned} &a(u_h^n, \Pi_h^* u_h^n) + \frac{1}{\tau}(u_h^n, \Pi_h^* u_h^n) \\ &\geq \alpha \|u_h^n\|_1^2, \quad \forall u_h^n \in U_h. \end{aligned}$$

This guarantees the existence and uniqueness of the solution $u_h^n$ to (5.2.1a) for a given $u_h^{n-1}$.

If we discretize the semi-discrete scheme at time $t_{n-1/2} = \left(n - \frac{1}{2}\right)\tau$ in a symmetric fashion, then we have another fully-discrete scheme as follows: Find $u_h^n \in U_h$ $(n = 1, 2, \cdots)$ such that

$$\begin{cases} (\bar{\partial}_t u_h^n, v_h) + a\Big(\dfrac{u_h^n + u_h^{n-1}}{2}, v_h\Big) = \\ \qquad \Big(\dfrac{f(t_n) + f(t_{n-1})}{2}, v_h\Big), \quad \forall v_h \in V_h, & (5.2.2a) \\ \qquad n = 1, 2, \cdots, \\ u_h^0 = u_{0h}. & (5.2.2b) \end{cases}$$

This scheme is the so-called Crank-Nicolson generalized difference scheme. The existence and uniqueness of its solution can be readily proved similarly as above.

### 5.2.2 Error estimates for backward Euler generalized difference schemes

**Theorem 5.2.1** *Let $u$ and $\{u_h^n\}$ be the solutions to the parabolic equation (5.1.3) and the backward Euler generalized difference scheme*

*(5.2.1), respectively. Then*

$$\begin{aligned}&\|u(t_n) - u_h^n\|_0\\ \le\ & C\Big\{\|u_0 - u_{0h}\|_0 + h^2\Big[\|u_0\|_{3,p} + \int_0^{t_n} \|u_t\|_{3,p}\mathrm{d}t\Big]\\ &+\tau\int_0^{t_n} \|u_{tt}\|_0\mathrm{d}t\Big\},\quad n = 1, 2, \cdots.\ (p > 1)\end{aligned} \tag{5.2.3}$$

**Proof** Set

$$\rho^n = u(t_n) - P_h u(t_n),\ e^n = P_h u(t_n) - u_h^n,$$

then

$$u(t_n) - u_h^n = \rho^n + e^n. \tag{5.2.4}$$

It follows from (5.1.23) that

$$\|\rho^n\|_0 \le Ch^2\|u(t_n)\|_{3,p} \le Ch^2\Big[\|u_0\|_{3,p} + \int_0^{t_n} \|u_t\|_{3,p}\mathrm{d}t\Big]. \tag{5.2.5}$$

Set $t = t_n$ in (5.1.3), and subtract it and (5.2.1), then we have

$$(u_t(t_n) - \bar{\partial}_t u_h^n, v_h) + a(\rho^n + e^n, v_h) = 0,\ \ \forall v_h \in V_h. \tag{5.2.6}$$

By virtue of (5.1.21) we have

$$\begin{aligned}&(\bar{\partial}_t e^n, v_h) + a(e^n, v_h)\\ =\ & (\bar{\partial}_t P_h u(t_n) - u_t(t_n), v_h),\ \ \forall v_h \in V_h.\end{aligned} \tag{5.2.7}$$

Write $r^n = \bar{\partial}_t P_h u(t_n) - u_t(t_n)$, set $v_h = \Pi_h^* e^n$, and use (5.1.15), then we have

$$(\bar{\partial}_t e^n, \Pi_h^* e^n) \le (r^n, \Pi_h^* e^n).$$

So

$$|||e^n|||_0^2 \le (e^{n-1}, \Pi_h^* e^n) + \tau(r^n, \Pi_h^* e^n),$$
$$|||e^n|||_0^2 \le (|||e^{n-1}|||_0 + \tau|||r^n|||_0)|||e^n|||_0.$$

Eliminating $|||e^n|||_0$ and using the above recursion relation, we have

$$|||e^n|||_0 \le |||e^0|||_0 + \tau\sum_{j=1}^{n} |||r^j|||_0.$$

Making use of the equivalence of the norms we get

$$\|e^n\|_0 \le C\Big(\|e^0\|_0 + \tau \sum_{j=1}^{n} \|r^j\|_0\Big). \tag{5.2.8}$$

Write $r^j = r_1^j + r_2^j$, where

$$r_1^j = \bar{\partial}_t P_h u(t_j) - \bar{\partial}_t u(t_j) = \frac{1}{\tau}\int_{t_{j-1}}^{t_j} (P_h - I)u_t \mathrm{d}t,$$

$$r_2^j = \bar{\partial}_t u(t_j) - u_t(t_j) = -\frac{1}{\tau}\int_{t_{j-1}}^{t_j} (t - t_{j-1})u_{tt}\mathrm{d}t.$$

Then by (5.1.23)

$$\begin{aligned}\sum_{j=1}^{n} \|r_1^j\|_0 &\le \frac{1}{\tau}\sum_{j=1}^{n}\int_{t_{j-1}}^{t_j} Ch^2 \|u_t\|_{3,p}\mathrm{d}t \\ &= C\tau^{-1}h^2 \int_0^{t_n} \|u_t\|_{3,p}\mathrm{d}t.\end{aligned} \tag{5.2.9}$$

Similarly

$$\sum_{j=1}^{n} \|r_2^j\|_0 \le \sum_{j=1}^{n}\int_{t_{j-1}}^{t_j} \|u_{tt}\|_0 \mathrm{d}t = \int_0^{t_n} \|u_{tt}\|_0\mathrm{d}t. \tag{5.2.10}$$

Again by (5.1.23)

$$\begin{aligned}\|e^0\|_0 &\le \|P_h u_0 - u_0\|_0 + \|u_0 - u_{0h}\|_0 \\ &\le Ch^2\|u_0\|_{3,p} + \|u_0 - u_{0h}\|_0.\end{aligned} \tag{5.2.11}$$

Substituting (5.2.9)-(5.2.11) into (5.2.8) yields

$$\begin{aligned}\|e^n\|_0 \le \quad & C\{\|u_0 - u_{0h}\|_0 + h^2[\|u_0\|_{3,p} \\ & + \int_0^{t_n} \|u_t\|_{3,p}\mathrm{d}t] + \tau\int_0^{t_n} \|u_{tt}\|_0\mathrm{d}t\}.\end{aligned} \tag{5.2.12}$$

Finally (5.2.3) follows from (5.2.5) and (5.2.12). This completes the proof. □

Next we deal with the $H^1$-estimate.

**Theorem 5.2.2** *Let $u$ and $\{u_h^n\}$ be the solutions to the parabolic equation (5.1.3) and the backward Euler generalized difference scheme respectively. Then*

$$\begin{aligned}
&\|u(t_n) - u_h^n\|_1 \\
\le\ & C\Big\{\|u_0 - u_{0h}\|_1 + h\Big[\|u_0\|_2 + \int_0^{t_n} \|u_t\|_2 dt \\
&+\Big(\int_0^{t_n} \|u_t\|_2^2 dt\Big)^{1/2}\Big] + \tau\Big(\int_0^{t_n} \|u_{tt}\|_0^2 dt\Big)^{1/2}\Big\}, \\
&n = 1, 2, \cdots.
\end{aligned} \tag{5.2.13}$$

**Proof** As in the proof to Theorem 5.2.1, we can again obtain (5.2.7). To proceed, we set $v_h = \Pi_h^* \bar{\partial}_t e^n$ to get

$$|||\bar{\partial}_t e^n|||_0^2 + a(e^n, \Pi_h^* \bar{\partial}_t e^n) = (r^n, \Pi_h^* \bar{\partial}_t e^n). \tag{5.2.14}$$

By the equivalence of the norms we have a constant $C_0 > 0$ satisfying

$$|||\bar{\partial}_t e^n|||_0^2 \ge C_0 \|\bar{\partial}_t e^n\|_0^2. \tag{5.2.15}$$

It follows from Lemmas 5.1.2 and 5.1.3 that

$$\begin{aligned}
&a(e^n, \Pi_h^* \bar{\partial}_t e^n) \\
=\ & \frac{1}{2\tau}[a(e^n + e^{n-1}, \Pi_h^*(e^n - e^{n-1})) \\
&+ a(e^n - e^{n-1}, \Pi_h^*(e^n - e^{n-1}))] \\
\ge\ & \frac{1}{2\tau}[a_h(e^n + e^{n-1}, \Pi_h^*(e^n - e^{n-1})) \\
&+ b_h(e^n + e^{n-1}, \Pi_h^*(e^n - e^{n-1}))] \\
\ge\ & \frac{1}{2\tau}[|||e^n|||_1^2 - |||e^{n-1}|||_1^2] \\
&- C\|e^n + e^{n-1}\|_1 \|\bar{\partial}_t e^n\|_0 \\
\ge\ & \frac{1}{2\tau}[(1 - C'\tau)|||e^n|||_1^2 - (1 + C'\tau)|||e^{n-1}|||_1^2] \\
&- \frac{C_0}{2}\|\bar{\partial}_t e^n\|_0^2.
\end{aligned} \tag{5.2.16}$$

Also note

$$|(r^n, \Pi_h^* \bar{\partial}_t e^n)| \leq C\|r^n\|_0^2 + \frac{C_0}{2}\|\bar{\partial}_t e^n\|_0^2. \tag{5.2.17}$$

Combining (5.2.14)-(5.2.17) gives

$$|||e^n|||_1^2 \leq \frac{1 + C'\tau}{1 - C'\tau}|||e^{n-1}|||_1^2 + C\tau\|r^n\|_0^2. \tag{5.2.18}$$

This recursion relation leads to

$$|||e^n|||_1^2 \leq C\Big(|||e^0|||_1^2 + \tau\sum_{j=1}^n \|r^j\|_0^2\Big). \tag{5.2.19}$$

Note

$$r^j = r_1^j + r_2^j,$$

$$r_1^j = \bar{\partial}_t P_h u(t_j) - \bar{\partial}_t u(t_j) = \frac{1}{\tau}\int_{t_{j-1}}^{t_j} (P_h u_t - u_t)\mathrm{d}t,$$

$$r_2^j = \bar{\partial}_t u(t_j) - u_t(t_j) = -\frac{1}{\tau}\int_{t_{j-1}}^{t_j} (t - t_{j-1})u_{tt}\mathrm{d}t.$$

So we have

$$\begin{aligned} \sum_{j=1}^n \|r_1^j\|_0^2 &\leq C\tau^{-2}\sum_{j=1}^n\Big(\int_{t_{j-1}}^{t_j} h\|u_t\|_2 \mathrm{d}t\Big)^2 \\ &\leq C\tau^{-1}h^2\int_0^{t_n}\|u_t\|_2^2\mathrm{d}t, \end{aligned} \tag{5.2.20}$$

$$\begin{aligned} \sum_{j=1}^n \|r_2^j\|_0^2 &\leq \sum_{j=1}^n\Big(\int_{t_{j-1}}^{t_j} \|u_{tt}\|_0 \mathrm{d}t\Big)^2 \\ &\leq \tau\int_0^{t_n}\|u_{tt}\|_0^2\mathrm{d}t. \end{aligned} \tag{5.2.21}$$

Also notice

$$\begin{aligned} \|e^0\|_1^2 &\leq \|P_h u_0 - u_0\|_1^2 + \|u_0 - u_{0h}\|_1^2 \\ &\leq Ch^2\|u_0\|_2^2 + \|u_0 - u_{0h}\|_1^2. \end{aligned} \tag{5.2.22}$$

A combination of (5.2.19)-(5.2.22) yields

$$\begin{aligned}\|e^n\|_1 \le\ & C\Big[\|u_0-u_{0h}\|_1+h\|u_0\|_2+h\Big(\int_0^{t_n}\|u_t\|_2^2\mathrm{d}t\Big)^{1/2}\\ & +\tau\Big(\int_0^{t_n}\|u_{tt}\|_0^2\mathrm{d}t\Big)^{1/2}\Big].\end{aligned} \tag{5.2.23}$$

On the other hand,

$$\begin{aligned}\|\rho^n\|_1\ &=\|u(t_n)-P_hu(t_n)\|_1\le Ch\|u(t_n)\|_2\\ &\le Ch\Big[\|u_0\|_2+\int_0^{t_n}\|u_t\|_2\mathrm{d}t\Big].\end{aligned} \tag{5.2.24}$$

Finally, (5.2.13) results from (5.2.23) and (5.2.24). This completes the proof. □

### 5.2.3 Error estimates for Crank-Nicolson generalized difference schemes

**Theorem 5.2.3** *Let $u$ and $\{u_h^n\}$ be the solutions to the parabolic problem (5.1.3) and the Crank-Nicolson generalized difference scheme (5.2.2), respectively, then*

$$\begin{aligned}&\|u(t_n)-u_h^n\|_0\\ \le\ & C\Big\{\|u_0-u_{0h}\|_0+h^2\Big[\|u_0\|_{3,p}+\int_0^{t_n}\|u_t\|_{3,p}\mathrm{d}t\Big]\\ & +\tau^2\int_0^{t_n}\|u_{ttt}\|_0\mathrm{d}t\Big\},\quad n=1,2,\cdots.\ (p>1)\end{aligned} \tag{5.2.25}$$

**Proof** As before we set

$$u(t_n)-u_h^n=\rho^n+e^n, \tag{5.2.26}$$

where

$$\rho^n=u(t_n)-P_hu(t_n),\ e^n=P_hu(t_n)-u_h^n.$$

For $\rho^n$ we have

$$\begin{aligned}\|\rho^n\|_0\ &\le Ch^2\|u(t_n)\|_{3,p}\\ &\le Ch^2\Big[\|u_0\|_{3,p}+\int_0^{t_n}\|u_t\|_{3,p}\mathrm{d}t\Big].\end{aligned} \tag{5.2.27}$$

On the other hand, by (5.1.3), (5.1.21) and (5.2.2), $e^n$ satisfies

$$(\bar{\partial}_t e^n, v_h) + a\Big(\frac{e^n + e^{n-1}}{2}, v_h\Big) = (r^n, v_h); \quad \forall v_h \in V_h, \tag{5.2.28}$$

where

$$r^n = \bar{\partial}_t P_h u(t_n) - \frac{u_t(t_n) + u_t(t_{n-1})}{2}.$$

Take $v_h = \Pi_h^* \frac{e^n + e^{n-1}}{2}$ in (5.2.28) to get

$$\Big(\bar{\partial}_t e^n, \Pi_h^* \frac{e^n + e^{n-1}}{2}\Big) \le \Big(r^n, \Pi_h^* \frac{e^n + e^{n-1}}{2}\Big).$$

By virtue of Lemma 5.1.5 we have

$$\frac{1}{2\tau}(|||e^n|||_0^2 - |||e^{n-1}|||_0^2) \le \frac{1}{2}|||r^n|||_0(|||e^n|||_0 + |||e^{n-1}|||_0).$$

Thus,

$$|||e^n|||_0 \le |||e^{n-1}|||_0 + C\tau\|r^n\|_0.$$

By this recursion relation we have

$$|||e^n|||_0 \le |||e^0|||_0 + C\tau \sum_{j=1}^{n} \|r^j\|_0,$$

and hence

$$\|e^n\|_0 \le C\Big(\|e^0\|_0 + \tau \sum_{j=1}^{n} \|r^j\|_0\Big). \tag{5.2.29}$$

Write

$$r^j = r_1^j + r_2^j,$$

$$r_1^j = \bar{\partial}_t P_h u(t_j) - \bar{\partial}_t u(t_j) = \frac{1}{\tau}\int_{t_{j-1}}^{t_j} (P_h u_t - u_t)\mathrm{d}t,$$

$$r_2^j = \bar{\partial}_t u(t_j) - \frac{u_t(t_j) + u_t(t_{j-1})}{2}.$$

By Lemma 5.1.6

$$\sum_{j=1}^{n} \|r_1^j\|_0 \le C\tau^{-1}h^2 \int_0^{t_n} \|u_t\|_{3,p}\mathrm{d}t. \tag{5.2.30}$$

By the Taylor expansion we have

$$\sum_{j=1}^{n}\|r_2^j\|_0 \le C\tau\int_0^{t_n}\|u_{ttt}\|_0 \mathrm{d}t. \tag{5.2.31}$$

Also note

$$\begin{aligned}\|e^0\|_0 &\le \|P_h u_0 - u_0\|_0 + \|u_0 - u_{0h}\|_0 \\ &\le Ch^2\|u_0\|_{3,p} + \|u_0 - u_{0h}\|_0.\end{aligned} \tag{5.2.32}$$

This together with (5.2.29)-(5.2.31) gives

$$\begin{aligned}\|e^n\|_0 \le \ & C\Big\{\|u_0 - u_{0h}\|_0 + h^2\Big[\|u_0\|_{3,p} \\ &+ \int_0^{t_n}\|u_t\|_{3,p}\mathrm{d}t\Big] + \tau^2\int_0^{t_n}\|u_{ttt}\|_0\mathrm{d}t\Big\}.\end{aligned} \tag{5.2.33}$$

Now, (5.2.25) follows from (5.2.27) and (5.2.33), This completes the proof. □

**Theorem 5.2.4** *Let $u$ and $\{u_h^n\}$ be the solutions to the parabolic problem (5.1.3) and the Crank-Nicolson generalized difference scheme (5.2.2), respectively, then*

$$\begin{aligned}&\|u(t_n) - u_h^n\|_1 \\ \le \ & C\Big\{\|u_0 - u_{0h}\|_1 + h\Big[\|u_0\|_2 + \int_0^{t_n}\|u_t\|_2\mathrm{d}t \\ &+\Big(\int_0^{t_n}\|u_t\|_2^2\mathrm{d}t\Big)^{1/2}\Big] + \tau^2\Big(\int_0^{t_n}\|u_{ttt}\|_0^2\mathrm{d}t\Big)^{1/2}\Big\}, \\ &n = 1, 2, \cdots.\end{aligned} \tag{5.2.34}$$

**Proof** As in the proof of the last theorem, we have

$$(\bar{\partial}_t e^n, v_h) + a\Big(\frac{e^n + e^{n-1}}{2}, v_h\Big) = (r^n, v_h), \quad \forall v_h \in V_h. \tag{5.2.35}$$

Choosing $v_h = \Pi_h^*\bar{\partial}_t e^n$ leads to

$$|||\bar{\partial}_t e^n|||_0^2 + a\Big(\frac{e^n + e^{n-1}}{2}, \Pi_h^*\bar{\partial}_t e^n\Big) = (r^n, \Pi_h^*\bar{\partial}_t e^n). \tag{5.2.36}$$

As in the proof to Theorem 5.2.2, we use Lemma 5.2.2 and the inverse property of the finite element space $U_h$, and note the equivalence of the norms to obtain

$$\begin{aligned}
& a\Big(\frac{e^n+e^{n-1}}{2}, \Pi_h^*\bar{\partial}_t e^n\Big) \\
= \ & \frac{1}{2\tau}[a_h(e^n+e^{n-1}, \Pi_h^*(e^n-e^{n-1})) \\
& +b_h(e^n+e^{n-1}, \Pi_h^*(e^n-e^{n-1}))] \\
\geq \ & \frac{1}{2\tau}[|||e^n|||_1^2 - |||e^{n-1}|||_1^2] \\
& -C\|e^n+e^{n-1}\|_1\|\bar{\partial}_t e^n\|_0 \\
\geq \ & \frac{1}{2\tau}[(1-C'\tau)|||e^n|||_1^2-(1+C'\tau)|||e^{n-1}|||_1^2] \\
& -\frac{1}{2}\|\bar{\partial}_t e^n\|_0^2,
\end{aligned} \tag{5.2.37}$$

$$|(r^n, \Pi_h^*\bar{\partial}_t e^n)| \leq C\|r^n\|_0^2+\frac{1}{2}|||\bar{\partial}_t e^n|||_0^2. \tag{5.2.38}$$

It follows from (5.2.36)-(5.2.38) that

$$|||e^n|||_1^2 \leq \frac{1+C'\tau}{1-C'\tau}|||e^{n-1}|||_1^2+C\tau\|r^n\|_0^2.$$

This implies

$$|||e^n|||_1^2 \leq C\Big(|||e^0|||_1^2+\tau\sum_{j=1}^n\|r^j\|_0^2\Big). \tag{5.2.39}$$

Now, similar to (5.2.22) we have

$$\|e^0\|_1^2 \leq Ch^2\|u_0\|_2^2+\|u_0-u_{0h}\|_1^2. \tag{5.2.40}$$

As for $r^j = r_1^j + r_2^j$, we imitate (5.2.30) and (5.2.31) to get

$$\begin{aligned}
\sum_{j=1}^n\|r_1^j\|_0^2 & \leq C\tau^{-2}h^2\sum_{j=1}^n\Big(\int_{t_{j-1}}^{t_j}\|u_t\|_2\mathrm{d}t\Big)^2 \\
& \leq C\tau^{-1}h^2\int_0^{t_n}\|u_t\|_2^2\mathrm{d}t,
\end{aligned} \tag{5.2.41}$$

$$\begin{aligned}\sum_{j=1}^{n}\|r_2^j\|_0^2 &\le C\tau^2\sum_{j=1}^{n}\Big(\int_{t_{j-1}}^{t_j}\|u_{ttt}\|_0\mathrm{d}t\Big)^2\\ &\le C\tau^3\int_0^{t_n}\|u_{ttt}\|_0^2\mathrm{d}t.\end{aligned}\tag{5.2.42}$$

Combining (5.2.39)-(5.2.42) yields

$$\begin{aligned}\|e^n\|_1\le\ & C\Big\{\|u_0-u_{0h}\|_1+h\Big[\|u_0\|_2+\Big(\int_0^{t_n}\|u_t\|_2^2\mathrm{d}t\Big)^{1/2}\Big]\\ &+\tau^2\Big(\int_0^{t_n}\|u_{ttt}\|_0^2\mathrm{d}t\Big)^{1/2}\Big\}.\end{aligned}\tag{5.2.43}$$

Similar to (5.2.24) we have

$$\|\rho^n\|_1\le Ch\Big[\|u_0\|_2+\int_0^{t_n}\|u_t\|_2\mathrm{d}t\Big].\tag{5.2.44}$$

Finally, (5.2.34) results from (5.2.26), (5.2.43) and (5.2.44). This completes the proof. □

## 5.3 Mass Concentration Methods

This section is devoted to a variety of the generalized difference method, a mass concentration method, for parabolic equations. This method simplifies the computation and enjoys a satisfactory convergence.

### 5.3.1 Construction of schemes

Let us recall the semi-discrete generalized difference scheme (5.1.5):

$$\begin{cases}(u_{h,t},v_h)+a(u_h,v_h)=(f,v_h), & \forall v_h\in V_h, t>0, \quad (5.3.1\text{a})\\ u_h(x,0)=u_{0h}(x), & x\in\Omega. \quad (5.3.1\text{b})\end{cases}$$

Its equivalent matrix form is:

$$\begin{cases}M\dfrac{\partial\mathbf{u}}{\partial t}+K\mathbf{u}=F, & (5.3.1\text{a})'\\ \mathbf{u}(0)=\alpha, & (5.3.1\text{b})'\end{cases}$$

where $M$ is a mass matrix, and $K$ a stiff matrix.

The idea of the so-called mass concentration method is to concentrate all the entries on each row of the mass matrix $M = [m_{ij}]$ to the diagonal position, such that the inverse of $M$ is extremely easy to get and hence greatly simplifies the computation. To elaborate, the scheme of semi-discrete mass concentration method is:

$$\begin{cases} \overline{M}\dfrac{\partial \mathbf{u}}{\partial t} + K\mathbf{u} = F, & \text{(5.3.2a)} \\ \mathbf{u}(0) = \alpha. & \text{(5.3.2b)} \end{cases}$$

where $\overline{M} = [\bar{m}_{ij}]$ is a diagonal matrix

$$\bar{m}_{ij} = \begin{cases} 0, & \text{when } j \neq i, \\ \sum_{k=1}^{n} m_{ik}, & \text{when } j = i. \end{cases} \qquad (5.3.3)$$

Now, we deduce an equivalent form of the above scheme, which will be used later on for error estimates.

Define a semi-discrete problem: Find $u_h = \sum_{j=1}^{n} \mu_j(t)\phi_j \in U_h$ such that

$$\begin{cases} (\Pi_h^* u_{h,t}, v_h) + a(u_h, v_h) = (f, v_h), \ v_h \in V_h, \ t > 0, & \text{(5.3.4a)} \\ u_h(x, 0) = u_{0,h}(x), \ x \in \Omega. & \text{(5.3.4b)} \end{cases}$$

**Lemma 5.3.1** *Problems (5.3.2) and (5.3.4) are equivalent.*

**Proof** Write (5.3.4) into a matrix form:

$$\begin{cases} \tilde{M}\dfrac{\mathrm{d}\mathbf{u}}{\mathrm{d}t} + K\mathbf{u} = F, \\ \mathbf{u}(0) = \alpha. \end{cases}$$

Apparently $K, f$ and $\alpha$ here are identical to those in (5.3.2). It merely remains to show $\tilde{M} = \overline{M}$. The $ij$-entry of $\tilde{M}$ is

$$\tilde{m}_{ij} = (\Pi_h^* \phi_j, \psi_i) = (\psi_j, \psi_i) = \begin{cases} 0, & \text{when } j \neq i, \\ S_{P_i}^*, & \text{when } j = i. \end{cases}$$

Here $\psi_i$ is the characteristic function of the dual element $K_{P_i}^*$, and $S_{P_i}^*$ is the area of $K_{P_i}^*$. By (5.3.3)

$$\tilde{m}_{ij} = \begin{cases} 0, & \text{when } j \neq i, \\ \sum_{k=1}^{n} (\phi_k, \psi_i) = (1, \psi_i) = S_{P_i}^*, & \text{when } j = i. \end{cases}$$

Thus $\tilde{M} = \overline{M}$. This completes the proof. □

The fully-discrete mass concentration scheme is: Find $u_h^n \in U_h$ $(n = 1, 2, \cdots)$ such that

$$\begin{cases} (\Pi_h^* \bar{\partial}_t u_h^n, v_h) + a(\theta u_h^n + (1-\theta)u_h^{n-1}, v_h) \\ \quad = (\theta f(t_n) + (1-\theta) f(t_{n-1}), v_h), & (5.3.5\text{a}) \\ \qquad v_h \in V_h, \ n = 1, 2, \cdots, \\ u_h^0 = u_{0h}. & (5.3.5\text{b}) \end{cases}$$

Its matrix form is

$$\begin{cases} (\overline{M} + \theta\tau K)\mathbf{u}^n \\ \quad = [\overline{M} - (1-\theta)\tau K]\mathbf{u}^{n-1} + \tau[\theta f^n + (1-\theta) f^{n-1}], \\ \qquad n = 1, 2, \cdots, \\ \mathbf{u}^0 = \alpha. \end{cases}$$

This leads to a backward Euler fully-discrete scheme of the mass concentration method when $\theta = 1$, and a Crank-Nicolson scheme when $\theta = \frac{1}{2}$.

### 5.3.2 Error estimates for semi-discrete schemes

**Theorem 5.3.1** *Let $u$ and $u_h$ be the solutions to the problem (5.1.3) and the semi-discrete, mass concentration, generalized difference scheme (5.3.4), respectively. Then we have*

$$\|u - u_h\|_1 \leq C\Big\{\|u_0 - u_{0h}\|_1 + h\Big[\|u_0\|_2 + \int_0^t \|u_\tau\|_2 \mathrm{d}\tau + \Big(\int_0^t \|u_\tau\|_2^2 \mathrm{d}\tau\Big)^{1/2}\Big]\Big\}. \tag{5.3.6}$$

**Proof** As in §5.2, we write

$$u - u_h = \rho + e,\ \rho = u - P_h u,\ e = P_h u - u_h, \tag{5.3.7}$$

where $P_h$ is the elliptic projection operator. By (5.1.22)

$$\|\rho\|_1 \le Ch\|u\|_2 \le Ch\Big(\|u_0\|_2 + \int_0^t \|u_\tau\|_2 \mathrm{d}\tau\Big). \tag{5.3.8}$$

Since $u$ and $u_h$ satisfy (5.1.3) and (5.3.4) respectively, we have

$$(u_t - \Pi_h^* u_{h,t}, v_h) + a(u - u_h, v_h) = 0,\ v_h \in V_h. \tag{5.3.9}$$

This together with (5.1.21) gives

$$(\Pi_h^* e_t, v_h) + a(e, v_h) = -(r, v_h),$$

where

$$r = u_t - \Pi_h^* P_h u_t.$$

Set $v_h = \Pi_h^* e_t$ and use Lemma 5.1.2 to obtain

$$\|e_t\|_{0,h}^2 + a_h(e, \Pi_h^* e_t) = -b_h(e, \Pi_h^* e_t) - (r, \Pi_h^* e_t).$$

We have the following estimates for the above terms.

$$\|e_t\|_{0,h}^2 \ge C_0 \|e_t\|_0^2,$$

$$a_h(e, \Pi_h^* e_t) = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|||e|||_1^2,$$

$$\begin{aligned} |b_h(e, \Pi_h^* e_t)| &\le Ch\|e\|_1\|e_t\|_1 \le C\|e\|_1\|e_t\|_0 \\ &\le C|||e|||_1^2 + \frac{C_0}{2}\|e_t\|_0^2, \end{aligned}$$

$$|(r, \Pi_h^* e_t)| \le C\|r\|_0^2 + \frac{C_0}{2}\|e_t\|_0^2.$$

Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}t}|||e|||_1^2 \le C(|||e|||_1^2 + \|r\|_0^2).$$

Integrate on $t$ and note

$$\|e(0)\|_1 \le Ch\|u_0\|_2 + \|u_0 - u_{0h}\|_1,$$

then we have

$$|||e|||_1^2 \le \; C\Big[\|u_0 - u_{0h}\|_1^2 + h^2\|u_0\|_2^2 + \int_0^t (|||e|||_1^2 + \|r\|_0^2)\mathrm{d}\tau\Big].$$

Make use of Gronwall's inequality to get

$$\|e\|_1^2 \le C\Big(\|u_0 - u_{0h}\|_1^2 + h^2\|u_0\|_2^2 + \int_0^t \|r\|_0^2 \mathrm{d}\tau\Big). \tag{5.3.10}$$

Write

$$r = r_1 + r_2,$$

$$r_1 = u_t - P_h u_t, \; r_2 = P_h u_t - \Pi_h^* P_h u_t.$$

Then it is easy to see that

$$\|r_1\|_0 \le Ch\|u_t\|_2,$$

$$\|r_2\|_0 \le Ch\|P_h u_t\|_1 \le Ch\|u_t\|_2.$$

Inserting the above two estimates into (5.3.10) yields

$$\|e\|_1 \le C\Big[\|u_0 - u_{0h}\|_1 + h\|u_0\|_2 + h\Big(\int_0^t \|u_\tau\|_2^2 \mathrm{d}\tau\Big)^{1/2}\Big]. \tag{5.3.11}$$

Finally, (5.3.6) follows from (5.3.7), (5.3.8) and (5.3.11). This completes the proof. □

### 5.3.3 Error estimates for fully-discrete schemes

**Theorem 5.3.2** *Let $u$ and $\{u_h^n\}$ be the solutions to the parabolic equation (5.1.3) and the backward Euler, mass concentration, generalized difference scheme (5.3.5), respectively. Then*

$$\begin{aligned} &\|u(t_n) - u_h^n\|_1 \\ \le\; & C\Big\{\|u_0 - u_{0h}\|_1 + h\Big[\|u_0\|_2 + \int_0^{t_n} \|u_t\|_2 \mathrm{d}t \\ &+\Big(\int_0^{t_n} \|u_t\|_2^2 \mathrm{d}t\Big)^{1/2}\Big] + \tau\Big(\int_0^{t_n} \|u_{tt}\|_0^2 \mathrm{d}t\Big)^{1/2}\Big\}, \\ & n = 0, 1, 2, \cdots. \end{aligned} \tag{5.3.12}$$

**Proof** Write

$$u(t_n) - u_h^n = \rho^n + e^n, \tag{5.3.13}$$

where

$$\rho^n = u(t_n) - P_h u(t_n),\ e^n = P_h u(t_n) - u_h^n.$$

It is obvious that

$$\|\rho^n\|_1 \le Ch\|u(t_n)\|_2 \le Ch\Big[\|u_0\|_2 + \int_0^{t_n} \|u_t\|_2 \mathrm{d}t\Big]. \tag{5.3.14}$$

It is easy to check that $e^n$ satisfies

$$(\Pi_h^* \bar{\partial}_t e^n, v_h) + a(e^n, v_h) = (r^n, v_h),\ v_h \in V_h, \tag{5.3.15}$$

where

$$r^n = \Pi_h^* \bar{\partial}_t P_h u(t_n) - u_t(t_n). \tag{5.3.16}$$

Setting $v_h = \Pi_h^* \bar{\partial}_t e^n$ yields

$$\|\bar{\partial}_t e^n\|_{0,h}^2 + a(e^n, \Pi_h^* \bar{\partial}_t e^n) = (r^n, \Pi_h^* \bar{\partial}_t e^n).$$

We have the following estimates for the above terms.

$$\|\bar{\partial}_t e^n\|_{0,h}^2 \ge C_0 \|\bar{\partial}_t e^n\|_0^2,$$

$$\begin{aligned}
& a(e^n, \Pi_h^* \bar{\partial}_t e^n) \\
\ge\ & \frac{1}{2\tau}[a_h(e^n + e^{n-1}, \Pi_h^*(e^n - e^{n-1})) \\
& + b_h(e^n + e^{n-1}, \Pi_h^*(e^n - e^{n-1}))] \\
\ge\ & \frac{1}{2\tau}[(1 - C\tau)|||e^n|||_1^2 - (1 + C\tau)|||e^{n-1}|||_1^2] \\
& - \frac{C_0}{2}\|\bar{\partial}_t e^n\|_0^2,
\end{aligned}$$

$$|(r^n, \Pi_h^* \bar{\partial}_t e^n)| \le C\|r^n\|_0^2 + \frac{C_0}{2}\|\bar{\partial}_t e^n\|_0^2.$$

Consequently

$$|||e^n|||_1^2 \le \frac{1 + C\tau}{1 - C\tau}|||e^{n-1}|||_1^2 + C\tau\|r^n\|_0^2.$$

This implies

$$|||e^n|||_1^2 \le C\Big(|||e^0|||_1^2 + \tau\sum_{j=1}^{n}\|r^j\|_0^2\Big). \tag{5.3.17}$$

Set

$$r^j = r_0^j + r_1^j + r_2^j,$$

where

$$r_0^j = \Pi_h^*\bar\partial_t P_h u(t_j) - \bar\partial_t P_h u(t_j) = (\Pi_h^* - I)\tau^{-1}\int_{t_{j-1}}^{t_j} P_h u_t \mathrm{d}t,$$

$$r_1^j = \bar\partial_t P_h u(t_j) - \bar\partial_t u(t_j),\ r_2^j = \bar\partial_t u(t_j) - u_t(t_j).$$

Correspondingly we have the following estimates:

$$\begin{aligned}\sum_{j=1}^{n}\|r_0^j\|_0^2 &\le Ch^2\tau^{-2}\sum_{j=1}^{n}\Big(\int_{t_{j-1}}^{t_j}\|P_h u_t\|_2\mathrm{d}t\Big)^2\\ &\le Ch^2\tau^{-1}\int_0^{t_n}\|u_t\|_2^2\mathrm{d}t,\end{aligned}$$

$$\sum_{j=1}^{n}\|r_1^j\|_0^2 \le Ch^2\tau^{-1}\int_0^{t_n}\|u_t\|_2^2\mathrm{d}t,$$

$$\sum_{j=1}^{n}\|r_2^j\|_0^2 \le \tau\int_0^{t_n}\|u_{tt}\|_0^2\mathrm{d}t.$$

Substituting these estimates into (5.3.17) gives

$$\begin{aligned}\|e^n\|_1 \le\ & C\Big\{\|u_0 - u_{0h}\|_1 + h\Big[\|u_0\|_2 + \Big(\int_0^{t_n}\|u_t\|_2^2\mathrm{d}t\Big)^{1/2}\Big]\\ &+\tau\Big(\int_0^{t_n}\|u_{tt}\|_0^2\mathrm{d}t\Big)^{1/2}\Big\}.\end{aligned} \tag{5.3.18}$$

A combination of (5.3.13), (5.3.14) and (5.3.18) leads to (5.3.12). This completes the proof. □

## 5.4 High Order Element Difference Schemes

This section is concerned with high order element difference schemes for parabolic equations. First we discuss a cubic element difference scheme for parabolic equations in one dimension, and present its error estimate. Then we consider a quadratic element difference scheme for parabolic equations in two dimensions, for which a numerical example is also provided.

### 5.4.1 Cubic element difference schemes for one-dimensional parabolic equations

Consider the mixed problem of the one-dimensional parabolic equation:

$$\begin{cases} \dfrac{\partial u}{\partial t} + Lu = f(x,t), & x \in (a,b),\ 0 < t \le T, \quad (5.4.1a) \\ u(a,t) = 0,\ \dfrac{\partial u(b,t)}{\partial x} = 0, & 0 < t \le T, \quad (5.4.1b) \\ u(x,0) = u_0(x), & x \in (a,b), \quad (5.4.1c) \end{cases}$$

where

$$Lu \equiv -\frac{\partial}{\partial x}\Big(p\frac{\partial u}{\partial x}\Big) + r\frac{\partial u}{\partial x} + qu,$$

$p \in C^1[a,b]$, $p \ge p_{\min} > 0$, $q, r \in C[a,b]$, and $f \in L^2(a,b)$.

Let us place a quasi-uniform grid $T_h$ and a corresponding barycenter dual grid $T_h^*$ on $[a,b]$. Take the trial function space $U_h$ as the Hermite cubic element space related to $T_h$, and the test function space $V_h$ as the piecewise linear function space with respect to $T_h^*$. For details, see §2.4.

The cubic element semi-discrete difference scheme reads: Find $u_h = u_h(\cdot,t) \in U_h$ $(0 < t \le T)$ such that

$$\begin{cases} \Big(\dfrac{\partial u_h}{\partial t}, v_h\Big) + (Lu_h, v_h) = (f, v_h), \quad (5.4.2a) \\ \qquad \forall v_h \in V_h,\ 0 < t \le T, \\ u_h(x,0) = u_{0h}(x),\ x \in (a,b), \quad (5.4.2b) \end{cases}$$

where $u_{0h} \in U_h$ is some approximation of $u_0$.

The cubic element fully-discrete difference scheme is: Find $u_h^n \in U_h$ $(n = 1, 2, \cdots)$ such that

$$\begin{cases} (\bar{\partial}_t u_h^n, v_h) + (L u_h^{n,\theta}, v_h) = (f^{n,\theta}, v_h), & (5.4.3a) \\ \qquad \forall v_h \in V_h,\ n = 1, 2, \cdots, & \\ u_h^0 = u_{0h}, & (5.4.3b) \end{cases}$$

where ($\tau$ is the time step size, and $t_n = n\tau$)

$$\bar{\partial}_t u_h^n = \frac{u_h^n - u_h^{n-1}}{\tau},\ u_h^{n,\theta} = \theta u_h^n + (1-\theta) u_h^{n-1},$$

$$f^{n,\theta} = \theta f^n + (1-\theta) f^{n-1},\ f^n = f(t_n).$$

(5.4.3) leads to a backward Euler fully-discrete scheme when $\theta = 1$, and a Crank-Nicolson fully-discrete scheme when $\theta = \frac{1}{2}$.

The following lemmas will be used later on for the error estimates.

**Lemma 5.4.1** *The elliptic projection $P_h u \in U_h$ of $u \in H^2(\Omega) \cap H_0^1(\Omega)$ is uniquely defined by*

$$(L P_h u, v_h) = (Lu, v_h),\ v_h \in V_h, \tag{5.4.4}$$

*and satisfies*

$$\|P_h u - u\|_m \le C h^{4-m} |u|_4,\ m = 0, 1. \tag{5.4.5}$$

**Proof** The conclusion is a consequence of Theorems 2.4.1, 2.4.3 and 2.5.2. □

**Lemma 5.4.2** *Let $\Pi_h^* u_h$ denote the interpolation projection of $u_h$ onto $V_h$. Then, $(L u_h, \Pi_h^* \bar{u}_h)$ can be expressed as*

$$(L u_h, \Pi_h^* \bar{u}_h) = a_1(u_h, \Pi_h^* \bar{u}_h) + a_2(u_h, \Pi_h^* \bar{u}_h), \tag{5.4.6}$$

*where the leading term satisfies ($c_1$ and $c_2$ are positive constants)*

$$a_1(u_h, \Pi_h^* \bar{u}_h) = a_1(\bar{u}_h, \Pi_h^* u_h),\ \ \forall \bar{u}_h, u_h \in U_h, \tag{5.4.7}$$

$$c_1\|u_h\|_1^2 \le a_1(u_h, \Pi_h^* u_h) \le c_2\|u_h\|_1^2, \quad \forall u_h \in U_h, \tag{5.4.8}$$

*and the remainder term satisfies*

$$|a_2(u_h, \Pi_h^* \bar{u}_h)| \le ch\|u_h\|_1\|\bar{u}\|_1, \quad \forall \bar{u}_h, u_h \in U_h. \tag{5.4.9}$$

*Set*

$$|||u_h|||_1 = [a_1(u_h, \Pi_h^* u_h)]^{1/2}, \ u_h \in U_h, \tag{5.4.10}$$

*then* $|||\cdot|||_1$ *is equivalent to the* $H^1$*-norm* $\|\cdot\|_1$.

**Proof** Identify $a_1(u_h, \Pi_h^* \bar{u}_h)$ with $b_h(u_h, \Pi_h^* \bar{u}_h)$ in §2.4. Imitating the proofs to (2.4.13), (2.4.17) and (2.4.18), we can show (5.4.7), (5.4.8) and (5.4.9) respectively. This completes the proof. □

A straightforward calculation verifies the following lemma.

**Lemma 5.4.3** *There exists a constant* $\beta > 0$ *independent of the subspace* $U_h$ *such that*

$$(u_h, \Pi_h^* u_h) \ge \beta\|u_h\|_0^2, \quad \forall u_h \in U_h. \tag{5.4.11}$$

**Theorem 5.4.1** *Let* $u$ *and* $u_h$ *be the solutions to the problem (5.4.1) and the semi-discrete cubic element generalized difference scheme (5.4.2), respectively. Then we have*

$$\begin{aligned} \|u - u_h\|_1 \le \ & C\Big\{\|u_0 - u_{0h}\|_1 + h^3\Big[\|u_0\|_4 \\ & + \int_0^t \|u_\tau\|_4 \mathrm{d}\tau + h\Big(\int_0^t \|u_\tau\|_4^2 \mathrm{d}\tau\Big)^{1/2}\Big]\Big\}. \end{aligned} \tag{5.4.12}$$

**Proof** Set

$$u - u_h = \rho + e, \tag{5.4.13}$$

where

$$\rho = u - P_h u, \ e = P_h u - u_h.$$

By Lemma 5.4.1 we have

$$\|\rho\|_1 \le Ch^3\|u\|_4 \le Ch^3\Big(\|u_0\|_4 + \int_0^t \|u_\tau\|_4 \mathrm{d}\tau\Big). \tag{5.4.14}$$

Since $u$ and $u_h$ satisfy (5.4.1) and (5.4.2) respectively, we have

$$\left(\frac{\partial u}{\partial t}-\frac{\partial u_h}{\partial t},v_h\right)+(Lu-Lu_h,v_h)=0,\quad \forall v_h\in V_h. \tag{5.4.15}$$

Thus it follows from (5.4.4) that

$$(e_t,v_h)+(Le,v_h)=-(\rho_t,v_h),\quad \forall v_h\in V_h. \tag{5.4.16}$$

Choosing $v_h=\Pi_h^* e_t$ leads to

$$(e_t,\Pi_h^*e_t)+a_1(e,\Pi_h^*e_t)=-(\rho_t,\Pi_h^*e_t)-a_2(e,\Pi_h^*e_t).$$

Now, it results from (5.4.11), (5.4.7), (5.4.10) and the inverse property of the finite element space that

$$\begin{aligned}&\beta\|e_t\|_0^2+\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|||e|||_1^2\\ \le\ &\|\rho_t\|_0\|e_t\|_0+C\|e\|_1\|e_t\|_0\\ \le\ &C(\|\rho_t\|_0^2+\|e\|_1^2)+\beta\|e_t\|_0^2.\end{aligned}$$

Simplify it and integrate it, then we have

$$|||e|||_1^2\le|||e(0)|||_1^2+C\int_0^t(\|\rho_t\|_0^2+\|e\|_1^2)\mathrm{d}\tau. \tag{5.4.17}$$

Notice the equivalence of the norms and the inequality

$$\begin{aligned}\|e(0)\|_1&\le\|P_hu_0-u_0\|_1+\|u_0-u_{0h}\|_1\\ &\le Ch^3\|u_0\|_4+\|u_0-u_{0h}\|_1.\end{aligned}$$

Hence it follows from the Gronwall's inequality that

$$\|e\|_1^2\le C\Big\{\|u_0-u_{0h}\|_1^2+h^6\|u_0\|_4^2+\int_0^t\|\rho_\tau\|_0^2\mathrm{d}\tau\Big\}. \tag{5.4.18}$$

By Lemma 5.4.1 we have

$$\|\rho_\tau\|_0\le Ch^4\|u_\tau\|_4. \tag{5.4.19}$$

(5.4.18) and (5.4.19) imply

$$\begin{aligned}\|e\|_1 \le\ & C\Big\{\|u_0 - u_{0h}\|_1 + h^3\|u_0\|_4 \\ & + h^4\Big(\int_0^t \|u_\tau\|_4^2 \mathrm{d}\tau\Big)^{1/2}\Big\}.\end{aligned} \tag{5.4.20}$$

A combination of (5.4.13), (5.4.14) and (5.4.20) implies (5.4.12). This completes the proof. □

**Theorem 5.4.2** *Let $u$ and $\{u_h^n\}$ be the solutions to the parabolic equation (5.4.1) and the Crank-Nicolson cubic element generalized difference scheme (5.4.3), respectively. Then*

$$\begin{aligned}& \|u(t_n) - u_h^n\|_1 \\ \le\ & C\Big\{\|u_0 - u_{0h}\|_1 + h^3\Big[\|u_0\|_4 + \int_0^{t_n} \|u_t\|_4 \mathrm{d}t \\ & + h\Big(\int_0^{t_n} \|u_t\|_4^2 \mathrm{d}t\Big)^{1/2}\Big] + \tau^2\Big(\int_0^{t_n} \|u_{ttt}\|_0^2 \mathrm{d}t\Big)^{1/2}\Big\}.\end{aligned} \tag{5.4.21}$$

**Proof** By (5.4.1) and (5.4.3)

$$\Big(\frac{\partial u}{\partial t}, v_h\Big) + (Lu, v_h) = (f, v_h), \quad \forall v_h \in V_h, \tag{5.4.22}$$

$$(\bar{\partial}_t u_h^n, v_h) + \Big(L\frac{u_h^n + u_h^{n-1}}{2}, v_h\Big) = \Big(\frac{f^n + f^{n-1}}{2}, v_h\Big), \quad \forall v_h \in V_h. \tag{5.4.23}$$

Set $t = t_n$ and $t = t_{n-1}$ respectively in (5.4.22), combine them with (5.4.23) and use (5.4.4), then we have

$$(\bar{\partial}_t e^n, v_h) + \Big(L\frac{e^n + e^{n-1}}{2}, v_h\Big) = (r^n, v_h), \quad \forall v_h \in V_h, \tag{5.4.24}$$

where

$$e^n = P_h u(t_n) - u_h^n, \ r^n = \bar{\partial}_t P_h u(t_n) - \frac{\partial}{\partial t}\frac{u(t_n) + u(t_{n-1})}{2}.$$

Choosing $v_h = \Pi_h^*\bar{\partial}_t e^n$ in (5.4.24) yields

$$(\bar{\partial}_t e^n, \Pi_h^*\bar{\partial}_t e^n) + \frac{1}{2\tau}[a_1(e^n + e^{n-1}, \Pi_h^*(e^n - e^{n-1}))$$
$$+a_2(e^n + e^{n-1}, \Pi_h^*(e^n - e^{n-1}))] = (r^n, \Pi_h^*\bar{\partial}_t e^n).$$

It follows from (5.4.11), (5.4.7), (5.4.10) and (5.4.9) that

$$\begin{aligned}&\beta\|\bar{\partial}_t e^n\|_0^2 + \frac{1}{2\tau}(|||e^n|||_1^2 - |||e^{n-1}|||_1^2) - C\|e^n + e^{n-1}\|_1\|\bar{\partial}_t e^n\|_0\\ \leq\ & C\|r^n\|_0\|\bar{\partial}_t e^n\|_0.\end{aligned}$$

So there is a constant $C' > 0$ such that

$$\begin{aligned}&\beta\|\bar{\partial}_t e^n\|_0^2 + \frac{1}{2\tau}[(1 - C'\tau)(|||e^n|||_1^2\\ &-(1 + C'\tau)|||e^{n-1}|||_1^2] - \frac{\beta}{2}\|\bar{\partial}_t e^n\|_0^2\\ \leq\ & C\|r^n\|_0^2 + \frac{\beta}{2}\|\bar{\partial}_t e^n\|_0^2.\end{aligned}$$

Thus

$$|||e^n|||_1^2 \leq \frac{1 + C'\tau}{1 - C'\tau}|||e^{n-1}|||_1^2 + C\tau\|r^n\|_0^2.$$

This recursion relation implies the existence of a constant $C > 0$ such that

$$|||e^n|||_1^2 \leq C\Big(|||e^0|||_1^2 + \tau\sum_{j=1}^{n}\|r^j\|_0^2\Big). \tag{5.4.25}$$

By virtue of (5.4.5) we have

$$\begin{aligned}|||e^0|||_1 &\leq C\|e^0\|_1 = C\|P_h u^0 - u_{0h}\|_1\\ &\leq C(\|u_0 - P_h u_0\|_1 + \|u_0 - u_{0h}\|_1)\\ &\leq Ch^3\|u_0\|_4 + C\|u_0 - u_{0h}\|_1.\end{aligned} \tag{5.4.26}$$

As before, we write

$$r^j = r_1^j + r_2^j, \tag{5.4.27}$$

$$r_1^j = \bar{\partial}_t P_h u(t_j) - \bar{\partial}_t u(t_j) = \frac{1}{\tau}\int_{t_{j-1}}^{t_j}(P_h u_t - u_t)\mathrm{d}t,$$

$$r_2^j = \bar{\partial}_t u(t_j) - \frac{u_t(t_j) + u_t(t_{j-1})}{2}.$$

So by (5.4.5) we have

$$\begin{aligned} \sum_{j=1}^{n} \|r_1^j\|_0^2 &\leq C\tau^{-2} \sum_{j=1}^{n} \Big( \int_{t_{j-1}}^{t_j} h^4 \|u_t\|_4 \mathrm{d}t \Big)^2 \\ &\leq C\tau^{-1} h^8 \int_0^{t_n} \|u_t\|_4^2 \mathrm{d}t. \end{aligned} \tag{5.4.28}$$

Employ the Taylor expansion to get

$$\begin{aligned} \sum_{j=1}^{n} \|r_2^j\|_0^2 &\leq \sum_{j=1}^{n} \Big( C\tau \int_{t_{j-1}}^{t_j} \|u_{ttt}\|_0 \mathrm{d}t \Big)^2 \\ &\leq C\tau^3 \int_0^{t_n} \|u_{ttt}\|_0^2 \mathrm{d}t. \end{aligned} \tag{5.4.29}$$

It follows from (5.4.25)-(5.4.29) that

$$\begin{aligned} \|e^n\|_1 \leq \; & C\Big\{ \|u_0 - u_{0h}\|_1 + h^3 \|u_0\|_4 + h^4 \Big( \int_0^{t_n} \|u_t\|_4^2 \mathrm{d}t \Big)^{1/2} \\ & + \tau^2 \Big( \int_0^{t_n} \|u_{ttt}\|_0^2 \mathrm{d}t \Big)^{1/2} \Big\}. \end{aligned} \tag{5.4.30}$$

This together with

$$\begin{aligned} & \|u(t_n) - P_h u(t_n)\|_1 \\ \leq \; & Ch^3 \|u(t_n)\|_4 \leq Ch^3 \Big[ \|u_0\|_4 + \int_0^{t_n} \|u_t\|_4 \mathrm{d}t \Big] \end{aligned}$$

yields (5.4.21). This completes the proof. □

### 5.4.2 Quadratic element difference schemes for two-dimensional parabolic equations

Consider the following initial and boundary values problem:

$$\begin{cases} \dfrac{\partial u}{\partial t} - \Delta u = f(x,y,t), & (x,y) \in \Omega,\ 0 < t \le T, \qquad (5.4.31a) \\ u(x,y,t) = 0, & (x,y) \in \partial\Omega,\ 0 < t \le T, \quad (5.4.31b) \\ u(x,y,0) = u_0(x,y), & (x,y) \in \Omega, \qquad\qquad (5.4.31c) \end{cases}$$

where $\Omega$ is a planar polygon, and $f, u_0 \in L^2(\Omega)$. A corresponding variational problem is: Find $u = u(\cdot,t) \in H_0^1(\Omega)$ such that

$$\begin{cases} \left(\dfrac{\partial u}{\partial x}, v\right) + a(u,v) = (f,v),\ \forall v \in H_0^1(\Omega),\ 0 < t \le T, & (5.4.32a) \\ (u(\cdot,0),v) = (u_0,v),\ \forall v \in H_0^1(\Omega), & (5.4.32b) \end{cases}$$

where

$$a(u,v) = \int_\Omega \left(\frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y}\right)\mathrm{d}x\mathrm{d}y,\ u,v \in H_0^1(\Omega).$$

As in §3.4, we place a triangulation $T_h = \{K_Q : Q \in \Omega_h^*\}$, and a corresponding dual grid $T_h^* = \{K_{P_0}^*, K_M^* : P_0 \in \overline{\Omega}_h, M \in \overline{M}_h\}$. (See §3.4 and Figg. 5.4.1 and 5.4.2 below for details and notations.)

The trial function space $U_h$ is chosen as the Lagrange quadratic element space related to the triangulation $T_h$. The test space $V_h$ is taken as the piecewise constant function space corresponding to the dual grid $T_h^*$, of which the basis functions are the characteristic functions $\psi_{P_0}$ and $\psi_M$ of $K_{P_0}^*$ and $K_M^*$, respectively.

The semi-discrete quadratic difference scheme reads: Find $u_h = u_h(\cdot,t) \in U_h$ $(0 < t \le T)$ such that

$$\begin{cases} \left(\dfrac{\partial u_h}{\partial x}, v_h\right) + a(u_h,v_h) = (f,v_h),\ \forall v_h \in V_h, 0 < t \le T, & (5.4.33a) \\ (u_h(\cdot,0),v_h) = (u_0,v_h),\ \forall v_h \in V_h, & (5.4.33b) \end{cases}$$

where $a(\cdot,\cdot)$ is interpreted in the sense of generalized functions. In particular, when $v_h$ is taken as the basis functions $\psi_{P_0}$ and $\psi_M$ respectively, we have

$$a(u_h,\psi_{P_0}) = \int_{\partial K_{P_0}^*} \left(-\frac{\partial u_h}{\partial x}\mathrm{d}y + \frac{\partial u_h}{\partial y}\mathrm{d}x\right),$$

$$a(u_h,\psi_M) = \int_{\partial K_M^*} \left(-\frac{\partial u_h}{\partial x}\mathrm{d}y + \frac{\partial u_h}{\partial y}\mathrm{d}x\right).$$

The fully-discrete quadratic element difference scheme is: Find $u_h^n \in U_h$ $(n = 1, 2, \cdots)$ such that

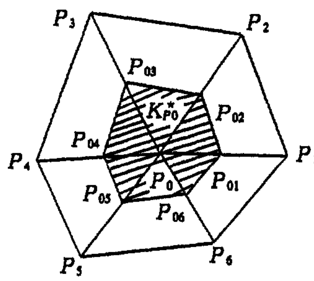



Fig. 5.4.1

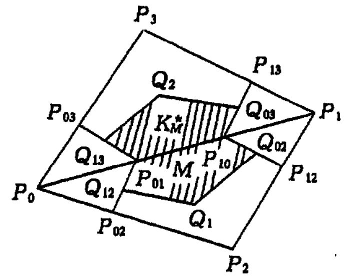



Fig. 5.4.2

$$\begin{cases} (\bar{\partial}_t u_h^n, v_h) + a(u_h^{n,\theta}, v_h) = (f^{n,\theta}, v_h), & (5.4.34a) \\ \qquad \forall v_h \in V_h,\ n = 1, 2, \cdots, & \\ (u_h^0, v_h) = (u_0, v_h), \ \ \forall v_h \in V_h, & (5.4.34b) \end{cases}$$

where ($\tau$ is the time step size and $t_n = n\tau$)

$$\bar{\partial}_t u_h^n = \frac{u_h^n - u_h^{n-1}}{\tau}, \ u_h^{n,\theta} = \theta u_h^n + (1-\theta) u_h^{n-1},$$

$$f^{n,\theta} = \theta f^n + (1-\theta) f^{n-1}, \ f^n = f(t_n).$$

(5.4.34) leads to a backward Euler fully-discrete scheme when $\theta = 1$, and a Crank-Nicolson fully-discrete scheme when $\theta = \frac{1}{2}$.

As in finite element methods, we can first compute the element mass matrices and the element stiff matrices, then pile them up to form the overall mass and stiff matrices respectively.

A direct computation gives

$$\left(\frac{\partial u_h}{\partial t}, v_h\right) = \sum_{K \in T_h} \{v_h\}_K^T B \{\dot{u}_h\}_K,$$

where (cf. Fig. 5.4.3))

$$\{v_h\}_K^T = \ [v_h(P_i), v_h(P_j), v_h(P_k), v_h(M_i), v_h(M_j), v_h(M_k)],$$

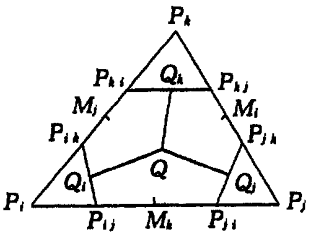



Fig. 5.4.3

$$
\begin{aligned}
\{\dot{u}_h\}_K = \ & \Big[\frac{\partial u_h}{\partial t}(P_i), \frac{\partial u_h}{\partial t}(P_j), \frac{\partial u_h}{\partial t}(P_k), \\
& \frac{\partial u_h}{\partial t}(M_i), \frac{\partial u_h}{\partial t}(M_j), \frac{\partial u_h}{\partial t}(M_k)\Big]^T,
\end{aligned}
$$

$$
B = \frac{S_K}{1944}\begin{bmatrix} 96 & -16 & -16 & 8 & 72 & 72 \\ -16 & 96 & -16 & 72 & 8 & 72 \\ -16 & -16 & 96 & 72 & 72 & 8 \\ -38 & -13 & -13 & 300 & 98 & 98 \\ -13 & -38 & -13 & 98 & 300 & 98 \\ -13 & -13 & -38 & 98 & 98 & 300 \end{bmatrix}.
$$

It is this $B$ that is called the element mass matrix.

Also note that

$$
a(u_h, v_h) = \sum_{K \in T_h} I_K(u_h, v_h),
$$

$$
I_K(u_h, v_h) = \{v_h\}_K^T A \{u_h\}_K,
$$

where $A$ is the element stiff matrix defined as follows:

$$
A = \frac{1}{36 S_K}[\bar{a}_{ij}]_{6\times 6},
$$

$$
\begin{aligned}
&\bar{a}_{11} = 10c^2, && \bar{a}_{12} = a^2 - b^2 + c^2, \\
&\bar{a}_{13} = -a^2 + b^2 + c^2, && \bar{a}_{14} = -4c^2,
\end{aligned}
$$

$$\begin{aligned}
\bar{a}_{15} &= 8a^2 - 8b^2 - 4c^2, & \bar{a}_{16} &= -8a^2 + 8b^2 - 4c^2,\\
\bar{a}_{21} &= a^2 - b^2 + c^2, & \bar{a}_{22} &= 10a^2,\\
\bar{a}_{23} &= a^2 + b^2 - c^2, & \bar{a}_{24} &= -4a^2 - 8b^2 + 8c^2,\\
\bar{a}_{25} &= -4a^2, & \bar{a}_{26} &= -4a^2 + 8b^2 - 8c^2,\\
\bar{a}_{31} &= -a^2 + b^2 + c^2, & \bar{a}_{32} &= a^2 + b^2 - c^2,\\
\bar{a}_{33} &= 10b^2, & \bar{a}_{34} &= -8a^2 - 4b^2 + 8c^2,\\
\bar{a}_{35} &= 8a^2 - 4b^2 - 8c^2, & \bar{a}_{36} &= -4b^2,\\
\bar{a}_{41} &= -2c^2, & \bar{a}_{42} &= -5a^2 - 3b^2 + 3c^2,\\
\bar{a}_{43} &= -3a^2 - 5b^2 + 3c^2, & \bar{a}_{44} &= 8a^2 + 8b^2 + 4c^2,\\
\bar{a}_{45} &= -8a^2 + 8b^2 - 4c^2, & \bar{a}_{46} &= 8a^2 - 8b^2 - 4c^2,\\
\bar{a}_{51} &= 3a^2 - 3b^2 - 5c^2, & \bar{a}_{52} &= -2a^2,\\
\bar{a}_{53} &= 3a^2 - 5b^2 - 3c^2, & \bar{a}_{54} &= -4a^2 + 8b^2 - 8c^2,\\
\bar{a}_{55} &= 4a^2 + 8b^2 + 8c^2, & \bar{a}_{56} &= -4a^2 - 8b^2 + 8c^2,\\
\bar{a}_{61} &= -3a^2 + 3b^2 - 5c^2, & \bar{a}_{62} &= -5a^2 + 3b^2 - 3c^2,\\
\bar{a}_{63} &= -2b^2, & \bar{a}_{64} &= 8a^2 - 4b^2 - 8c^2,\\
\bar{a}_{65} &= -8a^2 - 4b^2 + 8c^2, & \bar{a}_{66} &= 8a^2 + 4b^2 + 8c^2,
\end{aligned}$$

where $a = |\overline{P_iP_k}|$, $b = |\overline{P_iP_j}|$, and $c = |\overline{P_jP_k}|$.

The results of a numerical experiment are given in Table 5.4.1, where the Crank-Nicolson fully-discrete generalized difference scheme (5.4.34) (GDM) is compared with the linear finite element method (FEM1) and the quadratic finite element method (FEM2), for the following initial and boundary values problem:

$$\begin{cases}
\dfrac{\partial u}{\partial t} = \dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2}, \ (x,y) \in \Omega = (0,\pi) \times (0,\pi), \ 0 < t \le 1,\\
u|_{\partial\Omega} = 0, \ 0 < t \le 1,\\
u|_{t=0} = \sin x \cdot \sin y, \ (x,y) \in \Omega.
\end{cases}$$

Place a right angle triangulation with a space step size $\pi/4$ and a time step size $\tau = 0.001$. The average error and the maximum error (in absolute values) of the approximate solutions, at all the nodes when $t = \frac{1}{2}$, and the true solution $u = e^{-2t}\sin x \cdot \sin y$ are given in Table 5.4.1.

**Table 5.4.1** Comparison of approximation errors

| | GDM | FEM1 | FEM2 |
|---|---|---|---|
| average error | 0.005064 | 0.036336 | 0.000770 |
| maximum error | 0.009187 | 0.054020 | 0.001568 |

## 5.5 Generalized Difference Methods for Nonlinear Parabolic Equations

### 5.5.1 Problem and schemes

Let us consider the following initial and boundary values problem of nonlinear parabolic equations:

$$\begin{cases} \dfrac{\partial u}{\partial t} + Au = f(x,y,t), & (x,y) \in \Omega,\ 0 < t \le T, & (5.5.1a) \\ u = 0, & (x,y) \in \partial\Omega,\ 0 < t \le T, & (5.5.1b) \\ u = u_0(x,y), & (x,y) \in \Omega,\ t = 0, & (5.5.1c) \end{cases}$$

where

$$Au = -\nabla(a(x,y,u)\nabla u),$$

$\Omega$ is a planar polygonal region, $u_0$ a smooth function on $\bar{\Omega}$, $f$ a smooth function on $\bar{\Omega} \times [0,T]$, and $a(x,y,u)$ a smooth function on $\bar{\Omega}\times$R.

Place on $\Omega$ a triangulation $T_h$ and its corresponding circumcenter dual grid $T_h^*$. Assume that any inner angle of each triangular element is not greater than $\frac{\pi}{2}$, and that $T_h$ and $T_h^*$ are quasi-uniform. Also assume the following "quasi-parallelogram condition" holds: there exists a constant $\mu > 0$ such that for any adjacent triangular elements $K_Q$ and $K_{Q'}$ (cf. Fig. 5.5.1)

$$|\, \Delta_j^Q - \Delta_i^Q + \Delta_j^{Q'} - \Delta_i^{Q'} \,| \le \mu h^3,$$

where, e.g., $\Delta_j^Q$ denotes the area of the triangle with vertexes $Q, P_i$ and $P_k$.

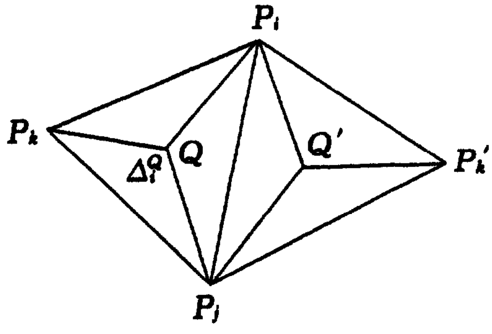



Fig. 5.5.1

Let $U_h$ be the linear element space corresponding to $T_h$, and $V_h$ the piecewise constant function space related to $T_h^*$. The semi-discrete generalized difference scheme approximating (5.5.1) is: Find $u_h = u_h(\cdot, t) \in U_h$ $(0 < t \le T)$ such that

$$\begin{cases} \left(\dfrac{\partial u_h}{\partial t}, v_h\right) + A(u_h; u_h, v_h) = (f, v_h), & (5.5.2a) \\ \qquad \forall v_h \in V_h,\ 0 < t \le T, & \\ u_h(x, y, 0) = u_{0h}(x, y),\ (x, y) \in \Omega, & (5.5.2b) \end{cases}$$

where

$$A(w; u, v) = \int_\Omega a(x, y, w) \nabla u \cdot \nabla v \mathrm{d}x\mathrm{d}y.$$

For $\bar{u}_h, u_h \in U_h$, let $\Pi_h^*$ be the interpolation projection operator from $U_h$ onto $V_h$, then we can write

$$\begin{aligned} & A(w; u_h, \Pi_h^* \bar{u}_h) \\ = & \sum_{K \in T_h} \sum_{l=i,j,k} |\overline{P_{l+1}P_{l+2}}| \int_{\overline{M_l Q}} a(x, y, w) \frac{\partial u_h}{\partial \tau_l} \frac{\partial \bar{u}_h}{\partial \tau_l} \mathrm{d}s, \end{aligned} \tag{5.5.3}$$

where $\tau_l = \overline{P_{l+1}P_{l+2}} / |\overline{P_{l+1}P_{l+2}}|$ $(l = i, j, k;\ i+1 = j, j+1 = k, k+1 = i)$. (cf. §4.5 and Fig. 5.5.2.) Assume $u_{0h}$ is a certain approximat of $u_0$ in $U_h$, satisfying

$$\|u_0 - u_{0h}\|_0 \le Ch.$$

The Crank-Nicolson fully-discrete generalized difference scheme for (5.5.1) is: Find $u_h^n \in U_h$ $(n = 1, 2, \cdots, N)$ such that

$$\begin{cases} (\bar{\partial}_t u_h^n, v_h) + A(u_h^{n-1/2}; u_h^{n-1/2}, v_h) = (f^{n-1/2}, v_h), & (5.5.4a) \\ \qquad \forall v_h \in V_h,\ n = 1, 2, \cdots, N, & \\ u_h^0 = u_{0h}, & (5.5.4b) \end{cases}$$

where $\tau$ is the time step size, $N = T/\tau$, $t_n = n\tau$ and

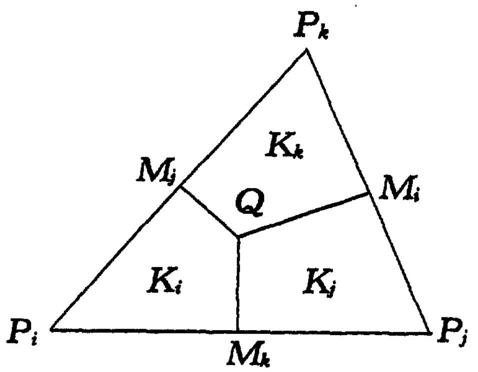



Fig. 5.5.2

$$\bar{\partial}_t u_h^n = \frac{u_h^n - u_h^{n-1}}{\tau},\ u_h^{n-1/2} = \frac{u_h^n + u_h^{n-1}}{2},$$

$$f^{n-1/2} = \frac{f^n + f^{n-1}}{2},\ f^n = f(x, y, t_n).$$

In the sequel we assume the following:

(i) $a(x, y, u), \dfrac{\partial}{\partial t} a(x, y, u) \in C(\bar{\Omega} \times [0, T]),$

$\qquad \forall (x, y) \in \Omega,\ u \in C(\bar{\Omega} \times [0, T]).$

(ii) $0 < a_0 \le a(x, y, u) \le a_1 < +\infty,\ \left|\dfrac{\partial}{\partial t} a(x, y, u)\right| \le a_1,$

$\forall (x, y) \in \Omega,\ u \in C(\bar{\Omega} \times [0, T]),\ t \in [0, T].$

(iii) $|a(x,y,u) - a(x,y,w)| \le M|u - w|$,
$\forall (x,y) \in \Omega,\ u,v \in C(\bar{\Omega} \times [0,T])$.

(iv) (5.5.1) has a unique solution $u$, and $u, u_t \in C([0,T]; C^2(\bar{\Omega}))$,
$u_{tt} \in L^2((0,T); H^1(\Omega)),\ u_{ttt} \in L^2((0,T); L^2(\Omega))$.

Here and below, the following Banach function spaces are used. Let $X$ be a Banach space equipped with a norm $\|\cdot\|_X$, $m$ a nonnegative integer, $-\infty \le a < b \le \infty$, and $1 \le p \le \infty$. We define

$$C^m([a,b];X) := \{u(t) :\ u(t) \text{ as a function from } [a,b] \text{ to } X \text{ is } m\text{-times continuously differentiable}\},$$

and

$$L^p((a,b);X) = \{u(t) : u(t) \in X,\ \ \forall t \in [a,b],\ \|u(t)\|_X \in L^p(a,b)\}.$$

In particular, we write

$$C([a,b];X) = C^0([a,b];X).$$

$C^m([a,b];X)$ and $L^p((a,b);X)$ become Banach spaces when supplied with the following norms respectively:

$$\|u\|_{C^m([a,b];X)} = \sum_{j=0}^{m} \max_{t\in[a,b]} \Big\| \frac{\partial^j}{\partial t^j} u(t) \Big\|_X,$$

$$\|u\|_{L^p((a,b);X)} = \begin{cases} \Big\{ \int_a^b \|u(t)\|_X^p \mathrm{d}t \Big\}^{1/p}, & 1 \le p < \infty, \\ \operatorname*{ess\ sup}_{t\in(a,b)} \|u(t)\|_X, & p = \infty. \end{cases}$$

Furthermore, if $X$ is a Hilbert space with an inner product $(\cdot,\cdot)_X$, then $L^2((a,b);X)$ is a Hilbert space as well with the inner product

$$(u,v)_{L^2((a,b);X)} = \int_a^b (u(t), v(t))_X \mathrm{d}t.$$

### 5.5.2 Some lemmas

In the error estimations later on, besides the results such as the equivalences of the norms given in §5.1, we also need some preliminary results presented below.

**Lemma 5.5.1** *The following estimates holds for any $w \in C(\Omega \times [0,T])$*

$$A(w; u_h, \Pi_h^* \bar{u}_h) = A(w; \bar{u}_h, \Pi_h^* u_h), \quad \forall \bar{u}_h, u_h \in U_h, \tag{5.5.5}$$

$$A(w; u_h, \Pi_h^* u_h) \geq \beta \|u_h\|_1^2, \quad \forall u_h \in U_h, \tag{5.5.6}$$

$$|A(w; u_h, \Pi_h^* \bar{u}_h)| \leq C \|u_h\|_1 \|\bar{u}_h\|_1, \quad \forall \bar{u}_h, u_h \in U_h, \tag{5.5.7}$$

*where $\beta$ and $C$ are positive constants independent of $U_h$.*

The above results can be found in Lemma 5.1.2 and Theorem 4.5.1. The next lemma reveals some properties of $(u_h, \Pi_h^* \bar{u}_h)$ for the circumcenter decomposition, which are similar to those in Lemma 5.1.5 for the barycenter decomposition discussed there.

**Lemma 5.5.2** *There exist positive constants $C_0$ and $C$ such that*

$$(u_h, \Pi_h^* u_h) \geq C_0 \|u_h\|_0^2, \quad \forall u_h \in U_h, \tag{5.5.8}$$

$$|(u_h, \Pi_h^* \bar{u}_h)| \leq C \|u_h\|_0 \|\bar{u}_h\|_0, \quad \forall \bar{u}_h, u_h \in U_h, \tag{5.5.9}$$

$$|(u_h, \Pi_h^* \bar{u}_h) - (\bar{u}_h, \Pi_h^* u_h)| \leq Ch \|u_h\|_0 \|\bar{u}_h\|_0, \tag{5.5.10}$$

$$\forall \bar{u}_h, u_h \in U_h.$$

**Proof** Write $K = \triangle P_i P_j P_k \in T_h$, $K_l = K \cap K_{P_l}^*$, $u_l = u_h(P_l)$ $(l = i, j, k)$, $u_Q = u_h(Q)$. (Cf. Fig. 5.5.2.) Since the inner angles of $K$ are not greater than $\frac{\pi}{2}$, it is easy to verify that

$$\triangle_{l-1} + \triangle_{l+1} \geq \triangle_l, \ (l = i, j, k) \tag{5.5.11}$$

and hence

$$\triangle_{l-1} + \triangle_{l+1} \geq \frac{1}{2} S_K, \ (l = i, j, k) \tag{5.5.12}$$

where $\triangle_i$, $\triangle_j$ and $\triangle_k$ denote the areas of $\triangle QP_jP_k$, $\triangle QP_kP_i$ and $\triangle QP_iP_j$, respectively.

Note that for any $u_h \in U_h$,

$$\int_{K_i} u_h \mathrm{d}x\mathrm{d}y = \frac{1}{3}\Big(u_i + \frac{1}{2}(u_i+u_j) + u_Q\Big)\frac{\triangle_k}{2} + \frac{1}{3}\Big(u_i + \frac{1}{2}(u_i+u_k) + u_Q\Big)\frac{\triangle_j}{2},$$

and

$$u_Q = (u_i\,\triangle_i + u_j\,\triangle_j + u_k\triangle_k)/S_K.$$

So we have

$$\begin{aligned}
&\int_{K_i} u_h \mathrm{d}x\mathrm{d}y \\
= {}& \frac{1}{6}\Big\{u_i(\triangle_j+\triangle_k)\Big(\frac{3}{2}+\frac{\triangle_i}{S_K}\Big) + u_j\Big[\frac{\triangle_k}{2} + (\triangle_j+\triangle_k)\frac{\triangle_j}{S_K}\Big] \\
&+u_k\Big[\frac{\triangle_j}{2} + (\triangle_j+\triangle_k)\frac{\triangle_k}{S_K}\Big]\Big\}.
\end{aligned} \tag{5.5.13}$$

This together with (5.5.11) and (5.5.12) gives

$$\begin{aligned}
&\sum_{l=i,j,k} u_l \int_{K_l} u_h \mathrm{d}x\mathrm{d}y \\
= {}& \frac{1}{6}\sum_{l=i,j,k}\Big\{u_l^2(\triangle_{l-1}+\triangle_{l+1})\Big(\frac{3}{2}+\frac{\triangle_l}{S_K}\Big) \\
&+u_lu_{l+1}\Big[\frac{\triangle_{l-1}}{2} + (\triangle_{l-1}+\triangle_{l+1})\frac{\triangle_{l+1}}{S_K}\Big] \\
&+u_lu_{l-1}\Big[\frac{\triangle_{l+1}}{2} + (\triangle_{l-1}+\triangle_{l+1})\frac{\triangle_{l-1}}{S_K}\Big]\Big\} \\
= {}& \frac{1}{6}\sum_{l=i,j,k}\Big\{u_l^2(\triangle_{l-1}+\triangle_{l+1})(\triangle_{l-1}+\triangle_{l+1}+2\triangle_l)/S_K \\
&+u_lu_{l+1}(\triangle_{l-1}\,\triangle_{l+1} + \triangle_l\,\triangle_{l-1} + \triangle_l^2 + \triangle_{l+1}^2)/S_K \\
&+\frac{1}{2}(u_l+u_{l+1})^2\triangle_{l-1}\Big\}
\end{aligned}$$

$$\begin{aligned}
&\geq \frac{1}{6S_K}\sum_{l=i,j,k}[u_l^2(\Delta_{l-1}^2+\Delta_{l+1}^2+2\,\Delta_l^2+2\,\Delta_{l-1}\,\Delta_{l+1})\\
&\quad +u_lu_{l+1}(\Delta_l^2+\Delta_{l+1}^2+\Delta_{l-1}\,\Delta_{l+1}+\Delta_l\,\Delta_{l-1})]\\
&= \frac{1}{12S_K}\sum_{l=i,j,k}[u_l^2(\Delta_{l-1}^2+\Delta_{l+1}^2+2\,\Delta_l^2+2\,\Delta_{l-1}\,\Delta_{l+1})\\
&\quad +u_{l+1}^2(\Delta_l^2+\Delta_{l-1}^2+2\,\Delta_{l+1}^2+2\,\Delta_l\,\Delta_{l-1})\\
&\quad +2u_lu_{l+1}(\Delta_l^2+\Delta_{l+1}^2+\Delta_{l-1}\,\Delta_{l+1}+\Delta_l\,\Delta_{l-1})]\\
&\geq \frac{1}{12S_K}\sum_{l=i,j,k}[(u_l+u_{l+1})^2(\Delta_l^2+\Delta_{l+1}^2)\\
&\quad +(u_l\,\Delta_{l-1}+u_{l+1}\Delta_{l+1})^2+(u_l\,\Delta_l+u_{l+1}\Delta_{l-1})^2]\\
&\geq \frac{1}{24S_K}\sum_{l=i,j,k}(u_l+u_{l+1})^2(\Delta_l+\Delta_{l+1})^2\\
&\geq \frac{S_K}{96}\sum_{l=i,j,k}(u_l+u_{l+1})^2\\
&= \frac{S_K}{96}\Big[\sum_{l=i,j,k}u_l^2+\Big(\sum_{l=i,j,k}u_l\Big)^2\Big]\\
&\geq \frac{S_K}{96}\sum_{l=i,j,k}u_l^2.
\end{aligned}\tag{5.5.14}$$

So by the equivalence of the norms there exists a constant $C_0 > 0$ such that

$$(u_h,\Pi_h^* u_h)=\sum_{K\in T_h}\sum_{l=i,j,k}u_l\int_{K_l}u_h\mathrm{d}x\mathrm{d}y\geq C_0\|u_h\|_0^2,\quad \forall u_h\in U_h.$$

It is easy to see that

$$\begin{aligned}
(u_h,\Pi_h^*\bar{u}_h) &= \sum_{K\in T_h}\sum_{l=i,j,k}\bar{u}_l\int_{K_l}u_h\mathrm{d}x\mathrm{d}y\\
&\leq C\sum_{K\in T_h}\sum_{l,m=i,j,k}\bar{u}_lu_mS_K\\
&\leq C\|u_h\|_0\|\bar{u}_h\|_0,\quad \forall \bar{u}_h,u_h\in U_h.
\end{aligned}$$

It follows from (5.5.13) that

$$
\begin{aligned}
&(u_h, \Pi_h^* w_h) - (w_h, \Pi_h^* u_h) \\
= \; & \frac{1}{6S_K} \sum_{K \in T_h} \sum_{l=i,j,k} [w_l u_{l+1} (\Delta_{l+1} \, \Delta_{l-1} + \Delta_{l+1}^2 - \Delta_l^2 - \Delta_l \, \Delta_{l-1}) \\
& + w_{l+1} u_l (\Delta_l^2 + \Delta_l \, \Delta_{l-1} - \Delta_{l+1} \, \Delta_{l-1} - \Delta_{l+1}^2)] \\
= \; & \frac{1}{6} \sum_{K \in T_h} \sum_{l=i,j,k} (w_l u_{l+1} - w_{l+1} u_l)(\Delta_{l+1} - \Delta_l).
\end{aligned}
$$

Summing the right-hand side over every side $L = \overline{P_i P_j}$ , and using the boundary condition and the "quasi-parallelogram condition" yield the following estimate (cf. Fig. 5.5.1):

$$
\begin{aligned}
& |(u_h, \Pi_h^* w_h) - (w_h, \Pi_h^* u_h)| \\
= \; & \Big| \sum_L (w_i u_j - w_j u_i)(\Delta_j^Q - \Delta_i^Q + \Delta_j^{Q'} - \Delta_i^{Q'}) \Big| \\
\le \; & Ch \Big( \sum u_i^2 h^2 \Big)^{1/2} \Big( \sum w_i^2 h^2 \Big)^{1/2} \\
\le \; & Ch \|u_h\|_0 \|w_h\|_0.
\end{aligned}
$$

This completes the proof. □

**Lemma 5.5.3** *Assume $w \in C(\bar{\Omega} \times [0,T])$ and $u \in C([0,T]; H_0^1(\Omega) \cap C^2(\bar{\Omega}))$. Then there exists a constant $C > 0$ independent of the subspace $U_h$ such that for all $\bar{u}_h, u_h \in U_h$*

$$|A(w; u - u_h, \Pi_h^* \bar{u}_h)| \le C(h + \|u - u_h\|_1) \|\bar{u}_h\|_1, \tag{5.5.15}$$

$$|A_t(w; u - u_h, \Pi_h^* \bar{u}_h)| \le C(h + \|u - u_h\|_1) \|\bar{u}_h\|_1, \tag{5.5.16}$$

*where*

$$
\begin{aligned}
& A_t(w; v, \Pi_h^* \bar{u}_h) \\
= \; & \sum_{K \in T_h} \sum_{l=i,j,k} \overline{P_{l+1} P_{l+2}} \int_{\overline{M_l Q}} \frac{\partial a(x,y,w)}{\partial t} \frac{\partial v}{\partial \tau_l} \frac{\partial \bar{u}_h}{\partial \tau_l} ds.
\end{aligned}
$$

**Proof** It follows from (5.5.7) that

$$\begin{aligned}
&|A(w;u-u_h,\Pi_h^*\bar{u}_h)| \\
\leq\ & |A(w;u-\Pi_h u,\Pi_h^*\bar{u}_h)| + |A(w;\Pi_h u-u_h,\Pi_h^*\bar{u}_h)| \\
\leq\ & C|\nabla(u-\Pi_h u)|_\infty \sum_{K\in T_h}\sum_{l=i,j,k}\overline{P_{l+1}P_{l+2}}\int_{\overline{M_lQ}}\Big|\frac{\partial\bar{u}_h}{\partial\tau_l}\Big|\mathrm{d}s \\
&+C\|\Pi_h u-u_h\|_1\|\bar{u}_h\|_1.
\end{aligned}\tag{5.5.17}$$

Notice the following estimates

$$|\nabla(u-\Pi_h u)|_\infty \leq Ch\|u\|_{C([0,T];C^2(\bar{\Omega}))},$$

$$\begin{aligned}
&\sum_{K\in T_h}\sum_{l=i,j,k}\overline{P_{l+1}P_{l+2}}\int_{\overline{M_lQ}}\Big|\frac{\partial\bar{u}_h}{\partial\tau_l}\Big|\mathrm{d}s \\
\leq\ & C\sum_{K_Q\in T_h}\Big(\Big|\frac{\partial\bar{u}_h(Q)}{\partial x}\Big|+\Big|\frac{\partial\bar{u}_h(Q)}{\partial y}\Big|\Big)h^2 \\
\leq\ & C\Big(\sum_{K_Q\in T_h}h^2\Big)^{1/2}\Big(\sum_{K_Q\in T_h}\Big(\Big|\frac{\partial\bar{u}_h(Q)}{\partial x}\Big|^2+\Big|\frac{\partial\bar{u}_h(Q)}{\partial y}\Big|^2\Big)h^2\Big)^{1/2} \\
\leq\ & C\|\bar{u}_h\|_1.
\end{aligned}$$

$$\begin{aligned}
\|\Pi_h u-u_h\|_1 &\leq \|\Pi_h u-u\|_1+\|u-u_h\|_1 \\
&\leq Ch+\|u-u_h\|_1.
\end{aligned}$$

So (5.5.15) holds. (5.5.16) can be similarly proved. □

**Lemma 5.5.4** *Let $P_h u$ be the elliptic projection of the solution $u = u(x,y,t)$ to (5.5.1), onto $U_h$, that is, let $P_h u\in U_h$ satisfy*

$$A(u;P_h u-u,v_h)=0,\ \ \forall v_h\in V_h,\ 0<t\leq T. \tag{5.5.18}$$

*Then there exists a constant $C>0$ independent of $U_h$ such that*

$$\|u-P_h u\|_1\leq Ch, \tag{5.5.19}$$

$$\Big\|\frac{\partial}{\partial t}(u-P_h u)\Big\|_1\leq Ch. \tag{5.5.20}$$

**Proof** Since (5.5.18) is a linear system of equations, its solution $P_h u$ uniquely exists by (5.5.6). It follows from (5.5.6), (5.5.18) and (5.5.15) that

$$\begin{aligned}&\|\Pi_h u - P_h u\|_1^2\\ \le\ & CA(u; \Pi_h u - P_h u, \Pi_h^*(\Pi_h u - P_h u))\\ \le\ & C(h + \|\Pi_h u - u\|_1)\|\Pi_h u - P_h u\|_1.\end{aligned}$$

Hence

$$\|u - P_h u\|_1 \le \|u - \Pi_h u\|_1 + \|\Pi_h u - P_h u\|_1 \le Ch.$$

(5.5.20) can be proved in like manner by virtue of (5.5.16). This completes the proof. □

**Lemma 5.5.5** *If $u \in C([0,T]; H_0^1(\Omega) \cap C^2(\bar{\Omega}))$, then there exists a constant $C > 0$ independent of $U_h$ such that*

$$\|\nabla P_h u\|_\infty \le C,\ 0 \le t \le T. \tag{5.5.21}$$

**Proof** It is obvious that

$$\begin{aligned}&\|\nabla P_h u\|_\infty\\ \le\ & \|\nabla(P_h u - \Pi_h u)\|_\infty + \|\nabla(\Pi_h u - u)\|_\infty + \|\nabla u\|_\infty.\end{aligned}$$

By the inverse property of the finite element and (5.5.19) we have

$$\begin{aligned}&\|\nabla(P_h u - \Pi_h u)\|_\infty \le Ch^{-1}\|P_h u - \Pi_h u\|_1\\ \le\ & Ch^{-1}(\|P_h u - u\|_1 + \|u - \Pi_h u\|_1) \le C.\end{aligned}$$

Also note

$$\|\nabla(\Pi_h u - u)\|_\infty \le Ch\|u\|_{2,\infty}.$$

Thus (5.5.21) holds. This completes the proof. □

**Lemma 5.5.6** *If $u \in C([0,T]; H_0^1(\Omega) \cap C^2(\bar{\Omega}))$, then there exists a constant $C > 0$ independent of $U_h$ such that*

$$\begin{aligned} & |A(u; P_h u, \Pi_h^* \bar{u}_h) - A(w_h; P_h u, \Pi_h^* \bar{u}_h)| \\ \leq\ & C(h + \|u - w_h\|_0)\|\bar{u}_h\|_1, \quad \forall w_h, \bar{u}_h \in U_h. \end{aligned} \tag{5.5.22}$$

**Proof** The following estimate follows from (5.5.3), the hypothesis (iii), (5.5.21) and (5.5.18):

$$\begin{aligned} & |A(u; P_h u, \Pi_h^* \bar{u}_h) - A(w_h; P_h u, \Pi_h^* \bar{u}_h)| \\ =\ & \Big| \sum_{K \in T_h} \sum_{l=i,j,k} |\overline{P_{l+1} P_{l+2}}| \\ & \cdot \int_{\overline{M_l Q}} [a(x,y,u) - a(x,y,w_h)] \frac{\partial P_h u}{\partial \tau_l} \frac{\partial \bar{u}_h}{\partial \tau_l} \mathrm{d}s \Big| \\ \leq\ & C \sum_{K \in T_h} \sum_{l=i,j,k} |\overline{P_{l+1} P_{l+2}}| \\ & \cdot \int_{\overline{M_l Q}} |u - w_h| \|\nabla P_h u\|_\infty \Big| \frac{\partial \bar{u}_h}{\partial \tau_l} \Big| \mathrm{d}s \\ \leq\ & C \sum_{K \in T_h} \sum_{l=i,j,k} |\overline{P_{l+1} P_{l+2}}| \\ & \cdot \int_{\overline{M_l Q}} (|u - \Pi_h u| + |\Pi_h u - w_h|) \Big| \frac{\partial \bar{u}_h}{\partial \tau_l} \Big| \mathrm{d}s \\ \leq\ & C\Big( h|u|_2 \sum_{K \in T_h} \sum_{l=i,j,k} |\overline{P_{l+1} P_{l+2}}| \cdot \int_{\overline{M_l Q}} \Big| \frac{\partial \bar{u}_h}{\partial \tau_l} \Big| \mathrm{d}s \\ & + \sum_{K \in T_h} \sum_{l=i,j,k} |\overline{P_{l+1} P_{l+2}}| \cdot \int_{\overline{M_l Q}} |\Pi_h u - w_h| \Big| \frac{\partial \bar{u}_h}{\partial \tau_l} \Big| \mathrm{d}s \Big) \\ \leq\ & C(h|u|_2 \|\bar{u}_h\|_1 + \|\Pi_h u - w_h\|_0 \|\bar{u}_h\|_1) \\ \leq\ & C(h + \|u - w_h\|_0) \|\bar{u}_h\|_1. \end{aligned}$$

This completes the proof. □

### 5.5.3 Error estimates

**Theorem 5.5.1** *Let $u$ and $u_h$ be the solutions to the problem (5.5.1) and the semi-discrete generalized difference scheme (5.5.2) respectively, satisfying $u, u_t \in C([0,T]; C^2(\bar{\Omega}))$, then*

$$\max_{0\le t\le T} \|u-u_h\|_0 + \Big(\int_0^t \|u-u_h\|_1^2 \mathrm{d}t\Big)^{1/2} \le Ch. \tag{5.5.23}$$

**Proof** It follows from (5.5.1), (5.5.2) and (5.5.19) that

$$\begin{aligned} & (e_t, v_h) + A(u_h; e, v_h) \\ = \ & -(\rho_t, v_h) + A(u_h; P_h u, v_h) - A(u; P_h u, v_h), \end{aligned} \tag{5.5.24}$$

where

$$e = P_h u - u_h, \ \rho = u - P_h u.$$

Setting $v_h = \Pi_h^* e$ and using (5.5.10) and (5.5.22) yield

$$\begin{aligned} & \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}(e, \Pi_h^* e) + A(u_h; e, \Pi_h^* e) \\ = \ & \frac{1}{2}[(e, \Pi_h^* e_t) - (e_t, \Pi_h^* e)] - (\rho_t, \Pi_h^* e) \\ & + A(u_h; P_h u, \Pi_h^* e) - A(u; P_h u, \Pi_h^* e) \\ \le \ & Ch\|e\|_0\|e_t\|_0 + C\|\rho_t\|_0\|e\|_0 \\ & + C(h + \|u-u_h\|_0)\|e\|_1. \end{aligned} \tag{5.5.25}$$

If we set $v_h = \Pi_h^* e_t$ in (5.5.24) and employ (5.5.8), (5.5.7), (5.5.9) and (5.5.22), then we have

$$\begin{aligned} & \|e_t\|_0^2 \\ \le \ & C(\|e\|_1\|e_t\|_1 + \|\rho_t\|_0\|e_t\|_0 + (h + \|u-u_h\|_0)\|e_t\|_1) \\ \le \ & C(h^{-1}\|e\|_1 + \|\rho_t\|_0 + h^{-1}(h + \|u-u_h\|_0))\|e_t\|_0. \end{aligned}$$

Thus

$$h\|e_t\|_0 \le C(\|e\|_1 + \|\rho_t\|_0 + h + \|\rho\|_0 + \|e\|_0). \tag{5.5.26}$$

Inserting (5.5.26) into (5.5.25) gives

$$\begin{aligned} & \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}(e,\Pi_h^* e) + A(u_h; e, \Pi_h^* e) \\ \le\ & C\|e\|_0(\|e\|_1 + \|\rho_t\|_0 + h + \|\rho\|_0) \\ & + C\|e\|_1(h + \|\rho\|_0 + \|e\|_0). \end{aligned}$$

This together with (5.5.20) and (5.5.21) implies

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}(e,\Pi_h^* e) + A(u_h; e, \Pi_h^* e) \le C(\|e\|_0 + h)\|e\|_1.$$

Integrate it on $t$ and use (5.5.8) and (5.5.6), then we have

$$\begin{aligned} & \|e\|_0^2 + \int_0^t \|e\|_1^2 \mathrm{d}t \\ \le\ & C\Big(\|e(0)\|_0^2 + \int_0^t (\|e\|_0 + h)\|e\|_1 \mathrm{d}t\Big) \\ \le\ & C\Big(\|e(0)\|_0^2 + \int_0^t \Big(C(\|e\|_0^2 + h^2) + \frac{1}{2C}\|e\|_1^2\Big)\mathrm{d}t\Big). \end{aligned}$$

This results in

$$\|e\|_0^2 + \int_0^t \|e\|_1^2 \mathrm{d}t \le C\Big(h^2 + \int_0^t \|e\|_0^2 \mathrm{d}t\Big).$$

So by the Gronwall's inequality we have

$$\|e\|_0^2 + \int_0^t \|e\|_1^2 \mathrm{d}t \le Ch^2.$$

Finally, this together with (5.5.20) and (5.5.21) leads to the desired result (5.5.23). This completes the proof. □

**Theorem 5.5.2** *Let $u$ and $u_h^n$ be the solutions to the problem (5.5.1) and the Crank-Nicolson fully-discrete generalized difference scheme (5.5.4) respectively, satisfying $u, u_t \in C([0,T]; C^2(\bar{\Omega}))$, $u_{tt} \in L^2((0, T); H^2(\Omega))$, and $u_{ttt} \in L^2((0,T); L^2(\Omega))$. Then we have*

$$\begin{aligned} & \max_{1\le n\le N}\Big\{\|u_h^n - u(t_n)\|_0^2 + \tau\sum_{i=1}^n \|u_h^{i-1/2} - u(t_{i-1/2})\|_1^2\Big\} \\ \le\ & C(h^2 + \tau^4). \end{aligned}$$

The proof to this theorem is omitted to save the space (cf. [A-52]).

## Bibliography and Comments

The development of the theory of generalized difference methods for parabolic equations is parallel to that for elliptic equations. Generalized difference methods for parabolic equations are proposed and discussed in [B-57]. [A-53] considers a Hermite type cubic element difference scheme for a one-dimensional parabolic equation (cf. §5.4). Discussed respectively in [A-20,21,43] are the generalized difference method and its variety—a mass concentration method for two-dimensional parabolic equations. [A-52,23,46] deal with the generalized difference methods for nonlinear parabolic equations. [A-6] is concerned with a quadratic element generalized difference scheme for a heat-transfer equation (§5.4). The extreme value property and the uniform convergence is studied in [A-58].

In some early references on generalized difference methods for parabolic equations, the proofs to the error estimates are not quite rigorous, due to a wrong presumption that the $L^2$-estimate (the dual argument) still holds for linear element generalized difference methods for elliptic equations. As regards the error estimates of semi- and fully-discrete generalized difference methods, we can borrow the theories and techniques of finite element methods to get basically parallel results. But there are certain difficulties requiring special treatments, such as the asymmetry of $(\cdot, \Pi_h^* \cdot)$. A method dealing with the asymmetry of $(\cdot, \Pi_h^* \cdot)$ is given in §5.5.

**Problem 1** Discuss the error estimates for high order element difference schemes for two-dimensional parabolic equations.

**Problem 2** Consider the generalized difference method for $u_{tt} = \Delta^2 u$ (cf. the first four sections of Chapter 4), of which the one-dimensional case has been discussed in [A-42].

# Chapter 6

# HYPERBOLIC EQUATIONS

Hyperbolic equations, especially first order hyperbolic systems, have important applications in fluid mechanics and light propagation. Due to the special properties of this class of equations, the difference method remains to be the most used method to solve them. In this chapter, we introduce an extension of the classical difference method, i.e., the generalized difference method, in particular the upwind generalized difference method.

## 6.1 Generalized Difference Methods for Second Order Hyperbolic Equations

Consider the mixed problem of second order hyperbolic equations:

$$\begin{cases} u_{tt} + Au = f(x,t), & x \in \Omega,\ 0 < t \le T, & \text{(6.1.1a)} \\ u = 0, & x \in \partial\Omega,\ 0 < t \le T, & \text{(6.1.1b)} \\ u = u_0(x),\ u_t = u_1(x), & x \in \Omega,\ t = 0, & \text{(6.1.1c)} \end{cases}$$

where $\Omega \subset R^n$ is a bounded region with a piecewise smooth boundary $\partial\Omega$; $u_{tt} = \frac{\partial^2 u}{\partial t^2}$; $A$ is a uniformly elliptic second order partial differential operator:

$$Au = -\sum_{i,j} \frac{\partial}{\partial x_i}\Big(a_{ij}\frac{\partial u}{\partial x_j}\Big),$$

where $a_{ij}(x) = a_{ji}(x)$ are sufficiently smooth. By the uniform ellipticity we mean the existence of a constant $\alpha > 0$ satisfying

$$a(u,u) \equiv \int_\Omega \Big(\sum_{i,j} a_{ij}\frac{\partial u}{\partial x_i}\frac{\partial u}{\partial x_j}\Big)\mathrm{d}x \geq \alpha|u|_1^2, \ \ \forall u \in H_0^1(\Omega). \qquad (6.1.2)$$

A variational form of (6.1.1) is: Find $u(\cdot,t) \in H_0^1(\Omega)$ $(0 < t \leq T)$ such that

$$\begin{cases} (u_{tt}, v) + a(u,v) = (f,v), \ \ \forall v \in H_0^1(\Omega), \ 0 < t \leq T, & (6.1.3a) \\ u(x,0) = u_0(x), \ u_t(x,0) = u_1(x), \ x \in \Omega, & (6.1.3b) \end{cases}$$

where $(\cdot,\cdot)$ denotes the $L^2(\Omega)$ inner product,

$$a(u,v) = \int_\Omega \Big(\sum_{i,j} a_{ij}\frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_j}\Big)\mathrm{d}x, \qquad (6.1.4)$$

and the solution to (6.1.3) is referred to as a generalized solution of (6.1.1).

### 6.1.1 Semi-discrete generalized difference scheme

For simplicity, let $\Omega$ be a planar convex polygonal region. As in the previous chapters, we place a quasi-uniform triangulation $T_h$ and a barycenter dual grid $T_h^*$ on $\Omega$, and accordingly construct a piecewise linear trial function space $U_h \subset H_0^1(\Omega)$ and a piecewise constant test function space $V_h \subset L^2(\Omega)$. Then, the semi-discrete scheme reads: Find $u_h = u_h(\cdot,t) \in U_h$ such that

$$\begin{cases} (u_{htt}, v_h) + a(u_h, v_h) = (f, v_h), & \forall v_h \in V_h, \quad (6.1.5a) \\ u_h(x,0) = u_{0h}(x), \ u_{ht}(x,0) = u_{1h}(x), & x \in \Omega, \quad (6.1.5b) \end{cases}$$

where $a(\cdot,\cdot)$ is the bilinear form defined by (6.1.4). But on $U_h \times V_h$, this is only a formal definition and calls for further explanations: It is obtained by integrating $(Au, v)$ in parts either on individual dual elements, or on the whole $\Omega$ in the sense of generalized functions (cf.

§3.1). $u_{0h}$ and $u_{1h}$ are certain approximations of $u_0(x)$ and $u_1(x)$ respectively, usually taken as their interpolation projections or $L^2$-projections into $U_h$. The latter is equivalent to replace (6.1.5b) by

$$\begin{cases} (u_h(\cdot,0),v_h)=(u_0,v_h), \\ (u_{ht}(\cdot,0),v_h)=(u_1,v_h), \end{cases} \quad \forall v_h \in V_h.$$

Let $\{\phi_j(x)\}_{j=1,2,\cdot,n}$ and $\{\psi_j(x)\}_{j=1,2,\cdot,n}$ be bases of $U_h$ and $V_h$ respectively. Then we can state (6.1.5) in the following fashion: Find an approximate solution in the form

$$u_h = \sum_{j=1}^{n} \mu_j(t)\phi_j(x)$$

such that its coefficients $\mu_1(t),\cdots,\mu_n(t)$ solve the following initial value problem of ordinary differential equations:

$$\begin{cases} \sum_{j=1}^{n}\Big[\dfrac{\mathrm{d}^2\mu_j(t)}{\mathrm{d}t^2}(\phi_j,\psi_i)+\mu_j(t)a(\phi_j,\psi_i)\Big]=(f,\psi_i), & \text{(6.1.5a)}' \\ \mu_i(0)=\alpha_i,\ \mu_{it}(0)=\beta_i, & \text{(6.1.5b)}' \end{cases}$$

where $0<t\le T$, $i=1,2,\cdots,n$, $\alpha_i$'s are the coefficients in $u_{0h}=\sum_{i=1}^{n}\alpha_i\phi_i$, and $\beta_i$'s are the coefficients in $u_{1h}=\sum_{i=1}^{n}\beta_i\phi_i$. It is easy to check that the matrix $M=\{(\phi_j,\psi_i)\}$ is symmetric and positive definite, so (6.1.5)$'$ admits a unique and smooth solution for each $f\in L^2(\Omega)$.

Now we deduce the $H^1$-estimate of the error $u-u_h$ in a way similar to that in §5.1. First let us define an elliptic projection operator $P_h$: $H^2(\Omega)\cap H_0^1(\Omega)\to U_h$ in terms of the following generalized difference equation:

$$a(P_hw,v_h)=a(w,v_h),\ \ \forall v_h\in V_h. \tag{6.1.6}$$

Recalling the results in §3.2 we have

$$\|w-P_hw\|_1\le Ch|w|_2. \tag{6.1.7}$$

Now let $u$ and $u_h$ be the solutions to (6.1.3) and (6.1.5) respectively. Then the error $(u-u_h)$ satisfies

$$(u_{tt}-u_{htt},v_h)+a(u-u_h,v_h)=0,\ \ \forall v_h\in V_h. \tag{6.1.8}$$

Set

$$\rho = u - P_h u, \; e = P_h u - u_h, \tag{6.1.9}$$

then $u - u_h = \rho + e$. We shall need the following estimates for $\rho$ and $e$ (see (6.1.7)):

$$\begin{aligned} \|\rho\|_1 &= \|u - P_h u\|_1 \le Ch\|u\|_2 \\ &\le Ch\Big(\|u_0\|_2 + \int_0^t \|u_t\|_2 \mathrm{d}t\Big), \end{aligned} \tag{6.1.10a}$$

$$\|\rho_{tt}\|_0 = \|P_h u_{tt} - u_{tt}\|_0 \le Ch\|u_{tt}\|_2, \tag{6.1.10b}$$

$$\begin{aligned} \|e(0)\|_1 &= \|P_h u_0 - u_{0h}\|_1 \\ &\le \|P_h u_0 - u_0\|_1 + \|u_0 - u_{0h}\|_1 \\ &\le Ch\|u_0\|_2 + \|u_0 - u_{0h}\|_1, \end{aligned} \tag{6.1.10c}$$

$$\begin{aligned} \|e_t(0)\|_0 &\le \|P_h u_1 - u_1\|_1 + \|u_1 - u_{1h}\|_0 \\ &\le Ch\|u_1\|_2 + \|u_1 - u_{1h}\|_0. \end{aligned} \tag{6.1.10d}$$

Rewrite (6.1.8) into

$$(e_{tt}, v_h) + (\rho_{tt}, v_h) + a(e, v_h) + a(\rho, v_h) = 0.$$

This together with (6.1.6) gives

$$(e_{tt}, v_h) + a(e, v_h) = -(\rho_{tt}, v_h), \quad \forall v_h \in V_h. \tag{6.1.11}$$

As in §5.1, let us introduce the interpolation projection operator $\Pi_h^* : H_0^1(\Omega) \to V_h$ and set $v_h = \Pi_h^* e_t$ in (6.1.11) to obtain

$$(e_{tt}, \Pi_h^* e_t) + a(e, \Pi_h^* e_t) = -(\rho_{tt}, \Pi_h^* e_t). \tag{6.1.12}$$

As in §5.1 we have

$$(u_h, \Pi_h^* \bar{u}_h) = (\bar{u}_h, \Pi_h^* u_h), \quad \forall \bar{u}_h, u_h \in U_h,$$

$$(u_h, \Pi_h^* u_h) > 0, \quad \forall u_h \neq 0.$$

Write $|||u_h|||_0 = (u_h, \Pi_h^* u_h)^{1/2}$, then the norms $|||u_h|||_0$ and $\|u_h\|_0$ are equivalent. Moreover, by the inverse property the following estimate holds:

$$|a(u_h, \Pi_h^* \bar{u}_h) - a(\bar{u}_h, \Pi_h^* u_h)| \le Ch\|u_h\|_1\|\bar{u}_h\|_1 \le C\|u_h\|_0\|\bar{u}_h\|_1,$$

$$\forall \bar{u}_h, u_h \in U_h. \tag{6.1.13}$$

(6.1.12) is equivalent to

$$\begin{aligned}&\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|||e_t|||_0^2 + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}a(e, \Pi_h^* e)\\ =\ &\frac{1}{2}[a(e_t, \Pi_h^* e) - a(e, \Pi_h^* e_t)] - (\rho_{tt}, \Pi_h^* e_t).\end{aligned}$$

It follows from (6.1.13) and $\|\Pi_h^* e_t\|_0 \le C\|e_t\|_0$ that

$$\begin{aligned}&\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|||e_t|||_0^2 + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}a(e, \Pi_h^* e)\\ \le\ &C\|e_t\|_0\|e\|_1 + \|\rho_{tt}\|_0\|\Pi_h^* e_t\|_0\\ \le\ &C[\|e_t\|_0^2 + \|e\|_1^2 + \|\rho_{tt}\|_0^2].\end{aligned}$$

Integrate it to obtain

$$\begin{aligned}&|||e_t|||_0^2 + a(e, \Pi_h^* e)\\ \le\ &|||e_t(0)|||_0^2 + a(e(0), \Pi_h^* e(0))\\ &+C\int_0^t [\|e_t\|_0^2 + \|e\|_1^2 + \|\rho_{tt}\|_0^2]\mathrm{d}t.\end{aligned} \tag{6.1.14}$$

Note

$$a(e, \Pi_h^* e) \ge \alpha\|e\|_1^2,\ \alpha > 0 \text{ a constant},$$
$$a(e(0), \Pi_h^* e(0)) \le C\|e(0)\|_1^2,$$
$$|||e_t(0)|||_0^2 \le C\|e_t(0)\|_0^2.$$

Therefore, by (6.1.10) and (6.1.14) we have

$$\begin{aligned}&\|e_t\|_0^2 + \|e\|_1^2\\ \le\ &C\Big\{\|P_h u_0 - u_{0h}\|_1^2 + \|P_h u_1 - u_{1h}\|_0^2 + h^2\int_0^t \|u_{tt}\|_2^2\mathrm{d}t\\ &+\int_0^t (\|e_t\|_0^2 + \|e\|_1^2)\mathrm{d}t\Big\}.\end{aligned}$$

So it follows from the Gronwall's inequality that

$$\|e\|_1^2 \le C\Big\{\|P_h u_0 - u_{0h}\|_1^2 + \|P_h u_1 - u_{1h}\|_0^2 + h^2\int_0^t \|u_{tt}\|_2^2\mathrm{d}t\Big\}.$$

Combining this with (6.1.10a) yields the error estimate:

$$\begin{aligned}\|u-u_h\|_1^2 \ &\leq C\Big\{\|P_hu_0-u_{0h}\|_1^2+\|P_hu_1-u_{1h}\|_0^2 \\ &+h^2\Big[\|u_0\|_2^2+\int_0^t\|u_t\|_2^2\mathrm{d}t+\int_0^t\|u_{tt}\|_2^2\mathrm{d}t\Big]\Big\}.\end{aligned} \tag{6.1.15}$$

### 6.1.2 Fully-discrete generalized difference scheme

Now we further discretize time $t$ of the semi-discrete difference scheme (6.1.5) to deduce fully-discrete schemes. Let the time step size be $\tau$ and $t_n = n\tau$ ($n = 0, 1, \cdots, N$; $N\tau = T$), $u_h^n = u_h(t_n)$. For a function $v$ well-defined at times $t = t_n$ ($n = 0, 1, \cdots, N$), we shall use the following symbols:

$$v^n = v|_{t=t_n},\ v^{n+1/2} = \frac{v^n + v^{n+1}}{2},\ \hat{\partial}_t v^{n+1/2} = \frac{v^{n+1}-v^n}{\tau},$$

$$v^{n,1/4} = \frac{1}{4}(v^{n+1}+2v^n+v^{n-1}) = \frac{1}{2}(v^{n+1/2}+v^{n-1/2}),$$

$$\hat{\partial}_t v^n = \frac{v^{n+1}-v^{n-1}}{2\tau} = \frac{1}{\tau}(v^{n+1/2}-v^{n-1/2}) = \frac{1}{2}(\hat{\partial}_t v^{n+1/2}+\hat{\partial}_t v^{n-1/2}),$$

$$\partial_{tt}v^n = \frac{v^{n+1}-2v^n+v^{n-1}}{\tau^2} = \frac{1}{\tau}(\hat{\partial}_t v^{n+1/2}-\hat{\partial}_t v^{n-1/2}).$$

Now, we use weighted averages of the values of $u_h$ and $f$ at $t_{n-1}$, $t_n$ and $t_{n+1}$ to construct the following fully-discrete generalized difference scheme:

$$(\partial_{tt}u_h^n, v_h) + a(u_h^{n,1/4}, v_h) = (f^{n,1/4}, v_h),\ \ \forall v_h \in V_h. \tag{6.1.16}$$

This is an implicit scheme, being absolutely stable as shown below. For readers familiar with finite difference methods, it is not difficult to recall the counterpart of (6.1.16) in finite difference methods. (cf. [A-27].)

Let us deduce the convergence estimate. Assume $u$ is a smooth solution of the continuous problem (6.1.3). By the Taylor expansion we have

$$u_{tt}^{n,1/4} = \partial_{tt}u^n - r_n, \tag{6.1.17a}$$

where the remainder $r_n$ satisfies the following estimate (cf. [B-27])

$$\|r_n\|_0^2 \le C\tau^3 \int_{t_{n-1}}^{t_{n+1}} \left\|\frac{\partial^4 u}{\partial t^4}\right\|_0^2 \mathrm{d}t. \tag{6.1.17b}$$

So by (6.1.3a) we have

$$(\partial_{tt}u^n - r_n, v) + a(u^{n,1/4}, v) = (f^{n,1/4}, v),$$

which gives by setting $v = v_h$ that

$$(\partial_{tt}u^n, v_h) + a(u^{n,1/4}, v_h) = (r_n, v_h) + (f^{n,1/4}, v_h).$$

Subtracting it with (6.1.16) yields the error equation

$$(\partial_{tt}(u^n - u_h^n), v_h) + a(u^{n,1/4} - u_h^{n,1/4}, v_h) = (r_n, v_h). \tag{6.1.18}$$

Set

$$u^n - u_h^n = (u^n - P_h u^n) + (P_h u^n - u_h^n) = \rho^n + e^n.$$

Then by (6.1.6) and (6.1.18) we have

$$(\partial_{tt}e^n, v_h) + a(e^{n,1/4}, v_h) = (r_n - \partial_{tt}\rho^n, v_h).$$

If in particular we choose $v_h = \Pi_h^* \hat{\partial}_t e^n$, then

$$(\partial_{tt}e^n, \Pi_h^*\hat{\partial}_t e^n) + a(e^{n,1/4}, \Pi_h^*\hat{\partial}_t e^n) = (r_n - \partial_{tt}\rho^n, \Pi_h^*\hat{\partial}_t e^n). \tag{6.1.19}$$

Let us deal with respectively these terms in the above equality. For the first term on the left-hand side we have

$$\begin{aligned}
&(\partial_{tt}e^n, \Pi_h^*\hat{\partial}_t e^n)\\
=\ & \frac{1}{2\tau}((e^{n+1} - 2e^n + e^{n-1})\tau^{-1}, \Pi_h^*(e^{n+1} - e^{n-1})\tau^{-1})\\
=\ & \frac{1}{2\tau}(\hat{\partial}_t e^{n+1/2} - \hat{\partial}_t e^{n-1/2}, \Pi_h^*(\hat{\partial}_t e^{n+1/2} + \hat{\partial}_t e^{n-1/2}))\\
=\ & \frac{1}{2\tau}[(\hat{\partial}_t e^{n+1/2}, \Pi_h^*\hat{\partial}_t e^{n+1/2}) - (\hat{\partial}_t e^{n-1/2}, \Pi_h^*\hat{\partial}_t e^{n-1/2})]\\
=\ & \frac{1}{2\tau}[|||\hat{\partial}_t e^{n+1/2}|||_0^2 - |||\hat{\partial}_t e^{n-1/2}|||_0^2].
\end{aligned}$$

For the second term on the left-hand side of (6.1.19) we have

$$\begin{aligned}
& a(e^{n,1/4}, \Pi_h^* \hat{\partial}_t e^n) \\
= \; & \frac{1}{2\tau}[a(e^{n+1/2}, \Pi_h^* e^{n+1/2}) - a(e^{n-1/2}, \Pi_h^* e^{n-1/2})] \\
& -\frac{1}{2\tau}[a(e^{n+1/2}, \Pi_h^* e^{n-1/2}) - a(e^{n-1/2}, \Pi_h^* e^{n+1/2})].
\end{aligned}$$

It follows from (6.1.13) that

$$\begin{aligned}
& \frac{1}{2\tau}|a(e^{n+1/2}, \Pi_h^* e^{n-1/2}) - a(e^{n-1/2}, \Pi_h^* e^{n+1/2})| \\
= \; & \frac{1}{2\tau}|a(e^{n+1/2} - e^{n-1/2}, \Pi_h^* e^{n-1/2}) \\
& -a(e^{n-1/2}, \Pi_h^*(e^{n+1/2} - e^{n-1/2}))| \\
= \; & \frac{1}{2}|a(\hat{\partial}_t e^n, \Pi_h^* e^{n-1/2}) - a(e^{n-1/2}, \Pi_h^* \hat{\partial}_t e^n)| \\
\leq \; & C\|e^{n-1/2}\|_1 \|\hat{\partial}_t e^n\|_0 \\
\leq \; & C(\|e^{n-1/2}\|_1^2 + \|\hat{\partial}_t e^{n+1/2}\|_0^2 + \|\hat{\partial}_t e^{n-1/2}\|_0^2).
\end{aligned}$$

For the right-hand side of (6.1.19) we have

$$\begin{aligned}
& |(r_n - \partial_{tt}\rho^n, \Pi_h^* \hat{\partial}_t e^n)| \\
\leq \; & \|r_n\|_0^2 + \|\partial_{tt}\rho^n\|_0^2 + \frac{1}{2}\|\Pi_h^* \hat{\partial}_t e^{n+1/2}\|_0^2 + \frac{1}{2}\|\Pi_h^* \hat{\partial}_t e^{n-1/2}\|_0^2.
\end{aligned}$$

Hence, (6.1.19) results in

$$\begin{aligned}
& \frac{1}{2\tau}[|||\hat{\partial}_t e^{n+1/2}|||_0^2 - |||\hat{\partial}_t e^{n-1/2}|||_0^2] \\
& +\frac{1}{2\tau}[a(e^{n+1/2}, \Pi_h^* e^{n+1/2}) - a(e^{n-1/2}, \Pi_h^* e^{n-1/2})] \\
\leq \; & C\{\|e^{n-1/2}\|_1^2 + \|\hat{\partial}_t e^{n+1/2}\|_0^2 + \|\hat{\partial}_t e^{n-1/2}\|_0^2 + \|r_n\|_0^2 + \|\partial_{tt}\rho^n\|_0^2 \\
& +\|\Pi_h^* \hat{\partial}_t e^{n+1/2}\|_0^2 + \|\Pi_h^* \hat{\partial}_t e^{n-1/2}\|_0^2\}.
\end{aligned}$$

Multiply it by $2\tau$, and sum it over $n = 1, 2, \cdots, N-1$ to obtain

$$\begin{aligned}
& |||\hat{\partial}_t e^{N-1/2}|||_0^2 + a(e^{N-1/2}, \Pi_h^* e^{N-1/2}) \\
\leq\ & |||\hat{\partial}_t e^{1/2}|||_0^2 + a(e^{1/2}, \Pi_h^* e^{1/2}) \\
& + C\tau \sum_{n=1}^{N-1} [\|e^{n-1/2}\|_1^2 + \|\hat{\partial}_t e^{n+1/2}\|_0^2 + \|\hat{\partial}_t e^{n-1/2}\|_0^2 \\
& + \|r_n\|_0^2 + \|\partial_{tt}\rho^n\|_0^2 + \|\Pi_h^* \hat{\partial}_t e^{n+1/2}\|_0^2 + \|\Pi_h^* \hat{\partial}_t e^{n-1/2}\|_0^2].
\end{aligned} \tag{6.1.20}$$

Notice (6.1.17b) and

$$a(e^{n-1/2}, \Pi_h^* e^{n-1/2}) \geq \alpha \|e^{n-1/2}\|_1^2,$$

$$\begin{aligned}
& \sum_{n=1}^{N-1} \|\partial_{tt}\rho^n\|_0^2 = \frac{1}{\tau^2} \sum_{n=1}^{N-1} \left\| \int_{t^n}^{t^{n+1}} \int_{t-\tau}^{t} \rho_{tt}(s) \mathrm{d}s \mathrm{d}t \right\|_0^2 \\
\leq\ & \frac{1}{\tau} \sum_{n=1}^{N-1} \int_{t^{n-1}}^{t^{n+1}} \|\rho_{tt}\|_0^2 \mathrm{d}t \leq \frac{2}{\tau} \int_0^T \|\rho_{tt}\|_0^2 \mathrm{d}t \leq \frac{Ch^2}{\tau} \int_0^T \|u_{tt}\|_2^2 \mathrm{d}t.
\end{aligned}$$

Also note the equivalence of the norms $|||\cdot|||_0$ and $\|\cdot\|_0$ on $U_h$. Then, (6.1.20) leads to

$$\begin{aligned}
& \|\hat{\partial}_t e^{N-1/2}\|_0^2 + \|e^{N-1/2}\|_1^2 \\
\leq\ & C\Big\{ \|\hat{\partial}_t e^{1/2}\|_0^2 + \|e^{1/2}\|_1^2 + \tau^4 \int_0^T \|u_{tttt}\|_0^2 \mathrm{d}t \\
& + h^2 \int_0^T \|u_{tt}\|_2^2 \mathrm{d}t + \tau \sum_{n=1}^{N} (\|\hat{\partial}_t e^{n-1/2}\|_0^2 + \|e^{n-1/2}\|_1^2) \Big\}.
\end{aligned}$$

Finally, by virtue of Gronwall's theorem we have

$$\begin{aligned}
& \|\hat{\partial}_t e^{N-1/2}\|_0^2 + \|e^{N-1/2}\|_1^2 \\
\leq\ & C\Big\{ \|\hat{\partial}_t e^{1/2}\|_0^2 + \|e^{1/2}\|_1^2 + \tau^4 \int_0^T \|u_{tttt}\|_0^2 \mathrm{d}t \\
& + h^2 \int_0^T \|u_{tt}\|_2^2 \mathrm{d}t \Big\}.
\end{aligned}$$

This together with (6.1.10) validates the following error estimate for the fully-discrete scheme.

**Theorem 6.1.1** *Let $u$ and $u_h^n$ be the solutions to (6.1.3) and (6.1.16) respectively. Then the following error estimate holds:*

$$\begin{aligned}
& \|u(t_{n+1/2}) - u_h^{n+1/2}\|_1^2 \\
\leq\ & C\Big\{\|(P_hu - u_h)^{1/2}\|_1^2 + \|\hat{\partial}_t(P_hu - u_h)^{1/2}\|_0^2 + \tau^4\int_0^T \|u_{tttt}\|_0^2 dt \\
& + h^2\Big(\|u_0\|_2^2 + \int_0^T \|u_t\|_2^2 dt + \int_0^T \|u_{tt}\|_2^2 dt\Big)\Big\}.
\end{aligned} \tag{6.1.21}$$

**Remark** For the first term in the right-hand side of the inequality (6.1.21), we have

$$\begin{aligned}
& (P_hu - u_h)^{1/2} \\
=\ & \frac{1}{2}(P_hu^0 - u_h^0) + \frac{1}{2}(P_hu^1 - u_h^1) \\
=\ & \frac{1}{2}(P_hu_0 - u_{0h}) + \frac{1}{2}P_h(u^1 - u^0) + \frac{1}{2}P_h(u^0 - u_h^1) \\
=\ & \frac{1}{2}P_h(u_0 - u_{0h}) + \frac{\tau}{2}P_h(u^1 - u^0)/\tau - \frac{\tau}{2}P_h(u_h^1 - u^0)/\tau.
\end{aligned}$$

The accuracy of the first and third terms above is determined by the choices of $u_{0h}$ and $u_h^1$, and the second term is of order $O(\tau)$ thanks to the smoothness of the solution $u$.

## 6.2 Generalized Upwind Schemes for First Order Hyperbolic Equations

The classical upwind scheme occupies a very important position in the approximation of first order hyperbolic equations, due to its nice stability and monotonicity. But this scheme has only first order accuracy and suits solely rectangular grids. In this and the next sections, we construct a class of accurate generalized upwind schemes on irregular networks, including the classical upwind scheme as a special case.

### 6.2.1 Generalized Upwind Schemes

Let $\Omega \subset R^2$ be a polygonal region with boundary $\partial\Omega$. Corresponding to a vector function $a = a(x) \in R^2$, $x = (x_1, x_2) \in \bar{\Omega}$, we divide $\partial\Omega$ into two parts

$$\begin{cases} (\partial\Omega)_- = \{x \in \partial\Omega : a \cdot \nu \leq 0\} & \text{(flow in)} \quad (6.2.1a) \\ (\partial\Omega)_+ = \{x \in \partial\Omega : a \cdot \nu > 0\} & \text{(flow out)} \quad (6.2.1b) \end{cases}$$

where $\nu$ stands for the unit outer normal vector of $\partial\Omega$.

Consider a mixed problem of first order partial differential equations:

$$\begin{cases} \dfrac{\partial u(x,t)}{\partial t} + a \cdot \nabla u(x,t) + \sigma(x,t)u(x,t) = f(x,t), \\ \qquad (x,t) \in \Omega \times [0,T], & (6.2.2a) \\ u(x,t) = 0, \ (x,t) \in (\partial\Omega)_- \times [0,T], & (6.2.2b) \\ u(x,0) = \phi(x), \ x \in \overline{\Omega}, & (6.2.2c) \end{cases}$$

where $\nabla = \left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}\right)$; $\sigma$, $f$ and $\phi$ are smooth (scalar) functions; and $a$ is a vector function. If these functions are sufficiently smooth, the problem (6.2.2) has a unique and smooth solution (See [B-53]).

As in §3.2, we place a quasi-uniform triangulation $T_h = \{K\}$ of $\Omega$. Choose $T_h^*$ to be the barycenter dual grid relative to $T_h$. Let $P_0$ be a node of $T_h$ (cf. Fig. 3.2.1) with neighbouring nodes $P_i$ $(1 \leq i \leq 6)$, $M_i$ the midpoint of $\overline{P_0P_i}$, $Q_i$ the circumcenter of $\triangle P_0P_iP_{i+1}$ $(1 \leq i \leq 6$ and $P_7 = P_1)$, and $K_{P_0}^*$ the dual element surrounding $P_0$.

Recalling that $\mathcal{P}_r$ is the polynomial family of degree $r$, let us construct a finite element space

$$V_h = \{v_h : v_h|_{K^*} \in \mathcal{P}_r, \ \forall K^* \subset T_h^*; \ v_h = 0, \text{ on } K_{P_0}^* \text{ for } P_0 \in (\partial\Omega)_-\}.$$

Its basis functions for an interior node $P_0 = (x_1^{(0)}, x_2^{(0)})$ of $T_h$ are taken as

$$v_{P_0}(x) = \begin{cases} \dfrac{1}{l!(m-l)!}(x_1 - x_1^{(0)})^l (x_2 - x_2^{(0)})^{m-l}, & x \in K_{P_0}^*, \\ 0, & \text{elsewhere}, \end{cases} \tag{6.2.3}$$

$$0 \le l \le m,\ 0 \le m \le r.$$

Due to the discontinuity of $V_h$ on the boundaries of the dual elements, one can not apply the Galerkin finite element method on the entire region $\Omega$. But it is feasible to apply it on a single dual element $K^*_{P_0}$. So we seek $u_h(\cdot, t) \in V_h$ satisfying

$$\int_{K^*_{P_0}} \Big[\frac{\partial u_h}{\partial t} + a \cdot \nabla u_h + \sigma u_h\Big] v_h \mathrm{d}x = \int_{K^*_{P_0}} f v_h \mathrm{d}x,\ v_h \in V_h. \tag{6.2.4}$$

Denote by $\nu$ the unit outer normal vector of $\partial K^*_{P_0}$, and employ Green's formula

$$\begin{aligned} &\int_{K^*_{P_0}} (a \cdot \nabla u_h) v_h \mathrm{d}x \\ = &-\int_{K^*_{P_0}} u_h \mathrm{div}(a v_h) \mathrm{d}x + \int_{\partial K^*_{P_0}} (a \cdot \nu) u_h v_h \mathrm{d}s, \end{aligned} \tag{6.2.5}$$

then we can rewrite (6.2.4) as

$$\begin{aligned} &\int_{K^*_{P_0}} \frac{\partial u_h}{\partial t} v_h \mathrm{d}x - \int_{K^*_{P_0}} u_h \mathrm{div}(a v_h) \mathrm{d}x \\ &+ \int_{K^*_{P_0}} \sigma u_h v_h \mathrm{d}x + \int_{\partial K^*_{P_0}} (a \cdot \nu) u_h v_h \mathrm{d}s \\ = &\int_{K^*_{P_0}} f v_h \mathrm{d}x,\ v_h \in V_h. \end{aligned} \tag{6.2.6}$$

Similarly as in (6.2.1), we can define $(\partial K^*_{P_0})_-$ and $(\partial K^*_{P_0})_+$. For $x \in \partial K^*_{P_0}$, set

$$u_h^+(x) = \begin{cases} \lim\limits_{\substack{x' \to x \\ x' \notin K^*_{P_0}}} u_h(x'), & \text{when } x \in (\partial K^*_{P_0})_-, \\ \lim\limits_{\substack{x' \to x \\ x' \in K^*_{P_0}}} u_h(x'), & \text{when } x \in (\partial K^*_{P_0})_+, \end{cases}$$

$$u_h^-(x) = \begin{cases} \lim\limits_{\substack{x' \to x \\ x' \in K^*_{P_0}}} u_h(x'), & \text{when } x \in (\partial K^*_{P_0})_-, \\ \lim\limits_{\substack{x' \to x \\ x' \notin K^*_{P_0}}} u_h(x'), & \text{when } x \in (\partial K^*_{P_0})_+. \end{cases}$$

They are referred to as the upwind and the downwind values of $u_h(x)$ at $x \in \partial K^*_{P_0}$, respectively. On the analogy of the classical upwind scheme, we replace $u_h(x)$ in the line integral of the left-hand side of (6.2.6) by $u_h^+$ to obtain

$$\begin{aligned}&\int_{K^*_{P_0}} \frac{\partial u_h}{\partial t} v_h \mathrm{d}x - \int_{K^*_{P_0}} u_h \mathrm{div}(a v_h)\mathrm{d}x + \int_{K^*_{P_0}} \sigma u_h v_h \mathrm{d}x \\ &+ \int_{\partial K^*_{P_0}} (a\cdot\nu) u_h^+ v_h \mathrm{d}s = \int_{K^*_{P_0}} f v_h \mathrm{d}x.\end{aligned} \tag{6.2.7}$$

It follows from (6.2.5) that

$$\begin{aligned}&-\int_{K^*_{P_0}} u_h \mathrm{div}(a v_h)\mathrm{d}x \\ =\ &\int_{K^*_{P_0}} (a\cdot\nabla u_h) v_h \mathrm{d}x - \int_{\partial K^*_{P_0}} (a\cdot\nu) u_h v_h \mathrm{d}s \\ =\ &\int_{K^*_{P_0}} (a\cdot\nabla u_h) v_h \mathrm{d}x - \int_{(\partial K^*_{P_0})_+} (a\cdot\nu) u_h^+ v_h \mathrm{d}s \\ &- \int_{(\partial K^*_{P_0})_-} (a\cdot\nu) u_h^- v_h \mathrm{d}s.\end{aligned}$$

Substituting it in (6.2.7) yields a semi-discrete upwind scheme:

$$\begin{aligned}&\int_{K^*_{P_0}} \frac{\partial u_h}{\partial t} v_h \mathrm{d}x + \int_{K^*_{P_0}} (a\cdot\nabla u_h) v_h \mathrm{d}x \\ &+ \int_{K^*_{P_0}} \sigma u_h v_h \mathrm{d}x + \int_{(\partial K^*_{P_0})_-} (a\cdot\nu)[u_h] v_h \mathrm{d}s \\ =\ &\int_{K^*_{P_0}} f v_h \mathrm{d}x,\ v_h \in V_h,\end{aligned} \tag{6.2.8a}$$

where $[u_h] = u_h^+ - u_h^-$ is the jump of $u_h$ across $(\partial K^*_{P_0})_-$. The initial and boundary value conditions are

$$\begin{cases} u_h(x,t) = 0, & x \in (\partial\Omega)_-, \quad (6.2.8\text{b}) \\ u_h(x,0) = \phi_h(x), & x \in \Omega, \quad (6.2.8\text{c}) \end{cases}$$

where $\phi_h(x)$ is a certain approximation of $\phi(x)$.

Equation (6.2.8a) can also be expressed in a symmetric form:

$$\begin{aligned}&\int_{K^*_{P_0}} \frac{\partial u_h}{\partial t} v_h \mathrm{d}x + \int_{K^*_{P_0}} (a\cdot\nabla u_h)v_h\mathrm{d}x \\ &+ \int_{K^*_{P_0}} \sigma u_h v_h \mathrm{d}x - \frac{1}{2}\int_{\partial K^*_{P_0}} (|a\cdot\nu| - a\cdot\nu)[u_h]v_h\mathrm{d}s \\ &= \int_{K^*_{P_0}} f v_h \mathrm{d}x,\ v_h\in V_h.\end{aligned} \tag{6.2.9}$$

Various kinds of finite difference quotients can be used to further discretize the time derivative $\frac{\partial u_h}{\partial t}$, such as forward difference, backward difference, or Crank-Nicolson difference. It will be illustrated in Section 3 below that our scheme here leads to a classical upwind scheme if the space dimension is one and $V_h$ consists of step functions. If on the other hand, $V_h$ consists of piecewise high degree ($> 1$) polynomials, then the convergence rate of the approximate solutions increase accordingly, resulting in highly accurate upwind schemes.

### 6.2.2 Semi-discrete error estimate

In the equation (6.2.2a), we may assume without loss of generality that $\bar{\sigma} = \sigma - \frac{1}{2}\mathrm{div}a \geq \sigma_0 > 0$. In fact, otherwise we only have to perform the transformation

$$\bar{u} = ue^{-\omega t},$$

$$\omega = \sigma_0 + \sup_{(x,t)\in\Omega\times[0,T]} |\sigma(x,t)| + \frac{1}{2}\sup_{x\in\Omega}|\mathrm{div}a(x)|$$

to validate this assumption. Now, we define a bilinear form

$$\begin{aligned}a(u,v) = &\sum_{P_0}\Big[\int_{K^*_{P_0}} (a\cdot\nabla u)v\mathrm{d}x \\ &+ \int_{(\partial K^*_{P_0})_-} (a\cdot\nu)[u]v\mathrm{d}s + \int_{K^*_{P_0}} \sigma uv\mathrm{d}x\Big],\end{aligned} \tag{6.2.10}$$

where $\sum\limits_{P_0}$ denotes the sum over all the interior nodes $P_0$ of $T_h$ and the boundary nodes on $(\partial\Omega)_+$. If the solution of (6.2.10) $u = u(\cdot, t) \in C^2((0,T); H^{r+1}(\Omega))$ for some $r > 0$, then the imbedding theorem guarantees that the jump of $u$ across the inner boundaries $[u] = u^+ - u^- = 0$. Therefore, we may write (6.2.10) in a equivalent form:

$$(u_t, v) + a(u,v) = (f,v), \quad \forall v \in L^2(\Omega). \tag{6.2.11}$$

In terms of $a(u,v)$ the semi-discrete scheme (6.2.8a) can be written as

$$\sum_{P_0}\int_{K^*_{P_0}} \frac{\partial u_h}{\partial t} v_h \mathrm{d}x + a(u_h, v_h) = \sum_{P_0}\int_{K^*_{P_0}} f v_h \mathrm{d}x, \quad \forall v_h \in V_h. \tag{6.2.12}$$

By means of Green's formula and

$$\mathrm{div}(a v_h) = v_h \mathrm{div} a + a \cdot \nabla v_h,$$

we have

$$\int_{\partial K^*_{P_0}} (a\cdot\nu) v_h^2 \mathrm{d}s = 2\int_{K^*_{P_0}} v_h (a\cdot\nabla v_h)\mathrm{d}x + \int_{K^*_{P_0}} v_h^2 \mathrm{div} a \mathrm{d}x.$$

Hence

$$\begin{aligned}
& a(v_h, v_h) \\
= \; & \sum_{P_0}\Big[\frac{1}{2}\int_{\partial K^*_{P_0}} (a\cdot\nu) v_h^2 \mathrm{d}s - \frac{1}{2}\int_{K^*_{P_0}} v_h^2 \mathrm{div} a \mathrm{d}x \\
& + \int_{(\partial K^*_{P_0})_-} (a\cdot\nu)[v_h] v_h \mathrm{d}s + \int_{K^*_{P_0}} \sigma v_h^2 \mathrm{d}x\Big] \\
= \; & \sum_{P_0}\Big[\frac{1}{2}\int_{(\partial K^*_{P_0})_+} (a\cdot\nu) v_h^2 \mathrm{d}s + \frac{1}{2}\int_{(\partial K^*_{P_0})_-} (a\cdot\nu) v_h^2 \mathrm{d}s \\
& + \int_{(\partial K^*_{P_0})_-} (a\cdot\nu)(v_h^+ - v_h^-) v_h \mathrm{d}s \\
& + \frac{1}{2}\int_{(\partial K^*_{P_0})_-} (a\cdot\nu)[(v_h^+)^2 - 2v_h^+ v_h^- + (v_h^-)^2]\mathrm{d}s \\
& - \frac{1}{2}\int_{(\partial K^*_{P_0})_-} (a\cdot\nu)[v_h]^2 \mathrm{d}s + \int_{K^*_{P_0}} \bar{\sigma} v_h^2 \mathrm{d}x\Big].
\end{aligned}$$

Notice

$$v_h|_{(\partial K^*_{P_0})_+} = v_h^+|_{\partial K^*_{P_0}},\ v_h|_{(\partial K^*_{P_0})_-} = v_h^-|_{\partial K^*_{P_0}}.$$

Thus,

$$\begin{aligned} & a(v_h, v_h) \\ = \ & \frac{1}{2}\sum_{P_0}\Big[\int_{(\partial K^*_{P_0})_-}(a\cdot\nu)(v_h^+)^2\mathrm{d}s + \int_{(\partial K^*_{P_0})_+}(a\cdot\nu)(v_h^+)^2\mathrm{d}s \\ & -\int_{(\partial K^*_{P_0})_-}(a\cdot\nu)[v_h]^2\mathrm{d}s + 2\int_{K^*_{P_0}}\bar\sigma v_h^2\mathrm{d}x\Big]. \end{aligned} \tag{6.2.13}$$

On the common side of $K^*_{P_0}$ and $K^*_{P_1}$

$$(\partial K^*_{P_0})_+ = (\partial K^*_{P_1})_-,\ (\partial K^*_{P_0})_- = (\partial K^*_{P_1})_+.$$

Hence the first and second terms on the right-hand side of (6.2.13) cancel out each other on the inner boundaries of the elements, resulting in

$$\begin{aligned} & a(v_h, v_h) \\ = \ & -\frac{1}{2}\int_{(\partial\Omega)_-}|a\cdot\nu|(v_h^+)^2\mathrm{d}s + \frac{1}{2}\int_{(\partial\Omega)_+}|a\cdot\nu|(v_h^+)^2\mathrm{d}s \\ & +\frac{1}{2}\sum_{P_0}\int_{(\partial K^*_{P_0})_-}|a\cdot\nu|[v_h]^2\mathrm{d}s + \sum_{P_0}\int_{K^*_{P_0}}\bar\sigma v_h^2\mathrm{d}x. \end{aligned} \tag{6.2.14}$$

But $v_h^+|_{(\partial\Omega)_-} = 0$, so $a(v_h, v_h)$ is positive definite:

$$a(v_h, v_h) \geq \gamma_0(\|v_h\|_0^2 + \|v_h\|_{\partial\Omega}^2), \tag{6.2.15}$$

where $\gamma_0 = \min(\sigma_0, \frac{1}{2})$, $\|v_h\|_0^2 = (v_h, v_h)$, and

$$\|v_h\|_{\partial\Omega}^2 = \sum_{P_0}\int_{(\partial K^*_{P_0})_-}|a\cdot\nu|[v_h]^2\mathrm{d}s + \int_{(\partial\Omega)_+}|a\cdot\nu|(v_h^+)^2\mathrm{d}s.$$

It is an easy matter to show the stability of the semi-discrete scheme (6.2.8) by means of the positive definiteness of $a(v_h, v_h)$ (cf.

[B-60]). Now, let us define for $u \in H^1(\Omega)$ the Ritz projection $R_h u \in V_h$ determined by the equation

$$a(R_h u, v_h) = a(u, v_h), \quad \forall v_h \in V_h. \tag{6.2.16}$$

Since $a(v_h, v_h)$ is positive definite, the Ritz projection exists and is unique. Moreover, if $u(\cdot, t) \in H^{r+1}(\Omega)$ $(r \geq 0)$, then there holds the following estimates (cf. [B-60]):

$$\|u - R_h u\|_0^2 + \|u - R_h u\|_{\partial\Omega}^2 \leq Ch\|u\|_1^2, \text{ when } r = 0, \tag{6.2.17}$$

$$|||u - R_h u||| \leq Ch^{r+1/2}\|u\|_{r+1}, \text{ when } r \geq 1, \tag{6.2.18}$$

where $|||\cdot|||$ is defined by

$$|||v|||^2 = \|v\|_0^2 + \|v\|_{\partial\Omega} + h\sum_{P_0}\int_{K^*_{P_0}} (a\cdot\nabla v)^2 \mathrm{d}x. \tag{6.2.19}$$

Using the arguments in §5.1 we can prove the following error estimate for the semi-discrete solution $u_h(t)$ (cf. [B-60]):

$$\begin{aligned}&\|u - u_h\|_0^2 + \int_0^t \|u - u_h\|_{\partial\Omega}^2 \mathrm{d}t \\ \leq\ & C\Big\{\|\phi - \phi_h\|_0^2 + h^{2r+1}\Big[\|\phi\|_{r+1}^2 + \|u\|_{r+1}^2 + \int_0^t \|u_t\|_{r+1}^2 \mathrm{d}t\Big]\Big\},\end{aligned} \tag{6.2.20}$$

where $\|v\|_{r+1}$ stands for the $H^{r+1}(\Omega)$ norm of $v(\cdot, t)$.

### 6.2.3 Fully-discrete error estimates

For sake of simplicity, we assume $\sigma(x,t) = 0$. So we consider the hyperbolic equation:

$$\frac{\partial u}{\partial t} + a(x)\cdot\nabla u = f(x,t), \ (x,t) \in \Omega \times (0,T],$$

subject to the initial and boundary conditions (6.2.1b) and (6.2.1c). The semi-discrete upwind scheme is (6.2.12), but now

$$\begin{aligned}&a(u_h, v_h) \\ =\ & \sum_{P_0}\Big[\int_{K^*_{P_0}} (a\cdot\nabla u_h)v_h \mathrm{d}x + \int_{(\partial K^*_{P_0})_-} (a\cdot\nu)[u_h]v_h \mathrm{d}s\Big].\end{aligned} \tag{6.2.21}$$

Take a time step size $\tau > 0$, and write $u_h^n = u_h(x, t_n)$, $f^n = f(x, t_n)$, $t_n = n\tau$. Using the backward differencing on the time direction yields a backward differencing implicit scheme:

$$\int_\Omega u_h^n v_h \mathrm{d}x + \tau a(u_h^n, v_h) = \int_\Omega (u_h^{n-1} + \tau f^n) v_h \mathrm{d}x, \quad \forall v_h \in V_h. \tag{6.2.22}$$

**Theorem 6.2.1** *Let $u$ and $u_h^n$ be the solutions to (6.2.1) and (6.2.22) respectively, satisfying $u_t \in H^{r+1}(\Omega)$, $u_{tt} \in L^2(\Omega)$, $u(x,0) = \phi(x) \in H^{r+1}(\Omega)$, and $u_h^0(x) = \phi_h(x) \in V_h$. Then there holds the following error estimate:*

$$\begin{aligned} &\|u(t_n) - u_h^n\|_0 \\ \le\ &\|\phi - \phi_h\|_0 + Ch^{r+1/2}\|\phi\|_{r+1} + \tau \int_0^{t_n} \|u_{tt}(t)\|_0 \mathrm{d}t \\ &+Ch^{r+1/2} \int_0^{t_n} \|u_t(t)\|_{r+1} \mathrm{d}t. \end{aligned} \tag{6.2.23}$$

**Proof** Note $u_h^n - u(t_n) = \rho^n + e^n$, where

$$\rho^n = R_h u(t_n) - u(t_n), \ e^n = u_h^n - R_h u(t_n).$$

It follows from (6.2.18) that

$$\|\rho^n\| \le Ch^{r+1/2}\|u(t_n)\|_{r+1}.$$

Also observe that

$$u(t_n) = u(0) + \int_0^{t_n} u_t(t) \mathrm{d}t,$$

$$\|u(t_n)\|_{r+1} \le \|u(0)\|_{r+1} + \int_0^{t_n} \|u_t(t)\|_{r+1} \mathrm{d}t.$$

Thus

$$\|\rho^n\|_0 \le Ch^{r+1/2}\Big(\|\phi\|_{r+1} + \int_0^{t_n} \|u_t(t)\|_{r+1} \mathrm{d}t\Big). \tag{6.2.24}$$

Next, we turn to deal with $e^n$. Write $\bar{\partial}_t u_h^n = (u_h^n - u_h^{n-1})/\tau$. It follows from (6.2.22) and the definition of $R_h$ that

$$\begin{aligned} & a(e^n, v_h) \\ = \ & a(u_h^n - R_h u(t_n), v_h) \\ = \ & -(\bar{\partial}_t u_h^n, v_h) + (f^n, v_h) - a(u(t_n), v_h) \\ = \ & (u_t(t_n) - \bar{\partial}_t u_h^n, v_h). \end{aligned}$$

Hence

$$\begin{aligned} & (\bar{\partial}_t e^n, e^n) + a(e^n, e^n) \\ = \ & (u_t(t_n) - R_h \bar{\partial}_t u(t_n), e^n) \\ = \ & (w_1^n + w_2^n, e^n), \end{aligned} \tag{6.2.25}$$

where

$$w_1^n = u_t(t_n) - \bar{\partial}_t u(t_n), \ w_2^n = \bar{\partial}_t u(t_n) - R_h \bar{\partial}_t u(t_n).$$

Notice

$$(e^n)^+|_{(\partial\Omega)_-} = 0, \ a(e^n, e^n) \geq 0.$$

So by (6.2.25) we have

$$\begin{aligned} \|e^n\|_0 & \leq \|e^{n-1}\|_0 + \tau \|w_1^n + w_2^n\|_0 \\ & \leq \|e^0\|_0 + \tau \sum_{j=1}^{n} \|w_1^j + w_2^j\|_0. \end{aligned} \tag{6.2.26}$$

It is obvious that

$$\begin{aligned} \|e^0\|_0 & = \|u_h^0 - R_h u(0)\|_0 \\ & \leq \|u_h^0 - u(x,0)\|_0 + \|u(x,0) - R_h u(x,0)\|_0 \\ & \leq \|\phi - \phi_h\|_0 + C h^{r+1/2} \|\phi\|_{r+1}. \end{aligned} \tag{6.2.27}$$

Also note

$$\begin{aligned} w_1^j & = u_t(t_j) - \tau^{-1}(u(t_j) - u(t_{j-1})) \\ & = \tau^{-1} \int_{t_{j-1}}^{t_j} (t - t_j) u_{tt}(t) \mathrm{d}t, \end{aligned}$$

$$\begin{aligned} w_2^j &= (I - R_h)\bar{\partial}_t u(t_j) \\ &= \tau^{-1}\int_{t_{j-1}}^{t_j} (I - R_h)u_t(t)\mathrm{d}t. \end{aligned}$$

So we have

$$\tau\sum_{j=1}^{n}\|w_1^j + w_2^j\|_0 \le \tau\int_0^{t_n}\|u_{tt}(t)\|_0\mathrm{d}t + Ch^{r+1/2}\int_0^{t_n}\|u_t(t)\|_{r+1}\mathrm{d}t. \tag{6.2.28}$$

Inserting (6.2.27) and (6.2.28) into (6.2.26) yields

$$\begin{aligned} \|e^n\|_0 \le \;& \|\phi - \phi_h\|_0 + Ch^{r+1/2}\|\phi\|_{r+1} + \tau\int_0^{t_n}\|u_{tt}(t)\|_0\mathrm{d}t \\ & + Ch^{r+1/2}\int_0^{t_n}\|u_t(t)\|_{r+1}\mathrm{d}t. \end{aligned} \tag{6.2.29}$$

Finally, a combination of (6.2.24) and (6.2.29) leads to the desired estimate

$$\begin{aligned} & \|u_h^n - u(t_n)\|_0 \\ \le\; & \|e^n\|_0 + \|\rho^n\|_0 \le \|\phi - \phi_h\|_0 + Ch^{r+1/2}\|\phi\|_{r+1} \\ & + \tau\int_0^{t_n}\|u_{tt}(t)\|_0\mathrm{d}t + Ch^{r+1/2}\int_0^{t_n}\|u_t(t)\|_{r+1}\mathrm{d}t. \end{aligned} \tag{6.2.30}$$

This completes the proof. □

If we approximate the derivative in (6.2.12) by a weighted differencing, then we have the following six-point difference scheme:

$$\begin{aligned} & (u_h^n, v_h) + \tau a(\theta u_h^n + (1-\theta)u_h^{n-1}, v_h) \\ =\; & (u_h^{n-1} + \tau\theta f^n + \tau(1-\theta)f^{n-1}, v_h), \quad \forall v_h \in V_h, \end{aligned} \tag{6.2.31}$$

where $\theta \in [0, 1]$ is a parameter. In particular, choosing $\theta = \frac{1}{2}$ results in a Crank-Nicolson scheme:

$$\begin{aligned} & (u_h^n, v_h) + \tau a((u_h^n + u_h^{n-1})/2, v_h) \\ =\; & (u_h^{n-1} + \tau(f^n + f^{n-1})/2, v_h), \quad \forall v_h \in V_h. \end{aligned} \tag{6.2.32}$$

**Theorem 6.2.2** *Assume the following:* $\frac{1}{2} \le \theta \le 1$; $u(x,t)$ *and* $u_h^n(x)$ *are the solutions to (6.2.1) and (6.2.31) respectively;* $u_t \in H^{r+1}(\Omega)$; $u_{tt} \in L^2(\Omega)$; $u(x,0) = \phi(x) \in H^{r+1}(\Omega)$; *and* $u_h^0(x) = \phi_h(x) \in V_h$. *Then (6.2.30) holds for the approximate solution* $u_h^n$.

The proof of this theorem is analogous to that of Theorem 6.2.1, and is left for interested readers.

## 6.3 Generalized Upwind Schemes for First Order Hyperbolic Systems

In this section we extend the generalized upwind difference scheme to solve positive symmetric hyperbolic systems.

### 6.3.1 Integral forms

Let $\Omega \subset R^2$ be a bounded region with a piecewise smooth boundary $\partial\Omega$, and $u(x,t) \in R^m$ ($x \in \overline{\Omega}$, $0 \le t \le T$). Consider the first order positive symmetric hyperbolic system:

$$\begin{cases} \dfrac{\partial u(x,t)}{\partial t} + A(x)\cdot\nabla u(x,t) + K(x,t)u(x,t) = f(x,t), \\ \qquad\qquad x = (x_1,x_2) \in \Omega,\ 0 < t \le T, & (6.3.1a) \\ (B-M)u(x,t) = 0,\ x \in \partial\Omega, & (6.3.1b) \\ u(x,0) = \phi(x),\ x \in \Omega, & (6.3.1c) \end{cases}$$

where $K(x,t)$ is an $m \times m$ coefficient matrix; $A = (A_1, A_2)$; $A_i$ ($i = 1,2$) are $m \times m$ real symmetric matrices; $B = \sum_{i=1}^{2} n_i A_i$; $n = (n_1, n_2)$ is the unit outer normal vector of the boundary $\partial\Omega$; $M = M(x,t)$ is an $m \times m$ matrix; $f(x,t)$ and $\phi(x)$ are $m$-dimensional vector function. All these matrices and vectors are given. We also assume the following conditions:

$$M + M^T \ge 0, \text{ on } \partial\Omega, \tag{6.3.2a}$$

$$K + K^T - \sum_{i=1}^{2} \frac{\partial A_i}{\partial x_i} \ge \sigma_0 I, \text{ on } \Omega, \tag{6.3.2b}$$

$$\ker(B-M)+\ker(B+M)=R^m, \text{ on } \partial\Omega, \tag{6.3.2c}$$

where $\sigma_0 > 0$ is a constant, and

$$\ker E = \{v \in R^m : Ev = 0\}. \tag{6.3.3}$$

Under the above conditions plus certain smoothness assumptions, problem (6.3.1) possesses a unique solution (cf. [B-31]).

Define an operator $L$ by

$$Lu = \sum_{i=1}^{2} A_i(x)\frac{\partial u}{\partial x_i} + K(x,t)u, \ x \in \Omega, \tag{6.3.4}$$

and its formal conjugate by

$$L^*u = -\sum_{i=1}^{2} \frac{\partial}{\partial x_i}(A_i(x)u) + K(x,t)^T u. \tag{6.3.5}$$

For $u, v \in (H^1(\Omega))^m$, we may use Green's formula to get an extended Green's formula

$$(Lu, v)_\Omega = (u, L^*v)_\Omega + (Bu, v)_{\partial\Omega}. \tag{6.3.6}$$

Here and below we adopt the symbols

$$(u,v)_\Omega = \int_\Omega \langle u, v\rangle \mathrm{d}x, \ (u,v)_{\partial\Omega} = \int_{\partial\Omega} \langle u, v\rangle \mathrm{d}s.$$

Here the symbol $\langle u, v\rangle$ denotes the $R^m$ inner product of $u, v$. In particular,

$$(Lv, v)_\Omega = \frac{1}{2}((L+L^*)v, v)_\Omega + \frac{1}{2}(Bv, v)_{\partial\Omega}. \tag{6.3.7}$$

In terms of (6.3.6) we can write (6.3.1a) in an integral form:

$$\left(\frac{\partial u}{\partial t}, v\right)_\Omega - (u, L^*v)_\Omega + (Bu, v)_{\partial\Omega} = (f, v)_\Omega, \tag{6.3.8}$$

### 6.3.2 Generalized upwind difference schemes

As in §6.2, we assume that $\Omega$ is a polygonal region, that $T_h = \{K\}$ is a triangulation of $\Omega$, and $T_h^* = \{K_{P_0}^*\}$ is a barycenter dual grid. The finite element space $V_h$ is defined as in §6.2, namely,

$$V_h = \{v_h : v_h|_{K^*} \in \mathcal{P}_r, \ \forall K^* \in T_h^*; \ v_h = 0 \text{ on } K_{P_0}^* \text{ for } P_0 \in (\partial\Omega)_-\}.$$

Since we are dealing with vector functions, we introduce $\bar{V}_h = [V_h]^m = V_h \times \cdots \times V_h$ (an $m$-multiplicative space). Employ (6.3.8) on each $K_{P_0}^*$ to get

$$\begin{aligned} &\sum_{P_0}\Big[\Big(\frac{\partial u_h}{\partial t}, v_h\Big)_{K_{P_0}^*} - (u_h, L^* v_h)_{K_{P_0}^*} + (Bu_h, v_h)_{\partial K_{P_0}^*}\Big] \\ = \ &\sum_{P_0}(f, v_h)_{K_{P_0}^*}, \end{aligned} \tag{6.3.9}$$

where $u_h, v_h \in \bar{V}_h$, $B = \sum_{i=1}^{2} \nu_i A_i$, and $\nu = (\nu_1, \nu_2)^T$ is the outer normal vector of $\partial K_{P_0}^*$. Assume that the $m$-order symmetric matrix $B$ has $m$ real eigenvalues: $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_m$, that there exists a constant $q \in [1, m)$ (dependent on $K_{P_0}^*$) and a constant $c_0 > 0$ (independent of $K_{P_0}^*$) such that

$$\lambda_1, \cdots, \lambda_q < -c_0, \ \lambda_{q+1}, \cdots, \lambda_m > c_0, \tag{6.3.10}$$

and that there exists an $m$-order orthogonal matrix $Q$ such that

$$B = Q\Lambda Q^T, \ \ (Q^T = Q^{-1},) \tag{6.3.11}$$

where $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_m)$.

Set $w_h = (w_h^1, \cdots, w_h^m)^T = Q^T u_h$ or $u_h = Qw_h$. For each side $\overline{Q_iQ_{i+1}}$ of $K_{P_0}^*$ (cf. Fig. 3.2.1), write $q_i$ for the $q$ validating (6.3.10), and define the upwind and the downwind values of $w_h^j$ as

$$(w_h^j)^+ = \begin{cases} \text{the value of } w_h^j \text{ outside of } \overline{Q_iQ_{i+1}}, & \text{when } 1 \le j \le q_i, \\ \text{the value of } w_h^j \text{ inside of } \overline{Q_iQ_{i+1}}, & \text{when } q_i + 1 \le j \le m, \end{cases} \tag{6.3.12a}$$

$$(w_h^j)^- = \begin{cases} \text{the value of } w_h^j \text{ inside of } \overline{Q_iQ_{i+1}}, & \text{when } 1 \le j \le q_i, \\ \text{the value of } w_h^j \text{ outside of } \overline{Q_iQ_{i+1}}, & \text{when } q_i + 1 \le j \le m. \end{cases} \tag{6.3.12b}$$

Since

$$\begin{aligned} [w_h] &= ([w_h^1], \cdots, [w_h^m]) \\ &= ((w_h^1)^+ - (w_h^1)^-, \cdots, (w_h^m)^+ - (w_h^m)^-), \end{aligned}$$

we have

$$[u_h] = Q[w_h]. \tag{6.3.13}$$

Set $|\Lambda|$ =diag$(|\lambda_1|, \cdots, |\lambda_m|)$, and extend (6.2.9) to the system of equations here, then we have a semi-discrete highly accurate upwind difference scheme approximating (6.3.1): Find $u_h \in \bar{V}_h$ such that

$$\begin{aligned} &\sum_{P_0}\Big[\Big(\frac{\partial u_h}{\partial t}, v_h\Big)_{K_{P_0}^*} + (A \cdot \nabla u_h, v_h)_{K_{P_0}^*} + (Ku_h, v_h)_{K_{P_0}^*} \\ &-\frac{1}{2}(Q(|\Lambda| - \Lambda)Q^T[u_h], v_h)_{\partial K_{P_0}^*}\Big] = \sum_{P_0}(f, v_h)_{K_{P_0}^*}, \end{aligned}$$

$$\forall v_h \in \bar{V}_h, \tag{6.3.14a}$$

$$(B - M)u_h = 0, \ x \in \partial\Omega, \tag{6.3.14b}$$

$$u_h(x, 0) = \phi_h(x), \ x \in \Omega. \tag{6.3.14c}$$

One can further approximate $\frac{\partial u_h}{\partial t}$ by a proper difference quotient to get forward, backward, or Crank-Nicolson fully-discrete highly accurate upwind schemes.

### 6.3.3 Estimation of a bilinear form

Let us introduce a bilinear form

$$\begin{aligned} a(u_h, v_h) = &\sum_{P_0}\Big[(A \cdot \nabla u_h, v_h)_{K_{P_0}^*} + (Ku_h, v_h)_{K_{P_0}^*} \\ &-\frac{1}{2}(Q(|\Lambda| - \Lambda)Q^T[u_h], v_h)_{\partial K_{P_0}^*}\Big]. \end{aligned} \tag{6.3.15}$$

By (6.3.7)

$$\begin{aligned}&(A\cdot\nabla v_h, v_h)_{K^*_{P_0}} + (Kv_h, v_h)_{K^*_{P_0}}\\ = \;& \frac{1}{2}\Big(\Big(K+K^T-\sum_{i=1}^{2}\frac{\partial A_i}{\partial x_i}\Big)v_h, v_h\Big)_{K^*_{P_0}} + \frac{1}{2}(Bv_h, v_h)_{\partial K^*_{P_0}}.\end{aligned}$$

So we have

$$\begin{aligned}&a(v_h, v_h)\\ = \;& \frac{1}{2}\sum_{P_0}\Big[\Big(\Big(K+K^T-\sum_{i=1}^{2}\frac{\partial A_i}{\partial x_i}\Big)v_h, v_h\Big)_{K^*_{P_0}} + (Q\Lambda Q^T v_h, v_h)_{\partial K^*_{P_0}}\\ &-(Q(|\Lambda|-\Lambda)Q^T[v_h], v_h)_{\partial K^*_{P_0}}\Big].\end{aligned} \tag{6.3.16}$$

It follows from the assumption (6.3.2b) that

$$\begin{aligned}a(v_h, v_h) \;&\geq \sigma_0\sum_{P_0}(v_h, v_h)_{K^*_{P_0}} + \frac{1}{2}\sum_{P_0}[(Q^TQ\Lambda\tilde{v}_h, \tilde{v}_h)_{\partial K^*_{P_0}}\\ &\quad -(Q^TQ(|\Lambda|-\Lambda)[\tilde{v}_h], \tilde{v}_h)_{\partial K^*_{P_0}}]\\ &= \sigma_0\sum_{P_0}(v_h, v_h)_{K^*_{P_0}} + \frac{1}{2}\sum_{P_0}[(\Lambda\tilde{v}_h, \tilde{v}_h)_{\partial K^*_{P_0}}\\ &\quad -((|\Lambda|-\Lambda)[\tilde{v}_h], \tilde{v}_h)_{\partial K^*_{P_0}}],\end{aligned} \tag{6.3.17}$$

where $\tilde{v}_h = Q^T v_h$.

Let $\overline{Q_iQ_{i+1}}$ be a side of a dual element $K^*_{P_0}$ (cf. Fig. 3.2.1). On $\overline{Q_iQ_{i+1}}$, decompose $\tilde{v}_h$ into a sum of $\tilde{v}_h^-$ and $\tilde{v}_h^+$, where the first $q_i$ entries of $\tilde{v}_h^-$ are equal to the counterpart of $\tilde{v}_h$ and the last $(m-q_i)$ entries are zero, while the last $(m-q_i)$ entries of $\tilde{v}_h^+$ are identical to those of $\tilde{v}_h$ and the first $q_i$ entries are zero. So $\tilde{v}_h = \tilde{v}_h^- + \tilde{v}_h^+$. Also decompose $\Lambda$ into a sum of $\Lambda^-$ and $\Lambda^+$, where $\Lambda^-$ =diag$(\lambda_1, \cdots, \lambda_{q_i}, 0, \cdots, 0)$ and $\Lambda^+$ =diag$(0, \cdots, 0, \lambda_{q_i+1}, \cdots, \lambda_m)$.

Then, the second term on the right-hand side of (6.3.17) is equal to

$$\begin{aligned}J = &\sum_{P_0}\sum_i \int_{\overline{Q_iQ_{i+1}}}\Big[\frac{1}{2}\langle\Lambda\tilde{v}_h,\tilde{v}_h\rangle - \frac{1}{2}\langle(|\Lambda|-\Lambda)[\tilde{v}_h],\tilde{v}_h\rangle\Big]\mathrm{d}s\\ = &\sum_{P_0}\sum_i \int_{\overline{Q_iQ_{i+1}}}\Big[\frac{1}{2}\langle\Lambda\tilde{v}_h,\tilde{v}_h\rangle - \frac{1}{2}\langle(|\Lambda|-\Lambda)[\tilde{v}_h],\tilde{v}_h\rangle\\ &+\frac{1}{4}\langle(|\Lambda|-\Lambda)[\tilde{v}_h],\tilde{v}_h\rangle - \frac{1}{4}\langle(|\Lambda|-\Lambda)[\tilde{v}_h],\tilde{v}_h\rangle\Big]\mathrm{d}s.\end{aligned} \tag{6.3.18}$$

But on $\overline{Q_iQ_{i+1}}$

$$\frac{1}{2}\langle\Lambda\tilde{v}_h,\tilde{v}_h\rangle = \frac{1}{2}\langle\Lambda^-\tilde{v}_h^-,\tilde{v}_h^-\rangle + \frac{1}{2}\langle\Lambda^+\tilde{v}_h^+,\tilde{v}_h^+\rangle,$$

$$-\frac{1}{2}\langle(|\Lambda|-\Lambda)[\tilde{v}_h],\tilde{v}_h\rangle = \langle\Lambda^-(\tilde{v}_h^+-\tilde{v}_h^-),\tilde{v}_h^-\rangle,$$

$$-\frac{1}{4}\langle(|\Lambda|-\Lambda)[\tilde{v}_h],[\tilde{v}_h]\rangle = \frac{1}{2}\langle\Lambda^-(\tilde{v}_h^+-\tilde{v}_h^-),\tilde{v}_h^+-\tilde{v}_h^-\rangle.$$

The sum of the left-hand sides of the above three equalities are

$$\frac{1}{2}[\langle\Lambda^-\tilde{v}_h^+,\tilde{v}_h^+\rangle + \langle\Lambda^+\tilde{v}_h^+,\tilde{v}_h^+\rangle].$$

Hence

$$\begin{aligned}J = &\frac{1}{2}\sum_{P_0}\sum_i\int_{\overline{Q_iQ_{i+1}}}[\langle\Lambda^-\tilde{v}_h^+,\tilde{v}_h^+\rangle\\ &+\langle\Lambda^+\tilde{v}_h^+,\tilde{v}_h^+\rangle + \langle|\Lambda^-|\tilde{v}_h^-,\tilde{v}_h^-\rangle]\mathrm{d}s.\end{aligned}$$

Notice that on the inner boundaries the first two integrals on the right-hand side of the above equality cancel each other out, and that on the boundary of $\Omega$ we have the zero boundary condition. Therefore we have

$$J = \frac{1}{2}\sum_{P_0}\sum_i\int_{\overline{Q_iQ_{i+1}}}\langle|\Lambda^-|\tilde{v}_h^-,\tilde{v}_h^-\rangle\mathrm{d}s + \int_{\partial\Omega}\langle M\tilde{v}_h^-,\tilde{v}_h^-\rangle\mathrm{d}s. \tag{6.3.19}$$

Write $\gamma_0 = \min\left(\sigma_0,\frac{1}{2}\right)$ and

$$\|v_h\|_{0,\Omega}^2 = (v_h,v_h),$$

$$\|v_h\|^2_{0,\partial\Omega} = \sum_{P_0}\sum_i \int_{\overline{Q_iQ_{i+1}}} \langle|\Lambda^-|\tilde{v}_h^-, \tilde{v}_h^-\rangle \mathrm{d}s + \int_{\partial\Omega} \langle M\tilde{v}_h^-, \tilde{v}_h^-\rangle \mathrm{d}s.$$

Then it follows from (6.3.17)-(6.3.19) that

$$a(v_h, v_h) \geq \gamma_0(\|v_h\|^2_{0,\Omega} + \|v_h\|^2_{0,\partial\Omega}). \tag{6.3.20}$$

This shows the positive definiteness of the bilinear form $a(u_h, v_h)$.

Analogous to the proof in §6.2, one can obtain the error estimates like (6.2.30) for semi- and fully-discrete approximations.

### 6.3.4 Some practical difference schemes

Consider the first order linear equation:

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0,\ 0 < x < L,\ t > 0, \tag{6.3.21}$$

where $a$ is a constant. Take a step size $h = L/N$ and nodes $x_j = jh$ $(0 \leq j \leq N)$, then we have a uniform grid $T_h$:

$$0 = x_0 < x_1 < \cdots < x_N = L.$$

Then, choose dual nodes $x_{j+1/2} = (j + \frac{1}{2})h$ $(j = 0, 1, \cdots, N-1)$ to obtain a dual grid $T_h^*$:

$$0 = x_0 < x_{1/2} < x_{3/2} < \cdots < x_{N-1/2} < x_N = L.$$

#### The classical upwind scheme

Let the finite element space $V_h$ be composed of piecewise constant functions relative to the dual grid. Obviously any $u_h \in V_h$ has the following expression:

$$u_h = u_h(x,t) = \sum_j u_j(t)\psi_j^{(0)}(x), \tag{6.3.22}$$

where

$$\psi_j^{(0)}(x) = \begin{cases} 1, & x \in [x_{j-1/2}, x_{j+1/2}], \\ 0, & \text{elsewhere}, \end{cases} \quad 1 \leq j \leq N-1,$$

$$\psi_0^{(0)}(x) = \begin{cases} 1, & x \in [0, x_{1/2}], \\ 0, & \text{elsewhere}, \end{cases}$$

$$\psi_N^{(0)}(x) = \begin{cases} 1, & x \in [x_{N-1/2}, x_N], \\ 0, & \text{elsewhere}. \end{cases}$$

(6.3.23)

Evidently $u_h(x_j, t) = u_j(t)$. Substitute (6.3.22) in (6.2.9), and choose $v_h = \psi_j^{(0)}$, then we have

$$\begin{aligned} & \int_{x_{j-1/2}}^{x_{j+1/2}} \frac{\partial u_h}{\partial t} \mathrm{d}x + \int_{x_{j-1/2}}^{x_{j+1/2}} a \frac{\partial u_h}{\partial x} \mathrm{d}x \\ & -\frac{1}{2}[(|a| - a)(u_{j+1}(t) - u_j(t)) \\ & +(|a| + a)(u_{j-1}(t) - u_j(t))] \\ = \; & h\frac{\mathrm{d}u_j(t)}{\mathrm{d}t} - \frac{1}{2}[(|a| - a)(u_{j+1}(t) - u_j(t)) \\ & +(|a| + a)(u_{j-1}(t) - u_j(t))] \\ = \; & 0. \end{aligned}$$

Take a time step size $\tau > 0$ and exploit the forward difference formula

$$\frac{\mathrm{d}u_j(t)}{\mathrm{d}t} \approx \frac{u_j^{n+1} - u_j^n}{\tau}, \; u_j^n \approx u_j(n\tau),$$

then we have

$$\frac{u_j^{n+1} - u_j^n}{\tau} + a\frac{u_j^n - u_{j-1}^n}{\tau} = 0, \text{ as } a \geq 0, \tag{6.3.24a}$$

$$\frac{u_j^{n+1} - u_j^n}{\tau} + a\frac{u_{j+1}^n - u_j^n}{\tau} = 0, \text{ as } a < 0. \tag{6.3.24b}$$

This is precisely the classical upwind scheme. The sufficient and necessary condition of its stability is $r = |a|\tau/h \leq 1$ (cf. [A-27] and [B-74]).

### A second order upwind scheme

Let the basis functions of $V_h$ be composed of two groups of functions, of which the first group is $\{\psi_j^{(0)}\}$ given in (6.3.23), and the other one is $\{\psi_j^{(1)}\}$ defined by

$$\psi_j^{(1)}(x) = \begin{cases} x - x_j, & x \in [x_{j-1/2}, x_{j+1/2}], \\ 0, & \text{elsewhere}, \end{cases} \quad 1 \le j \le N-1,$$

$$\psi_0^{(1)}(x) = \begin{cases} x, & x \in [0, x_{1/2}], \\ 0, & \text{elsewhere}, \end{cases}$$

$$\psi_N^{(0)}(x) = \begin{cases} x - x_N, & x \in [x_{N-1/2}, x_N], \\ 0, & \text{elsewhere}. \end{cases}$$

An element of $V_h$ is of the form

$$u_h = u_h(x,t) = \sum_j [u_{0j}(t)\psi_j^{(0)}(x) + u_{1j}(t)\psi_j^{(1)}(x)]. \tag{6.3.26}$$

Substitute $u_h$ into (6.2.9), choose $v_h = \psi_j^{(0)}$ and $\psi_j^{(1)}$, and approximate $\frac{\partial u_h}{\partial t}$ by a forward differencing, then we have the following two groups of equations:

$$\begin{cases} \dfrac{u_{0j}^{n+1} - u_{0j}^n}{\tau} + a\dfrac{u_{0j}^n - u_{0j-1}^n}{h} + \dfrac{a}{2}(u_{1j}^n - u_{1j-1}^n) = 0, \text{ as } a \ge 0, \\ \dfrac{u_{0j}^{n+1} - u_{0j}^n}{\tau} + a\dfrac{u_{0j+1}^n - u_{0j}^n}{h} + \dfrac{a}{2}(u_{1j+1}^n - u_{1j}^n) = 0, \text{ as } a < 0, \end{cases} \tag{6.3.27a}$$

$$\begin{cases} \dfrac{u_{1j}^{n+1} - u_{1j}^n}{\tau} - \dfrac{6a}{h^2}(u_{0j}^n - u_{0j-1}^n) + \dfrac{3a}{h}(u_{1j}^n + u_{1j-1}^n) = 0, \text{ as } a \ge 0, \\ \dfrac{u_{1j}^{n+1} - u_{1j}^n}{\tau} - \dfrac{6a}{h^2}(u_{0j+1}^n - u_{0j}^n) + \dfrac{3a}{h}(u_{1j+1}^n + u_{1j}^n) = 0, \text{ as } a < 0. \end{cases} \tag{6.3.27b}$$

One may use the variable separation method to deduce the corresponding amplification matrix:

$$G = \begin{bmatrix} 1 - \frac{a\tau}{h}(1 - e^{-\mathrm{i}\sigma h}) & -\frac{a\tau}{2h}(1 - e^{-\mathrm{i}\sigma h}) \\ \frac{6a\tau}{h^2}(1 - e^{-\mathrm{i}\sigma h}) & 1 - \frac{3a\tau}{h}(1 + e^{-\mathrm{i}\sigma h}) \end{bmatrix},$$

where $\sigma = 2\pi l$, $(l = 0, \pm 1, \cdots)$, i$= \sqrt{-1}$. It can proved that $G$ does not satisfy the von Neumann condition, since it always has an eigenvalue such that the lower bound of the absolute value of this eigenvalue is greater than 1, so (6.3.27) is absolutely unstable. But if we instead use backward or Crank-Nicolson difference approximations, then the resulting upwind schemes will be absolutely stable.

### 6.3.5 A numerical example

We use upwind schemes to solve the Riemann problem of Burger's equation:

$$\frac{\partial u}{\partial t} + \frac{1}{2}\frac{\partial}{\partial x}(u^2) = 0, \ -\infty < x < \infty, \tag{6.3.28a}$$

$$u(x, 0) = \begin{cases} 1, & \text{as } x \le 0, \\ 0, & \text{as } x > 0. \end{cases} \tag{6.3.28b}$$

The classical upwind scheme leads to the equations:

$$\begin{cases} u_j^{n+1} = u_j^n - r u_j^n (u_j^n - u_{j-1}^n), & \text{as } u_j^n \ge 0, \\ u_j^{n+1} = u_j^n - r u_j^n (u_{j+1}^n - u_j^n), & \text{as } u_j^n < 0. \end{cases} \tag{6.3.29}$$

where $r = \tau/h$. The numerical solution is given in Fig. 6.3.1(a). We observe that the shock wave is too flat and evolves too slow, indicating a notable error. Fig. 6.3.1(b) depicts the numerical results of the second order Crank-Nicolson upwind scheme (cf. [B-60]). Now the shock wave is steeper and its evolution is faster, very close to the true solution. But there appear oscillations after the wave.

**Remark** If we keep working with only one grid $T_h$, and replace the dual element by an element $K \in T_h$, then our strategy of

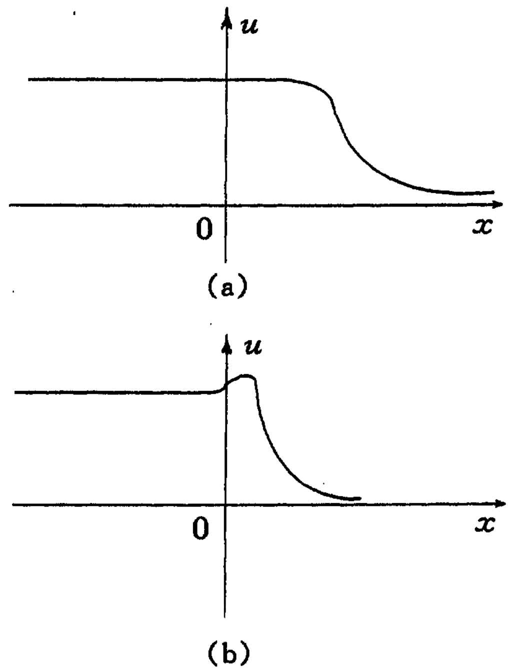



Fig. 6.3.1

constructing the generalized upwind scheme still works and results in some difference schemes called *box schemes*. Parallel results of convergence and error estimates can be similarly obtained for these box schemes (cf. [A-59] and [B-41]).

## 6.4 Finite Volume Methods for Nonlinear Conservative Hyperbolic Equations

The finite volume method (FVM for short), combined with, e.g., the Godunov scheme (see Example 1 below) or the TVD scheme (cf. [B-36]), has become one of the most popular methods for fluid computation in the last twenty years. In this section we introduce a FVM for the following system of conservative hyperbolic equations:

$$\frac{\partial u_j}{\partial t}+\sum_{i=1}^{n}\frac{\partial f_{ij}}{\partial x_i}=0,\ j=1,\cdots,d, \tag{6.4.1}$$

where $f_{ij}$'s are smooth functions of $\mathbf{u}=(u_1,\cdots,u_d)$. If we set $F=(f_{ij})_{n\times d}$ and $\nabla=\left(\frac{\partial}{\partial x_1},\cdots,\frac{\partial}{\partial x_n}\right)$, then (6.4.1) becomes

$$\frac{\partial \mathbf{u}}{\partial t}+\nabla\cdot F=0. \tag{6.4.1)'$$

An example of (6.4.1) is the Euler equation describing a one-dimensional non–steady-state flow ($n=1,\ d=3$):

$$\mathbf{u}=\begin{pmatrix}\rho\\ \rho u\\ \rho(e+\frac{1}{2}u^2)\end{pmatrix},\quad F=\begin{pmatrix}\rho u\\ p+\rho u^2\\ \rho u(e+\frac{p}{\rho}+\frac{1}{2}u^2)\end{pmatrix},$$

where $\rho$ is the density, $p$ the pressure, $u$ the velocity, and $e$ the internal energy. $\rho$, $p$ and $e$ satisfy a state equation $p=p(\rho,e)$. Let $\Omega\subset R^n$ be a bounded region. The initial-boundary value problem of the conservative equation (6.4.1) reads: Find $\mathbf{u}(x,t):\ \Omega\times[0,\infty)\to R^d$ satisfying (6.4.1) and

$$\begin{cases}\mathbf{u}(x,0)=\mathbf{u}_0(x), & x\in\Omega,\\ \mathbf{u}(x,t)=\phi(x), & x\in\partial\Omega.\end{cases} \tag{6.4.2}$$

Next, we consider the case of $d=1$ and $n=2$ to illustrate the idea. Now (6.4.1) reads

$$\frac{\partial u}{\partial t}+\nabla\cdot F(u)=0,\ \text{on}\ \Omega\subset R^2, \tag{6.4.3}$$

where $F=(f_1,f_2)^T$. Let $D$ be a subregion, e.g., a polygon, of $\Omega$. Integrate (6.4.3) on $D$ and make use of Green's formula, then we have

$$\int_D\frac{\partial u}{\partial t}\mathrm{d}x+\int_{\partial D}F\cdot\nu\mathrm{d}s=0, \tag{6.4.3)'$$

where $\nu$ is the unit outer normal vector. This is an integral form of (6.4.3). The so-called finite volume method is precisely the discretization method based on (6.4.3)$'$.

Since (6.4.3) is a conservative equation, it is natural to expect, and we shall try to ensure, its discretization to possess the conservative property as well.

As before, we place a quasi-uniform triangulation $T_h = \{K\}$ on $\Omega$. Suppose none of the triangular elements is an obtuse triangle. Denote by $T_h^*$ the circumcenter dual grid of $T_h$. As in Fig. 3.2.2, let $P_0$ be a node of $T_h$, $P_i$ $(i = 1, 2, \cdots, 6)$ the neighbouring nodes of $P_0$, and $K_{P_0}^*$ the dual element surrounding $P_0$ with vertexes $Q_i$ $(i = 1, 2, \cdots, 6)$. Choose in (6.4.3)$'$ $D = K_{P_0}^*$ as a control volume, and set $t = n\tau$ ($\tau > 0$ is the time step size), then we have

$$\int_{K_{P_0}^*} \frac{\partial u^n}{\partial t} \mathrm{d}x + \sum_{i=1}^{6} \int_{\overline{Q_i Q_{i+1}}} F^n \cdot \nu \mathrm{d}s = 0, \quad (Q_7 = Q_1) \qquad (6.4.4)$$

where the superscript $n$ denotes the function value on $t = t_n$. Note that $\overline{Q_i Q_{i+1}}$ is the perpendicular bisector of $\overline{P_0 P_{i+1}}$. So if we use $\nu_{0,i+1}$ to denote the unit outer normal vector on $\overline{Q_i Q_{i+1}}$ towards $P_{i+1}$, then

$$\mathcal{F}_{0,i+1} = \int_{\overline{Q_i Q_{i+1}}} F^n \cdot \nu_{0,i+1} \mathrm{d}s \qquad (6.4.5)$$

is the flux flowing out of $K_{P_0}^*$ and passing through the side $\overline{Q_i Q_{i+1}}$. Similarly consider the dual element $K_{P_{i+1}}^*$ surrounding $P_{i+1}$, then we have the fiux out of $K_{P_{i+1}}^*$ and passing through $\overline{Q_i Q_{i+1}}$:

$$\mathcal{F}_{i+1,0} = \int_{\overline{Q_i Q_{i+1}}} F^n \cdot \nu_{i+1,0} \mathrm{d}s.$$

Apparently

$$\mathcal{F}_{i+1,0} = -\mathcal{F}_{0,i+1}.$$

Now we can write (6.4.4) in the form

$$\int_{K_{P_0}^*} \frac{\partial u^n}{\partial t} \mathrm{d}x + \sum_{i=1}^{6} \mathcal{F}_{0,i+1} = 0. \qquad (6.4.6)$$

The time derivative is usually approximated by a forward (explicit) differencing. In the space direction, one often uses $u_0^n$ and $u_{i+1}^n$ to discretize (6.4.5) to obtain a so-called numerical flux. So let the numerical flux be of the form

$$\mathcal{F}_{0,i+1}^h = |\overline{Q_iQ_{i+1}}|g_{0,i+1}(u_0^n, u_{i+1}^n), \tag{6.4.7}$$

where $g_{0,i+1}(u_0^n, u_{i+1}^n)$ is suitably chosen. Then we have the finite volume equation:

$$u_0^{n+1} = u_0^n - \sum_i \frac{\tau|\overline{Q_iQ_{i+1}}|}{S_{P_0}^*} g_{0,i+1}(u_0^n, u_{i+1}^n), \tag{6.4.8}$$

where $S_{P_0}^*$ denotes the area of the dual element $K_{P_0}^*$. We require $g_{0,i+1}(u_0^n, u_{i+1}^n)$ to satisfy the conservative property

$$g_{0,i+1}(u_0^n, u_{i+1}^n) = -g_{i+1,0}(u_{i+1}^n, u_0^n) \tag{6.4.9}$$

and the consistency (cf. (6.4.3)′)

$$g_{0,i+1}(u, u) = F(u) \cdot \nu_{0,i+1}. \tag{6.4.10}$$

We also require $g_{0,i+1}(u_0^n, u_{i+1}^n)$ to possess a monotonicity, that is,

$$\frac{\partial g_{0,i+1}}{\partial u_0} \geq 0, \quad \frac{\partial g_{0,i+1}}{\partial u_{i+1}} \leq 0. \tag{6.4.11}$$

The numerical flux $g_{0,i+1}$ can be determined by, e.g., one-dimensional Godunov or Lax-Friedrichs schemes. In fact, if we define an $x'$-axis along $\overline{P_0P_{i+1}}$ with positive direction $\nu_{0,i+1}$, let any a point of $x \in R^2$ on $\overline{P_0P_{i+1}}$ correspond to $x' = x \cdot \nu_{0,i+1} \in R$, and set $w = w(x', t) = u(x, t)$, then $g_{0,i+1}$ becomes the numerical flux consistent with the following one-dimensional equation:

$$\frac{\partial w}{\partial t} + \frac{\partial}{\partial x'}(F(w) \cdot \nu_{0,i+1}) = 0. \tag{6.4.12}$$

**Example 1. Godunov scheme.** Set $f_{0,i+1} = F(w) \cdot \nu_{0,i+1}$, and define

$$g_{0,i+1} = f_{0,i+1}(w(0^+; u_0, u_{i+1}; f_{0,i+1})),$$

where $w(0^+; u_0, u_{i+1}; f_{0,i+1}) = w(x', t)$ is the solution of the following Riemann problem

$$\frac{\partial w}{\partial t} + \frac{\partial}{\partial x'} f_{0,i+1}(w) = 0, \quad t > 0, \ x' \in R,$$

$$w(x', 0) = \begin{cases} u_0, & \text{when } x' < 0, \\ u_{i+1}, & \text{when } x' > 0. \end{cases}$$

Here we require the time step size to be small enough such that the waves of the neighbouring Riemann problems will not interfere each other. One may show that the Godunov scheme indeed possesses properties (6.4.9)-(6.4.11). (cf. [B-19].)

**Example 2. Lax-Friedrichs scheme.** In this case the numerical flux is defined by

$$g_{0,i+1} = \frac{1}{2}(\nu_{0,i+1}F(u_0) + \nu_{0,i+1}F(u_{i+1})) - \frac{1}{2\lambda_{0,i+1}}(u_{i+1} - u_0),$$

where $\lambda_{0,i+1}$ is independent of $u_0$ and $u_{i+1}$, and satisfies

$$\lambda_{0,i+1} = \lambda_{i+1,0} > 0$$

as well as the CFL condition

$$\lambda_{0,i+1}\left\| \frac{\partial}{\partial u} F \cdot \nu_{0,i+1} \right\|_\infty \leq 1.$$

It can be shown (see [B-19] and [B-12]) that this scheme has the properties (6.4.9)-(6.4.11).

**Example 3** Suppose $F$ in equation (6.4.3) is of the form

$$F = b(x, t) f(u),$$

where $f : R \to R$ is a smooth function and $b(x,t) : R^2 \times [0, \infty) \to R^2$ a given vector function. Then (6.4.4) reads

$$\int_{K^*_{P_0}} \frac{\partial u^n}{\partial t} \mathrm{d}x + \sum_{i=1}^{6} \int_{\overline{Q_iQ_{i+1}}} (b \cdot \nu) f(u^n) \mathrm{d}s = 0.$$

Set

$$\beta_{ij} = \frac{1}{|\overline{Q_iQ_{i+1}}|}\int_{\overline{Q_iQ_{i+1}}} b\cdot\nu \mathrm{d}s,$$

$$g_{0,i+1} = \begin{cases} \beta_{ij}f(u_{i+1}), & \text{when } \beta_{ij} \le 0, \\ \beta_{ij}f(u_0), & \text{when } \beta_{ij} > 0, \end{cases}$$

$$\mathcal{F}^h_{0,i+1} = |\overline{Q_iQ_{i+1}}|g_{0,i+1}(u^n_0, u^n_{i+1}).$$

Then (6.4.8) is precisely the upwind scheme given in §6.2. Clearly when $\tau$ is sufficiently small such that the CFL condition holds, the upwind scheme satisfies the conditions (6.4.9) and (6.4.11). The consistency condition is now valid approximately.

**Remark 1** The FVM considered here adopts the nodes of $T_h$ as the center of the control volume, and hence is referred to be node-centred. Actually a node of $T_h$ is also a center of an element of $T^*_h$. So this FVM can as well be called cell-centred with respect to the dual grid $T^*_h$. Another strategy is to take a cell $K \in T_h$ (or $K^* \in T^*_h$) as the control center, and the vertexes of $K$ (or $K^*$) as nodes, resulting a FVM of cell-vertex type. (cf. [B-79].)

**Remark 2** FVM can be extended to solve a system of equations ($d > 1$). For instance, let us consider the case $n = 1$. (6.4.1) now reads

$$\frac{\partial \mathbf{u}}{\partial t} + F'(\mathbf{u})\frac{\partial \mathbf{u}}{\partial x} = 0.$$

Since we assume that this system of equations is of hyperbolic type, the Jacobian matrix $A = F'(\mathbf{u})$ can be reduced to a diagonal form. Accordingly, we can deduce this system to a system of scalar characteristic equations:

$$\frac{\partial w^k}{\partial t} + a^k\frac{\partial w^k}{\partial x} = 0, \ k = 1, 2, \cdots, d.$$

Now we can employ the FVM to each equation to obtain a numerical flux on $x$-direction ([B-36]). Furthermore, in the two-dimensional case ($n = 2$), one can similarly get the numerical flux on $y$-direction, and to deduce a FVM on a rectangular grid.

**Remark 3.** A scheme satisfying (6.4.11) is called a monotone scheme. There has been a great deal of research on the convergence of monotone FVM (cf. [A-57], [B-40], [B-19] and [B-12] etc.).

**Remark 4.** In [B-97], FVM has been successfully used to generalize TVD schemes to constructive quadrilateral grids, which have found wide applications in aerodynamics computations. (cf. [B-43,44,75,80,92].)

## Bibliography and Comments

[A-1,24] are among the earliest works for second order hyperbolic equations. An abstract framework is constructed in [A-1]. [A-24] deals with quasi-linear hyperbolic equations. The first section of this chapter is based upon [A-24]. An extension of the results in [A-24] to more general quasi-linear hyperbolic equations is presented in [A-45,47].

The second and third sections of this chapter are devoted to the study of generalized difference methods for first order system of hyperbolic equations, according in principle to [B-60]. Based on discontinuous finite element methods ([B-53],[A-59]), high accuracy generalized difference methods are proposed in [B-60], which differ from the usual discontinuous finite element methods in the following two aspects. Firstly, the discontinuous finite element methods are now used on the dual grid $T_h^*$ rather than the original grid $T_h$. Only in this way, one may end up with an extension of the upwind scheme. Secondly, an artificial viscosity, instead of the least square method as usual, is used in the extension of the results from a equation to a system of equations [B-41].

The high accuracy methods require further improvements such as: How to reduce the superfluous oscillations? And how to modify them for computing the discontinuous solutions (shock waves) of quasi-linear conservative equations. It seems necessary to introduce a proper diffusion term in the schemes.

The finite volume method was first used for computational fluid dynamics in the early seventies, resulting in a great number of references as well as software applications. §6.4 is only a rough introduction to this topic. For details, please see the corresponding references at the end of the book and certain journals such as J. Comput. Phys. and AIAA Journal.

# Chapter 7

# CONVECTION-DOMINATED DIFFUSION PROBLEMS

Convection-dominated diffusion problems often arise in mechanics, physics and other disciplines of applications. They are parabolic (non—steady-state) or elliptic (steady-state) equations. There have been many papers in recent years devoted to the numerical solution of this class of equations, aiming at constructing schemes that are stable, highly accurate, and suitable for small diffusion coefficients. Up to now, the schemes have been mainly various kinds of combinations of finite difference or finite element methods and characteristic methods (cf. [B-26]). In this chapter, we introduce some combinations of difference or generalized difference methods and characteristic methods, i.e., we use difference or generalized difference methods to discretize the diffusion term and characteristic methods to the convection term, resulting in various kinds of extensions of upwind schemes.

## 7.1 One-Dimensional Characteristic Difference Schemes

Consider the following Cauchy problem of a one-dimensional non–steady state convection-diffusion equation:

$$\begin{cases} \dfrac{\partial u}{\partial t} + b(x)\dfrac{\partial u}{\partial x} - \mu\dfrac{\partial^2 u}{\partial x^2} = f(x), & x \in \mathrm{R},\ t > 0, \quad (7.1.1\text{a}) \\ u(x,0) = u_0(x), & x \in \mathrm{R}. \quad (7.1.1\text{b}) \end{cases}$$

We assume $\mu > 0$ is a constant, and $\min_x |b(x)|$ is very large in comparison with $\mu$. Write

$$\psi(x) = [1 + b^2(x)]^{1/2}. \tag{7.1.2}$$

The characteristic direction with respect to the operator $\frac{\partial u}{\partial t} + b(x)\frac{\partial u}{\partial x}$ is

$$\tau = \tau(x) = \Big(\frac{1}{\psi(x)}, \frac{b(x)}{\psi(x)}\Big).$$

The directional derivative along $\tau$ is

$$\frac{\partial}{\partial \tau(x)} = \frac{1}{\psi(x)}\frac{\partial}{\partial t} + \frac{b(x)}{\psi(x)}\frac{\partial}{\partial x}. \tag{7.1.3}$$

Thus (7.1.1a) can be written as

$$\psi(x)\frac{\partial u}{\partial \tau} - \mu\frac{\partial^2 u}{\partial x^2} = f(x),\ x \in \mathrm{R},\ t > 0. \tag{7.1.4}$$

Take a time step size $\Delta t > 0$, and place a grid on $t$-axis with nodes $t_n = n\Delta t$ $(n = 0, 1, \cdots)$. The characteristic direction starting from $(x, t_n)$ crosses the straight line $t = t_{n-1}$ at

$$\bar{x} = x - b(x)\Delta t. \tag{7.1.5}$$

Naturally we use the following formula to approximate the characteristic directional derivative:

$$\begin{aligned} \psi(x)\frac{\partial u}{\partial \tau} &\approx \psi(x)\frac{u(x,t_n) - u(\bar{x},t_{n-1})}{[(x-\bar{x})^2 + (\Delta t)^2]^{1/2}} \\ &= \frac{u(x,t_n) - u(\bar{x},t_{n-1})}{\Delta t}. \end{aligned} \tag{7.1.6}$$

Correspondingly (7.1.4) is approximated by

$$\frac{u^n(x) - u^{n-1}(\bar{x})}{\Delta t} - \mu \frac{\partial^2 u^n(x)}{\partial x^2} = f(x). \qquad (7.1.7)$$

It remains to discretize the space variable. Since $\bar{x}$ is not necessarily a node, we need to evaluate the approximate solution $u_h(\bar{x}, t_{n-1})$. This is an easy matter for Galerkin finite element methods or generalized difference methods. As for finite difference methods, a linear or quadratic interpolation in terms of nodal values is often adopted to compute $u_h(\bar{x}, t_{n-1})$.

### 7.1.1 Difference methods based on algebraic interpolations

Take a space step size $h > 0$ and nodes $x_i = ih$. Noticing (7.1.5), we set $\bar{x}_i = x_i - b_i \Delta t$, where $b_i = b(x_i)$. Denote by $u_i^n$ a nodal function and $u^n(x)$ the piecewise linear function with nodal values $u_i^n$. Let $\bar{u}_i^n = u^n(\bar{x}_i)$. Define the second order central difference quotient

$$\bar{\partial}_{xx} u_i^n = \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{h^2}.$$

Then, the simplest difference scheme approximating (7.1.1) or (7.1.7) is:

$$\begin{cases} \dfrac{u_i^n - \bar{u}_{i-1}^n}{\Delta t} - \mu \bar{\partial}_{xx} u_i^n = f_i^n, & i = 0, \pm 1, \cdots; \; n \geq 1, \quad (7.1.8a) \\ u_i^0 = u_0(x_i), & i = 0, \pm 1, \cdots. \quad (7.1.8b) \end{cases}$$

This is an implicit characteristic difference scheme. One can use it to compute $u_i^n$ iteratively, starting from the initial value $u_i^0$.

Let $u(x,t)$ be the solution to (7.1.1). Restricting it at the nodes and inserting it into (7.1.8a) yield

$$\frac{u(x_i, t_n) - u(\bar{x}_i, t_{n-1})}{\Delta t} - \mu \bar{\partial}_{xx} u(x_i, t_n) = f_i^n + r_i^n, \qquad (7.1.9)$$

where $r_i^n$ is the truncation error. A simple calculation gives

$$r_i^n = \frac{(1 + b_i^2)}{2} \frac{\partial^2 u^*}{\partial \tau^2} \Delta t + O(\|u(x, t_{n-1})\|_{3,\infty} h), \qquad (7.1.10)$$

$$\text{(or } O(\|u(x,t_{n-1})\|_{4,\infty}h^2))$$

where $\frac{\partial^2 u^*}{\partial \tau^2}$ is the tangent-directional second derivative of $u$ along the characteristic line segment between $(x_i, t_n)$ and $(\bar{x}_i, t_{n-1})$. When $\mu = 0$, $u$ varies linearly as $\tau$. If in addition we assume $f = 0$, then $u$ becomes a constant along the characteristic line. Hence in the convection-dominated case, the second derivative $\frac{\partial^2 u}{\partial \tau^2}$ is generally less than $\frac{\partial^2 u}{\partial t^2}$ and $\frac{\partial^2 u}{\partial x^2}$. Suppose $b(x)$ is bounded: $|b(x)| \le K$, then

$$|r_i^n| \le \frac{1}{2}(1+K^2)\Delta t \sup_x \left|\frac{\partial^2 u^*}{\partial \tau^2}\right| + Ch\|u(x,t_n)\|_{3,\infty}, \qquad (7.1.11)$$

where $C$ is the general constant. Therefore, the order of the truncation error is $O(\Delta t + h)$.

If we wish to obtain an error order $O(\Delta t + h^2)$, then we have to use a quadratic interpolation. In such a case we naturally require

$$\Delta t = O(h^2), \text{ as } h \to 0.$$

By virtue of $|b(x)| \le K$ and (7.1.5), we know that, for sufficiently small $\Delta t$, $\bar{x}_i$ will lie in between $x_{i-1}$ and $x_{i+1}$. So we may use $u_{i-1}^{n-1}$, $u_i^{n-1}$ and $u_{i+1}^{n-1}$ to get a quadratic interpolation function $u^{n-1}(x)$ so as to determine $u^{n-1}(\bar{x}_i)$:

$$u^{n-1}(\bar{x}_i) = \frac{1}{2}\alpha_i^2(u_{i+1}^{n-1} + u_{i-1}^{n-1}) + (1-\alpha_i^2)u_i^{n-1} + \frac{1}{2}\alpha_i(u_{i+1}^{n-1} - u_{i-1}^{n-1}), \qquad (7.1.12)$$

where

$$\alpha_i = -b_i \Delta t / h. \qquad (7.1.13)$$

Then we can similarly obtain a difference equation as (7.1.8), save that $\bar{u}_i^{n-1} = u^{n-1}(\bar{x}_i)$ is now computed according to (7.1.12). A simple calculation shows that the truncation error in (7.1.9) now reads:

$$|r_i^n| \le C\Big(\Delta t \sup_{x,n}\left|\frac{\partial^2 u^*}{\partial \tau^2}\right| + h^2 \sup_{t>0}\|u(x,t)\|_{4,\infty}\Big). \qquad (7.1.14)$$

Its order is $O(\Delta t + h^2)$.

### 7.1.2 Upwind difference schemes

Once again consider the difference scheme (7.1.8a), where $\bar{u}_i^{n-1} = u^{n-1}(\bar{x}_i)$, $u^{n-1}(x)$ is the piecewise linear function with nodal values $u_i^{n-1}$, and $\bar{x}_i = x_i - b_i\Delta t$. If $b_i \geq 0$ and $\Delta t$ is sufficiently small, then $\bar{x}_i \in [x_i - 1, x_i]$ and in such a case

$$\bar{u}_i^{n-1} = \frac{b_i\Delta t}{h}u_{i-1}^{n-1} + \frac{h - b_i\Delta t}{h}u_i^{n-1}.$$

So (7.1.8a) can be written as

$$\frac{u_i^n - u_i^{n-1}}{\Delta t} + b_i\frac{u_i^{n-1} - u_{i-1}^{n-1}}{h} - \mu\bar{\delta}_{xx}u_i^n = f_i^n. \tag{7.1.15}$$

This amounts to approximating the convection term by a backward differencing. Similarly, a forward differencing should be adopted to approximate the convection term when $b_i < 0$. Therefore, the difference equation (7.1.8) based on linear interpolations is a kind of upwind scheme. We recall that the usual upwind scheme takes simultaneously the differencing of convection and diffusion terms on either $(n-1)$ or $n$ level, resulting in explicit or implicit schemes respectively.

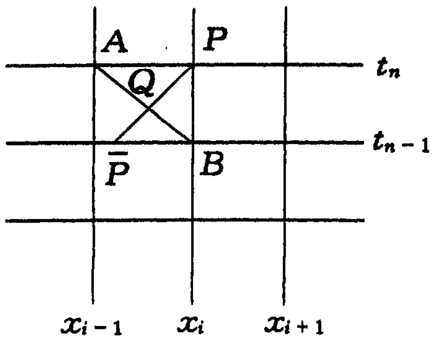



Fig. 7.1.1

Now we try to replace the explicit difference approximation (7.1.6) by an implicit one. Let a grid be given as in Fig. 7.1.1, where the nodes $P = (x_i, t_n)$, $A = (x_{i-1}, t_n)$ and $B = (x_i, t_{n-1})$. If $b_i \geq 0$, then the characteristic line starting from $P$ crosses the net line $t = t_{n-1}$ at

$\bar{P} = (\bar{x}_i, t_{n-1})$ on the left-hand side of $B$. Let the cross point of $\overline{P\bar{P}}$ and the diagonal line $\overline{AB}$ be $Q = (x', t')$. Take $u_A$ and $u_B$ as nodal values to construct a linear interpolation along $\overline{AB}$, and notice

$$\frac{|\overline{AQ}|}{|\overline{QB}|} = \frac{|\overline{AP}|}{|\overline{\bar{P}B}|} = \frac{h}{b_i \Delta t}.$$

Then we have

$$u_Q = \frac{b_i \Delta t}{h + b_i \Delta t} u_A + \frac{h}{h + b_i \Delta t} u_B = \frac{b_i \Delta t}{h + b_i \Delta t} u_{i-1}^n + \frac{h}{h + b_i \Delta t} u_i^{n-1}. \tag{7.1.16}$$

Note $|\overline{P\bar{P}}| = \sqrt{\Delta t^2 + b_i^2 \Delta t^2} = \psi \Delta t$ (cf. (7.1.2)). So it follows from the similarity of $\triangle APQ$ and $\triangle BQ\bar{P}$ that

$$\frac{b_i \Delta t}{h} = \frac{|\overline{\bar{P}Q}|}{|\overline{PQ}|} = \frac{|\overline{P\bar{P}}|}{|\overline{PQ}|} - 1 = \frac{\psi \Delta t}{|\overline{PQ}|} - 1.$$

Thus

$$|\overline{PQ}| = \psi h \Delta t / (h + b_i \Delta t). \tag{7.1.17}$$

Let us employ the following approximation instead of (7.1.6)

$$\psi(x) \frac{\partial u}{\partial \tau} \approx \psi(x) \frac{u_P - u_Q}{|\overline{PQ}|}.$$

Substituting (7.1.16) and (7.1.17) in the right-hand side yields

$$\psi(x_i) \frac{\partial u}{\partial \tau} \approx \frac{u_i^n - u_i^{n-1}}{\Delta t} + b_i \frac{u_i^n - u_{i-1}^n}{h}.$$

Finally we end up with an implicit upwind scheme for $b_i \geq 0$:

$$\frac{u_i^n - u_i^{n-1}}{\Delta t} + b_i \frac{u_i^n - u_{i-1}^n}{h} - \mu \bar{\partial}_{xx} u_i^n = f_i^n. \tag{7.1.18}$$

When $b_i < 0$, the convection term on the left-hand side should be approximated by a forward differencing.

For a steady-state problem

$$b(x) \frac{\partial u}{\partial x} - \mu \frac{\partial^2 u}{\partial x^2} = f(x), \tag{7.1.19}$$

a backward differencing should be adopted to approximate the convection term when $b_i \geq 0$, and a forward differencing when $b_i < 0$. Now let us consider the extension to multidimensional cases. Take the following two-dimensional convection-diffusion equation as an example:

$$\frac{\partial u}{\partial t} + b_1 \frac{\partial u}{\partial x_1} + b_2 \frac{\partial u}{\partial x_2} - \mu \Delta u = f, \tag{7.1.20}$$

where $\mu > 0$ and $\|b\| = \|(b_1, b_2)\| \gg \mu$. The character is now a curve

$$\frac{\mathrm{d}x_1}{\mathrm{d}t} = b_1(x), \ \frac{\mathrm{d}x_2}{\mathrm{d}t} = b_2(x).$$

Place a (triangular or rectangular) grid on $x$-plane, and a uniform grid on $t$-axis with a time step size $\Delta t > 0$. Assume that the characteristic line at a node $(x_{1i}, x_{2j}, t_n)$ crosses the plane $t = t_{n-1}$ at point $\bar{x}_{ij} = (\bar{x}_{1i}, \bar{x}_{2j})$. Evaluate $\bar{u}_{ij}^{n-1}$ as an interpolation in terms of certain neighbouring nodes of $\bar{x}_{ij}$. Then one may perform further discretization analogously as in the one-dimensional case. In particular, if a rectangular grid is used on $x$-plane, then the value of $\bar{u}_{ij}^{n-1}$ can be obtained by either a piecewise bilinear, or a biquadratic interpolation. One may also approximate $b_1 \partial u/\partial x_1$ (resp. $b_2 \partial u/\partial x_2$) along $x_1$-axis (resp. $x_2$-axis) as a one-dimensional convection term. Another strategy is to use the alternating direction method or the locally one-dimensional scheme to reduce the two-dimensional problem into one-dimensional problems along different directions, and then to further discretize the resulting one-dimensional convection terms.

## 7.2 Generalized Upwind Difference Schemes for Steady-state Problems

Let us consider a steady-state convection-diffusion problem:

$$\begin{cases} -\mu \Delta u + b \cdot \nabla u = f, & x \in \Omega, \quad (7.2.1\mathrm{a}) \\ u = 0, & x \in \partial\Omega, \quad (7.2.1\mathrm{b}) \end{cases}$$

where $\Omega \subset \mathrm{R}^2$ is a polygonal region, $\Gamma = \partial\Omega$ is the boundary of $\Omega$, $x = (x_1, x_2)$, $\mu > 0$ is the diffusion coefficient, and $b = b(x) = (b_1(x), b_2(x))$ is the convection velocity. By convection-dominated we mean $0 < \mu \ll \|b\|_\infty$. In this section we extend, from another angle, upwind schemes to generalized upwind difference schemes.

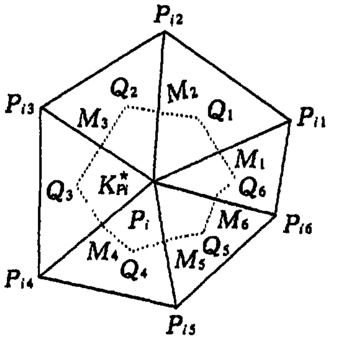



Fig. 7.2.1

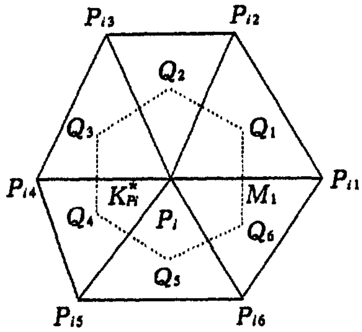



Fig. 7.2.2

### 7.2.1 Construction of the difference schemes

Place a suitable triangulation $T_h = \{K\}$ on $\Omega$, $K \in T_h$ being a triangular element. Let $\{P_i\}_{i=1}^M$ is the set of grid nodes, where $\{P_i\}_{i=1}^N$ are inner nodes and $\{P_i\}_{N+1}^M$ are boundary nodes. Use $h(K)$ and $\rho(K)$ respectively to denote the maximum side length and the diameter of the inscribed circle, and set $h = \max\limits_{K \in T_h} h(K)$. As usual we assume that $T_h$ is a quasi-uniform grid, namely there exist constants $\gamma_1, \gamma_2 > 0$ such that

$$h(K)/\rho(K) \leq \gamma_1, \; h(K)/h \geq \gamma_2, \;\; \forall K \in T_h. \tag{7.2.2}$$

Obviously we have $\Omega_h \equiv \bigcup\limits_{K \in T_h} K = \Omega$.

As in Chapter 3, we construct a barycenter or circumcenter dual grid $T_h^*$ of $T_h$. For any node $P_i$ (cf. Figg. 7.2.1 and 7.2.2), let $P_{ij}$ $(1 \leq j \leq 6)$ be the neighbouring nodes of $P_i$, $M_j$ the midpoint of $\overline{P_iP_{ij}}$, $Q_i$ the barycenter (Fig. 7.2.1) or circumcenter (cf. Fig. 7.2.2 where none of the elements is an obtuse triangle) of $\triangle P_iP_{ij}P_{ij+1}$. Connect successively $M_1, Q_1, M_2, Q_2, \cdots, M_6, Q_6$ and $M_1$ to form a polygon $K^*_{P_i}$ surrounding $P_i$, called a dual element. The entire dual elements constitute a new grid $T_h^* = \{K_{P_i}^*, 1 \le i \le M\}$ on $\Omega$, referred to as a barycenter or a circumcenter dual grid.

Let us introduce the following trial function space:

$$U_h = \{u_h(x) \in C(\Omega) : u_h(x)|_K \text{ is linear }, u_h(x)|_\Gamma = 0\}.$$

A basis function $\phi_i(x)$ of $U_h$ is equal to 1 at the inner node $P_i$ and 0 at other nodes. So any $u_h \in U_h$ has the following expression:

$$u_h(x) = \sum_{i=1}^{N} u_h(P_i)\phi_i(x).$$

The test function space $V_h$ is chosen as the piecewise constant function space related to $T_h^*$, subject to the boundary condition that $v_h(x) = 0$ on $K_{P_i}^*$ for any boundary node $P_i$ and any $v_h \in V_h$. Let $\psi_i(x)$ be the characteristic function of $K_{P_i}^*$. Then $\{\psi_i(x)\}$ is a basis of $V_h$. Denote by $\Pi_h$ and $\Pi_h^*$ the interpolation projectors from $C(\Omega)$ onto $U_h$ and $V_h$ respectively. Then for any $u \in C(\Omega)$

$$\Pi_h u = \sum_{i=1}^{N} u(P_i)\phi_i(x), \ \Pi_h^* u = \sum_{i=1}^{N} u(P_i)\psi_i(x).$$

It is clear that $\Pi_h^*\phi_i(x) = \psi_i(x)$ and consequently $V_h = \text{span}\{\Pi_h^*\phi_h : \phi_h \in U_h\}$.

Set $\Lambda_i = \{j : P_j \text{ is a neighbouring node of } P_i\}$. For adjacent nodes $P_i$ and $P_j$, write $\Gamma_{ij} = \partial K_{P_i}^* \cap \partial K_{P_j}^*$, and denote by $\gamma_{ij}$ the length of $\Gamma_{ij}$ and by $\nu_{ij}$ the unit outer normal direction of $\Gamma_{ij}$ (viewing $\Gamma_{ij}$ as a part of the boundary of $K_{P_i}^*$). Define

$$\beta_{ij} = \int_{\Gamma_{ij}} b(x) \cdot \nu_{ij} \mathrm{d}s. \tag{7.2.3}$$

Then we can divide $\partial K_{P_i}^*$ into a *flow in* part and a *flow out* part according to the sign of $\beta_{ij}$:

$$\begin{cases} (\partial K_{P_i}^*)_- = \bigcup\limits_{\substack{\beta_{ij} \le 0 \\ j \in \Lambda_i}} \Gamma_{ij} & \text{(Flow in)}, \\ (\partial K_{P_i}^*)_+ = \bigcup\limits_{\substack{\beta_{ij} > 0 \\ j \in \Lambda_i}} \Gamma_{ij} & \text{(Flow out)}. \end{cases} \tag{7.2.4}$$

The following facts are apparent

$$\beta_{ij} + \beta_{ji} = 0, \tag{7.2.5}$$

$$|\beta_{ij}| \le C\|b\|_\infty \gamma_{ij}. \tag{7.2.6}$$

Now we try to write (7.2.1) into a weak form. So we multiply (7.2.1a) by $v_h \in V_h$, integrate it on $\Omega$, and apply Green's formula to obtain

$$a(u, v_h) + b(u, v_h) = (f, v_h), \tag{7.2.7}$$

where

$$a(u, v_h) = -\sum_{j=1}^{N} v_h(P_j) \int_{\partial K^*_{P_j}} \mu \frac{\partial u}{\partial \nu} \mathrm{d}s, \tag{7.2.8}$$

$$b(u, v_h) = \sum_{j=1}^{N} v_h(P_j) \int_{\partial K^*_{P_j}} (b \cdot \nu) u \mathrm{d}s - \int_\Omega u v_h \mathrm{div} b \mathrm{d}x, \tag{7.2.9}$$

where $\nu$ is the unit outer normal direction of $\partial K^*_{P_j}$. The key point to construct a generalized upwind scheme lies in how to approximate the line integral of the first term on the right-hand side of (7.2.9). Write

$$\beta^+_{jl} = \max(\beta_{jl}, 0), \ \beta^-_{jl} = \max(-\beta_{jl}, 0). \tag{7.2.10}$$

Taking the *upwind* values leads to the following approximation:

$$\int_{\partial K^*_{P_j}} (b \cdot \nu) u \mathrm{d}s \approx \sum_{l \in \Lambda_j} \{\beta^+_{jl} u(P_j) - \beta^-_{jl} u(P_l)\}. \tag{7.2.11}$$

Let us use the following bilinear form to approximate $b(u, v_h)$:

$$b_h(u, v_h) = \sum_{j=1}^{N} v_h(P_j) \sum_{l \in \Lambda_j} \{\beta^+_{jl} u(P_j) - \beta^-_{jl} u(P_l)\} - \int_\Omega u v_h \mathrm{div} b \mathrm{d}x. \tag{7.2.12}$$

Then we have a generalized upwind difference scheme: Find $u_h \in U_h$ such that

$$a(u_h, v_h) + b_h(u_h, v_h) = (f, v_h), \ \forall v_h \in V_h. \tag{7.2.13}$$

This is equivalent to

$$\sum_{i=1}^{N} a(\phi_i, \psi_j)u_i + \sum_{i=1}^{N} b_h(\phi_i, \psi_j)u_i = (f, \psi_j), \ 1 \le j \le N, \quad (7.2.13)'$$

where $u_i = u_h(P_i)$, and

$$a(\phi_i, \psi_j) = -\int_{\partial K_{P_j}^*} \mu \frac{\partial \phi_i}{\partial \nu} \mathrm{d}s, \quad (7.2.14)$$

$$b_h(\phi_i, \psi_j) = \sum_{l \in \Lambda_j} \{\beta_{jl}^+ \delta_{ij} - \beta_{jl}^- \delta_{il}\} - \int_{K_{P_j}^*} \phi_i \mathrm{div} b \mathrm{d}x, \quad (7.2.15)$$

where $\delta_{ij}$ is the Kronecker delta.

### 7.2.2 Convergence and error estimate

Write the error as

$$u - u_h = (u - \Pi_h u) + (\Pi_h u - u_h). \quad (7.2.16)$$

The first term on the right-hand side is the error of a linear interpolation, satisfying

$$\|u - \Pi_h u\|_1 \le Ch|u|_2. \quad (7.2.17)$$

To estimate the second term, we note that by (3.2.24) and (3.2.46)

$$a(\bar{u}_h, \Pi_h^* \bar{u}_h) \ge \alpha \|\bar{u}_h\|_1^2, \ \ \forall \bar{u}_h \in U_h, \ \alpha > 0, \quad (7.2.18)$$

$$|a(u - \Pi_h u, \Pi_h^* \bar{u}_h)| \le Ch|u|_2 \|\bar{u}_h\|_1, \ \bar{u}_h \in U_h. \quad (7.2.19)$$

Next we turn to the estimation of $b_h(u_h, v_h)$. First we rewrite it as

$$b_h(u_h, v_h) = \sum_{j=1}^{N} \int_{\partial K_{P_j}^*} (\Pi_h^* u_h)^+ v_h (b \cdot \nu_j) \mathrm{d}s - \int_{K_{P_j}^*} u_h v_h \mathrm{div} b \mathrm{d}x, \quad (7.2.20)$$

where $\nu_j$ is the unit outer normal vector, $\Pi_h^* u_h \in V_h$, and $(\Pi_h^* u_h)^+$ is the upwind value of $\Pi_h^* u_h \in V_h$ across the boundary $\partial K_{P_j}^*$. As in §6.2, we may further express $b_h(u_h, v_h)$ as

$$b_h(u_h,v_h)=\sum_{j=1}^{N}\Big\{\int_{K_{P_j}^*}(b\cdot\nabla\Pi_h^* u_h)v_h\mathrm{d}x+\int_{(\partial K_{P_j}^*)_-}[\Pi_h^* u_h]v_h(b\cdot\nu_j)\mathrm{d}s\Big\}$$
$$-\int_{K_{P_j}^*} u_h v_h \mathrm{div} b \mathrm{d}x, \tag{7.2.21}$$

where [ ] denotes the jump value across the boundary, and we note that $\nabla\Pi_h^* u_h = 0$ on $K_{P_j}^*$. Comparing it with (6.2.8a), we find that $b_h(u_h, v_h)$ here turns out to be $a(u_h, v_h)$ there. So actually, the upwind difference scheme (7.2.13) uses the generalized difference method to discretize the diffusion term, and the discontinuous finite element method to discretize the convection term in (7.2.1a). If we assume

$$-\mathrm{div} b(x) \geq \sigma_0 > 0, \tag{7.2.22}$$

then it follows from (6.2.15) that

$$a(\bar{u}_h, \Pi_h^* \bar{u}_h) + b_h(\bar{u}_h, \Pi_h^* \bar{u}_h) \geq \alpha \|\bar{u}_h\|_1^2, \quad \forall \bar{u}_h \in U_h. \tag{7.2.23}$$

Notice

$$(u - \Pi_h u)(P) = 0, \text{ as } P = P_j, P_l.$$

Thus by (7.2.12) we have

$$b_h(u - \Pi_h u, \Pi_h^* \bar{u}_h) = -\int_\Omega (u - \Pi_h u)\Pi_h^* \bar{u}_h \mathrm{div} b \mathrm{d}x.$$

This results in the following estimate

$$|b_h(u - \Pi_h u, \Pi_h^* \bar{u}_h)| \leq C h^2 |u|_2 \|\bar{u}_h\|_0. \tag{7.2.24}$$

Now let $u$ and $u_h$ be the solutions to (7.2.7) and (7.2.13) respectively. Then the subtraction of these two equations leads to an error equation:

$$\begin{aligned} & a(u - u_h, \Pi_h^* \bar{u}_h) + b_h(u - u_h, \Pi_h^* \bar{u}_h) \\ = \ & -(b(u, \Pi_h^* \bar{u}_h) - b_h(u, \Pi_h^* \bar{u}_h)). \end{aligned} \tag{7.2.25}$$

It remains to show the error order of the right-hand side. To this end, define a function

$$H(x) = \begin{cases} 0, & \text{as } x < 0, \\ 1, & \text{as } x \geq 0. \end{cases}$$

Then we have

$$\begin{aligned} & \beta_{jl}^{+}u(P_j) - \beta_{jl}^{-}u(P_l) \\ = \;& (H(\beta_{jl})u(P_j) + (1 - H(\beta_{jl}))u(P_l))\beta_{jl} \\ = \;& \int_{\Gamma_{jl}} b \cdot \nu(H(\beta_{jl})u(P_j) + (1 - H(\beta_{jl}))u(P_l))\mathrm{d}s. \end{aligned}$$

Since $\Gamma_{jl} = \partial K_{P_j}^{*} \cap \partial K_{P_l}^{*}$, the integral along $\Gamma_{jl}$ in the summation on the right-hand side of (7.2.12) appears twice with opposite normal directions $\nu$. Write such two terms together to obtain

$$(v_h(P_j) - v_h(P_l)) \int_{\Gamma_{jl}} b \cdot \nu\{H(\beta_{jl})u(P_j) + (1 - H(\beta_{jl}))u(P_l)\}\mathrm{d}s.$$

So we have

$$\begin{aligned} & b_h(u, v_h) \\ = \;& \frac{1}{2}\sum_{j=1}^{N}\sum_{l\in\Lambda_j}(v_h(P_j) - v_h(P_l)) \\ & \cdot \int_{\Gamma_{jl}} b \cdot \nu\{H(\beta_{jl})u(P_j) + (1 - H(\beta_{jl}))u(P_l)\}\mathrm{d}s \\ & - \int_{\Omega} uv_h \mathrm{div} b \mathrm{d}x. \end{aligned} \tag{7.2.26a}$$

Similarly by (7.2.9) we have

$$\begin{aligned} b(u, v_h) = \;& \frac{1}{2}\sum_{j=1}^{N}\sum_{l\in\Lambda_j}(v_h(P_j) - v_h(P_l)) \int_{\Gamma_{jl}} b \cdot \nu u \mathrm{d}s \\ & - \int_{\Omega} uv_h \mathrm{div} b \mathrm{d}x. \end{aligned} \tag{7.2.26b}$$

Subtracting these two equations yields

$$\begin{aligned} & b(u,v_h)-b_h(u,v_h) \\ = \ & \frac{1}{2}\sum_{j=1}^{N}\sum_{l\in\Lambda_j}(v_h(P_j)-v_h(P_l))\int_{\Gamma_{jl}} b\cdot\nu\{H(\beta_{jl})(u-u(P_j)) \\ & +(1-H(\beta_{jl}))(u-u(P_l))\}\mathrm{d}s. \end{aligned}$$

As in the deduction of (3.2.46) one can show that

$$\begin{aligned} & |b(u,\Pi_h^*\bar{u}_h)-b_h(u,\Pi_h^*\bar{u}_h)| \\ \leq \ & C\|b\|_\infty h|u|_2|\bar{u}_h|_1, \quad \forall \bar{u}_h\in U_h. \end{aligned} \tag{7.2.27}$$

By the positive definiteness (7.2.23), the continuity (7.2.19), the consistency (7.2.27), and as in the error estimation in §3.2, one can easily prove the following theorem.

**Theorem 7.2.1** *Assume that $b$ satisfies (7.2.22), that $b\in H^1(\Omega)\times H^1(\Omega)\cap L^\infty(\Omega)\times L^\infty(\Omega)$, that $f\in L^2(\Omega)$, and that the solution $u$ of (7.2.1) belongs to $H^2(\Omega)\cap H_0^1(\Omega)$. Then, there holds the following error estimate:*

$$\|u-u_h\|_1\leq Ch|u|_2.$$

**Remark** The above theorem requires that $b$ satisfies (7.2.22). But this is not an essential restriction. In fact, if necessary, we can always validate (7.2.22) by performing the following transformation

$$\bar{u}=ue^{\omega t},\ \omega=\sigma_0+\frac{1}{2}|\mathrm{div}b|.$$

### 7.2.3 Extreme value theorem and uniform convergence

First let us have a look at the signs of the coefficients in the generalized difference equation (7.2.13). In Fig 7.2.3, $\triangle P_iP_jP_k$ and $\triangle P_iP_{k'}P_j$ are two adjacent triangle elements, and $Q$ and $Q'$ are their barycenters (or circumcenters) respectively. Write $K_Q=\triangle P_iP_jP_k$ and $K_{Q'}=\triangle P_iP_{k'}P_j$. Let $m_i$, $m_j$, and $m_k$ be the three midpoints

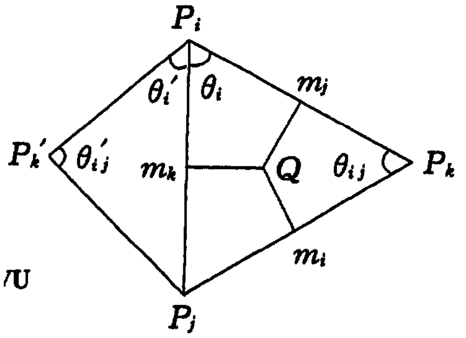



Fig. 7.2.3

of the sides of $K_Q$, and let $\theta_{ij}$, $\theta'_{ij}$, $\theta_i$ and $\theta'_i$ be the inner angles of the elements.

Let us try to evaluate

$$a(\phi_i, \psi_j) = -\int_{\partial K^*_{P_j}} \mu \frac{\partial \phi_i}{\partial \nu} \mathrm{d}s.$$

If $i \notin \Lambda_j$, then it is obvious that

$$a(\phi_i, \psi_j) = 0.$$

On the other hand, if $i \in \Lambda_j$ and $i \neq j$, then

$$a(\phi_i, \psi_j) = -\int_{\partial K^*_{P_j} \cap K_Q} \mu \frac{\partial \phi_i}{\partial \nu} \mathrm{d}s - \int_{\partial K^*_{P_j} \cap K_{Q'}} \mu \frac{\partial \phi_i}{\partial \nu} \mathrm{d}s.$$

Apply Green's formula on $\triangle m_i Q m_k$ to get

$$-\int_{\partial K^*_{P_j} \cap K_Q} \mu \frac{\partial \phi_i}{\partial \nu} \mathrm{d}s = \int_{\overline{m_k m_i}} \mu \frac{\partial \phi_i}{\partial \nu} \mathrm{d}s = \int_{\overline{m_k m_i}} \mu \nabla \phi_i \nu \mathrm{d}s.$$

Note that $\phi_i$ is actually an area coordinate on $\triangle P_i P_j P_k$. Hence according to the expression of area coordinates by rectangular coordinates, the length of $\nabla \phi_i$ is

$$|\overline{P_j P_k}|/(2S_Q), \ (S_Q \text{ being the area of } K_Q)$$

and it is perpendicular to $\overline{P_jP_k}$, pointing towards $P_i$. $\nu$ is a unit vector perpendicular to $\overline{m_im_k}$ (i.e. to $\overline{P_iP_k}$) and pointing towards $P_j$. Therefore we have

$$-\int_{\partial K^*_{P_j}\cap K_Q}\mu\frac{\partial\phi_i}{\partial\nu}ds = -\frac{\mu|\overline{P_jP_k}|\,|\overline{P_kP_i}|}{4S_Q}\cos\theta_{ij} = -\frac{\mu}{2}\text{ctg}\theta_{ij}.$$

Similarly

$$-\int_{\partial K^*_{P_j}\cap K_{Q'}}\mu\frac{\partial\phi_i}{\partial\nu}ds = -\frac{\mu}{2}\text{ctg}\theta'_{ij}.$$

So, for $i\in\Lambda_j$ but $i\neq j$, we have

$$a(\phi_i,\psi_j) = -\frac{\mu}{2}(\text{ctg}\theta_{ij}+\text{ctg}\theta'_{ij}). \tag{7.2.28}$$

In the case of $i=j$, i.e. $\phi_i=\phi_j$, we use Green's formula on the quadrilateral $P_jm_iQm_k$ to obtain

$$\begin{aligned}
&-\int_{\partial K^*_{P_j}\cap K_Q}\mu\frac{\partial\phi_j}{\partial\nu}ds\\
=\;& \int_{\overline{P_jm_i}}\mu\frac{\partial\phi_j}{\partial\nu}ds+\int_{\overline{m_kP_j}}\mu\frac{\partial\phi_j}{\partial\nu}ds\\
=\;& \frac{\mu|\overline{P_kP_i}|\,|\overline{P_kP_j}|}{4S_Q}\cos\theta_{ij}+\frac{\mu|\overline{P_kP_i}|\,|\overline{P_iP_j}|}{4S_Q}\cos\theta_i\\
=\;& \frac{\mu}{2}(\text{ctg}\theta_{ij}+\text{ctg}\theta_i).
\end{aligned}$$

Recall that $\theta_{ij}$ and $\theta_i$ are two inner angles of the element $K_Q$, of which the vertexes are not $P_j$. Thus we have

$$a(\phi_j,\psi_j) = -\int_{\partial K^*_{P_j}}\mu\frac{\partial\phi_j}{\partial\nu}ds = \frac{\mu}{2}\sum_{l\in\Lambda_j}(\text{ctg}\theta_l+\text{ctg}\theta'_l). \tag{7.2.29}$$

Now we assume $T_h$ is an acute triangulation, namely there exists a constant $\epsilon_0>0$ such that any inner angle $\theta$ of an element of $T_h$ satisfies $\theta\le\frac{\pi}{2}-\epsilon_0$. Then none of the angles appearing in (7.2.28) and (7.2.29) is greater than $\frac{\pi}{2}-\epsilon_0$, and hence

$$\begin{cases} a(\phi_i,\phi_j)\le 0, & \text{when } i\neq j,\\ a(\phi_i,\phi_j)\ge\delta_0>0, & \text{when } i=j.\end{cases} \tag{7.2.30}$$

To evaluate $b_h(\phi_i, \psi_j)$ we first note that by (7.2.15)

$$b_h(\phi_i, \psi_j) = 0, \text{ when } i \text{ and } j \text{ are not adjacent}, \tag{7.2.31}$$

$$b_h(\phi_j, \psi_j) = \sum_{l \in \Lambda_j} \beta_{jl}^+ - \int_{K_{P_j}^*} \phi_j \mathrm{div} b \mathrm{d}x \geq 0. \tag{7.2.32}$$

When $i$ and $j$ are adjacent but not identical we have

$$\begin{aligned} b_h(\phi_i, \psi_j) &= -\beta_{ji}^- - \int_{K_{P_j}^*} \phi_i \mathrm{div} b \mathrm{d}x \\ &= -\int_{(\partial K_{P_j}^*)_-} b \cdot \nu \mathrm{d}s - \int_{K_{P_j}^*} \phi_i \mathrm{div} b \mathrm{d}x. \end{aligned}$$

Notice that for a quasi-uniform grid $T_h$, the area $S_{P_j}^*$ of $K_{P_j}^*$ is of order $h^2$, that is, there exist positive constants $\alpha_1$ and $\alpha_2$ such that $\alpha_1 h^2 \leq S_{P_j}^* \leq \alpha_2 h^2$. Thus for sufficiently small $h$

$$0 \leq b_h(\phi_i, \psi_j) \leq \alpha_0 h \ \ (\alpha_0 > 0). \tag{7.2.33}$$

We also observe that by (7.2.15) and Green's formula

$$\sum_{i=1}^{N} b_h(\phi_i, \psi_j) = \sum_{l \in \Lambda_i} \beta_{jl} - \int_{K_{P_j}^*} \mathrm{div} b \mathrm{d}x = 0. \tag{7.2.34}$$

Now we are ready to study the extreme value property of (7.2.13) or (7.2.13)$'$. Set

$$a_{ij} = a(\phi_i, \psi_j) + b_h(\phi_i, \psi_j), \ b_j = (f, \psi_j).$$

Write (7.2.13)$'$ in the form

$$\sum_{i=1}^{N} a_{ij} u_i = b_j, \ j = 1, 2, \cdots, N.$$

It follows from (7.2.28)$-$(7.2.34) that if $T_h$ is an acute triangulation and $h$ is sufficiently small, then $a_{ij} \leq 0$ for $i \neq j$, $a_{jj} > 0$, and

$$\sum_{j=1}^{N} a_{ij} \geq 0, \ 1 \leq j \leq N. \tag{7.2.35a}$$

Observe that (7.2.29) remains to be true even if one of the neighbouring nodes of $P_j$ is on the boundary. But (7.2.28) is valid only when $i \in \Lambda_j$ and $P_i$ is not a boundary node. This term vanishes if $P_i$ is indeed a boundary node. These observations imply the existence of $d_0 > 0$ and $j_0$ such that

$$\sum_{i=1}^{N} a_{ij_0} \geq d_0. \tag{7.2.35b}$$

Hence, the extreme value theorem of difference equations (cf. [A-27] and [C-7]) gives

$$\max_{1\leq i\leq N} |u_i| \leq C \max_{1\leq i\leq N} |b_i|. \tag{7.2.36}$$

Since $u_h(x)$ is a piecewise linear function, furthermore we have (cf. [A-17])

**Theorem 7.2.2** *(Extreme value theorem) Let $T_h$ be an acute triangulation, and let $b$ satisfy the conditions of Theorem 7.2.1, then*

$$\|u_h\|_\infty \leq C \max_i |(f, \psi_i)|. \tag{7.2.37}$$

By the properties of the matrix $A = (a_{ij})$ mentioned above, we also know that $A^{-1}$ is a non-negative nonsingular matrix ([B-93]). So if the right-hand side term $f$ and the boundary value $g$ are non-negative, then the difference solution is non-negative as well. Thus the difference solution will not have unnecessary oscillations.

Next let us turn to the estimation in maximum norm. Let $u$ and $u_h$ be the solutions to the convection-diffusion equation and the difference equation respectively. Use the triangular inequality to get

$$\|u - u_h\|_\infty \leq \|u - \Pi_h u\|_\infty + \|\Pi_h u - u_h\|_\infty. \tag{7.2.38}$$

Take a fixed $p > 2$ (2 being the dimension of the plane), then by virtue of the Sobolev imbedding theorem we have

$$\|u - \Pi_h u\|_\infty \leq C\|u - \Pi_h u\|_{1,p}.$$

It follows from the interpolation approximation theory in Sobolev spaces ([B-17]) that

$$\|u - \Pi_h u\|_{1,p} \leq Ch|u|_{2,p}.$$

Thus

$$\|u - \Pi_h u\|_\infty \le Ch|u|_{2,p}. \tag{7.2.39}$$

To estimate the second term of the right-hand side of (7.2.38), we notice that $\Pi_h u - u_h$ satisfies the following equation

$$a(\Pi_h u - u_h, \Pi_h^* \phi_j) + b_h(\Pi_h u - u_h, \Pi_h^* \phi_j) = r_j,$$

where

$$\begin{aligned} r_j = \;& a(\Pi_h u - u, \Pi_h^* \phi_j) + b_h(\Pi_h u - u, \Pi_h^* \phi_j) \\ & -(b(u, \Pi_h^* \phi_j) - b_h(u, \Pi_h^* \phi_j)). \end{aligned}$$

A combination of (7.2.19), (7.2.24) and (7.2.26) leads to an estimation of $r_j$:

$$|r_j| \le Ch|u|_2.$$

Thus in terms of the extreme value property (7.2.37) we have

$$\|\Pi_h u - u_h\|_\infty \le Ch|u|_2. \tag{7.2.40}$$

Connecting (7.2.38)–(7.2.40) yields (cf. [A-17])

**Theorem 7.2.3** *Under the assumptions of Theorem 7.2.2, the following maximum estimate holds for any* $p > 2$

$$\|u - u_h\|_\infty \le Ch|u|_{2,p}. \tag{7.2.41}$$

### 7.2.4 Mass conservation

Consider a conservative equation:

$$\begin{cases} -\mu\Delta u + \nabla(bu) = f, & \text{in } \Omega, \qquad (7.2.42a) \\ u = 0, & \text{on } \partial\Omega. \qquad (7.2.42b) \end{cases}$$

The corresponding generalized upwind equation is

$$a(u_h, \psi_j) + b_h(u_h, \psi_j) = (f, \psi_j), \tag{7.2.43}$$

where

$$a(u_h, \psi_j) = -\int_{\partial K_{P_j}^*} \mu \frac{\partial u_h}{\partial \nu} \mathrm{d}s,$$

$$b_h(u_h, \psi_j) = \sum_{l \in \Lambda_j} [\beta_{jl}^+ u_h(P_j) - \beta_{jl}^- u_h(P_l)].$$

Integrate (7.2.42a) on $\Omega$ and exploit Green's formula to obtain

$$\int_{\partial\Omega} (-\mu \frac{\partial u}{\partial \nu} + b \cdot \nu u) \mathrm{d}s = \int_\Omega f \mathrm{d}x. \tag{7.2.44}$$

Here the left-hand side is the mass "flowing out" of $\Omega$ through the boundary $\partial\Omega$, and the right-hand side the mass out of the source $f$. Therefore, the equation (7.2.44) describes a mass conservation. Next, finding the sum of the difference equation (7.2.43) gives

$$-\sum_{j=1}^{N} \int_{\partial K_{P_j}^*} \mu \frac{\partial u_h}{\partial \nu} \mathrm{d}s + \sum_{j=1}^{N} \int_{\Gamma_{jl}} b \cdot \nu [H(\beta_{jl}) u_j + (1 - H(\beta_{jl})) u_l] \mathrm{d}s$$
$$= \int_{\Omega_h^*} f \mathrm{d}x,$$

where $\Omega_h^* = \bigcup_{j=1}^{N} K_{P_j}^*$. Obviously the left-hand side terms cancel each other on the inner boundaries, so the above equation becomes:

$$-\int_{\partial\Omega_h^*} \mu \frac{\partial u_h}{\partial \nu} \mathrm{d}s + \sum_{\Gamma_{ij} \in \partial\Omega_h^*} \int_{\Gamma_{ij}} b \cdot \nu [H(\beta_{ij}) u_i + (1 - H(\beta_{ij})) u_j] \mathrm{d}s$$
$$= \int_{\Omega_h^*} f \mathrm{d}x. \tag{7.2.45}$$

This is precisely a discrete mass conservation law.

## 7.3 Generalized Upwind Difference Schemes for Non–steady-state Problems

In this and the next sections we discuss generalized upwind difference solutions of the following non–steady-state convection-diffusion problem:

$$\begin{cases} \dfrac{\partial u}{\partial t} = \mu \Delta u - b \cdot \nabla u + f, & x \in \Omega,\ 0 < t \le T, & (7.3.1a) \\ u(x,t) = 0, & x \in \Gamma = \partial\Omega,\ 0 < t \le T, & (7.3.1b) \\ u(x,0) = u_0(x), & x \in \Omega, & (7.3.1c) \end{cases}$$

where $x = (x_1, x_2)$, $\mu > 0$ is the diffusion coefficient, $b = (b_1(x), b_2(x))$ is the convection speed. Usually $u = u(x,t)$ stands for the density or the temperature, and $f(x)$ some kind of source.

### 7.3.1 Construction of difference schemes

As in §7.2, we place a triangulation $T_h$ and a dual grid $T_h^*$ on $\Omega$, introduce a trial function space $U_h$ and a test function space $V_h$, and define the following bilinear forms:

$$a(u, v_h) = -\sum_{j=1}^{N} v_h(P_j) \int_{\partial K_{P_j}^*} \mu \frac{\partial u}{\partial \nu} \mathrm{d}s, \tag{7.3.2}$$

$$b(u, v_h) = \sum_{j=1}^{N} v_h(P_j) \int_{\partial K_{P_j}^*} (b \cdot \nu) u \mathrm{d}s - \int_\Omega u v_h \mathrm{div} b \mathrm{d}x, \tag{7.3.3}$$

$$b_h(u, v_h) = \sum_{j=1}^{N} v_h(P_j) \sum_{l \in \Lambda_j} \{\beta_{jl}^+ u(P_j) - \beta_{jl}^- u(P_l)\} - \int_\Omega u v_h \mathrm{div} b \mathrm{d}x, \tag{7.3.4}$$

where

$$\beta_{jl}^+ = \max(\beta_{jl}, 0),\ \beta_{jl}^- = \max(-\beta_{jl}, 0), \tag{7.3.5}$$

$$\beta_{ij} = \int_{\Gamma_{ij}} b(x) \cdot \nu_{ij} \mathrm{d}s. \tag{7.3.6}$$

A weak form of (7.3.1) is: Find $u \in C^1([0,T]; H_0^1(\Omega))$ such that

$$\left(\frac{\partial u}{\partial t}, v\right) + a(u,v) + b(u,v) = (f, v),\ v \in H_0^1(\Omega). \tag{7.3.7}$$

A semi-discrete generalized upwind difference scheme for (7.3.7) is: Find $u_h(\cdot, t) \in U_h$ such that

$$\left(\frac{\partial u_h}{\partial t}, v_h\right) + a(u_h, v_h) + b_h(u_h, v_h) = (f, v_h),\ v_h \in V_h. \tag{7.3.8}$$

Choose a time step size $\tau = T/N$ and the nodes $t_k = k\tau$ $(k = 0, 1, \cdots, N)$. A fully-discrete, explicit, generalized upwind difference scheme is: Find $u_h^k \in U_h$ such that

$$((u_h^k - u_h^{k-1})/\tau, v_h) + a(u_h^{k-1}, v_h) + b_h(u_h^{k-1}, v_h) = (f, v_h), \quad \forall v_h \in V_h. \tag{7.3.9}$$

In particular, choose $v_h$ as a basis function $\psi_j$ of $V_h$, and set

$$\bar{\partial}_t u_h^k = (u_h^k - u_h^{k-1})/\tau.$$

Then we have (cf. Figs. 7.2.1 and 7.2.2)

$$\int_{K_{P_j}^*} \bar{\partial}_t u_h^k \mathrm{d}x = \sum_{l\in\Lambda_j} \mu \frac{u_{jl}^{k-1} - u_j^{k-1}}{|\overline{P_j P_{jl}}|} \gamma_{jl} - \sum_{l\in\Lambda_j} \{\beta_{jl}^+ u_j^{k-1} - \beta_{jl}^- u_l^{k-1}\} - \int_{K_{P_j}^*} u_h^{k-1} \mathrm{div} b \mathrm{d}x + \int_{K_{P_j}^*} f \mathrm{d}x, \tag{7.3.10}$$

where $\gamma_{jl}$ is the length of $\Gamma_{jl}$. The initial value $u_h^0$ satisfies

$$(u_h^0, \psi_j) = (u_0, \psi_j). \tag{7.3.11}$$

Let us introduce the symbol $u_h^{k,\theta} = \theta u_h^k + (1-\theta) u_h^{k-1}$ $(0 \le \theta \le 1)$, then we can define a more general weighted scheme:

$$(\bar{\partial}_t u_h^k, v_h) + a(u^{k,\theta}, v_h) + b_h(u^{k,\theta}, v_h) = (f, v_h), \quad \forall v_h \in V_h. \tag{7.3.12}$$

This corresponds to a backward difference scheme if $\theta = 1$ and a Crank-Nicolson scheme if $\theta = \frac{1}{2}$. We remark that the difference equation (7.1.8) now can be written as

$$(\bar{\partial}_t u_h^k, v_h) + a(u_h^k, v_h) + b_h(u_h^{k-1}, v_h) = (f, v_h), \quad \forall v_h \in V_h. \tag{7.3.13}$$

**Remark 1** Baba and Tabata ([B-3]) proposed a upwind finite element scheme, which can be written in our symbols as

$$(\bar{\partial}_t u_h^k, \Pi_h^* \bar{u}_h) + a(u^{k,\theta}, \bar{u}_h) + b_h(u^{k,\theta}, \Pi_h^* \bar{u}_h) = (f, \Pi_h^* \bar{u}_h),$$

$$\forall \bar{u}_h \in U_h. \tag{7.3.14}$$

**Remark 2** In order to simplify the computation of the difference schemes, one often replaces $\bar{\partial}_t$ by $\Pi_h^* \bar{\partial}_t$, and $(\bar{\partial}_t u_h^{k-1}, v_h)$ by $(\Pi_h^* \bar{\partial}_t u_h^{k-1}, v_h)$. In such a case, we have

$$(\Pi_h^* \bar{\partial}_t u_h^{k-1}, \psi_j) = S_{P_j}^* (u_j^k - u_j^{k-1})/\tau,$$

where $S_{P_j}^*$ is the area of $K_{P_j}^*$.

### 7.3.2 Convergence and error estimate

In this subsection, we consider the case that $T_h^*$ is a barycenter dual grid. Taking $v_h = \Pi_h^* u_h$ in the semi-discrete scheme (7.3.8) yields

$$\frac{\partial}{\partial t} |||u_h|||_0^2 + a(u_h, \Pi_h^* u_h) + b_h(u_h, \Pi_h^* u_h) = (f, \Pi_h^* u_h), \tag{7.3.15}$$

where (cf §5.1)

$$|||u_h|||_0^2 = (u_h, \Pi_h^* u_h). \tag{7.3.16}$$

By §3.2, there is a constant $\alpha > 0$ such that

$$a(u_h, \Pi_h^* u_h) \geq \alpha |u_h|_1^2. \tag{7.3.17}$$

It follows from (7.2.26a) that

$$\begin{aligned} b_h(u_h, \Pi_h^* \bar{u}_h) = \ & \frac{1}{2} \sum_{j=1}^{N} \sum_{l \in \Lambda_j} (\bar{u}_h(P_j) - \bar{u}_h(P_l)) \cdot \\ & \cdot \int_{\Gamma_{jl}} b \cdot \nu \{ H(\beta_{jl}) u_h(P_j) + (1 - H(\beta_{jl})) u(P_l) \} \mathrm{d}s \\ & - \int_\Omega u_h \Pi_h^* \bar{u}_h \operatorname{div} b \mathrm{d}x. \end{aligned}$$

Evidently

$$\begin{aligned} & |b_h(u_h, \Pi_h^* \bar{u}_h)| \\ \leq \ & C_1 \|b\|_\infty |\bar{u}_h|_1 \|u_h\|_0 + C_2 \|\operatorname{div} b\|_\infty \|\bar{u}_h\|_0 \|u_h\|_0. \end{aligned} \tag{7.3.18}$$

Here we have used the equivalence of $\|\Pi_h^* \bar{u}_h\|_0$ and $\|\bar{u}_h\|_0$ (cf. §5.1). By virtue of (7.3.15), (7.3.17), (7.3.18) and the $\epsilon$-inequality we have

$$\frac{\partial}{\partial t}|||u_h|||_0^2 + \alpha_0 |u_h|_1^2 \leq C(\|u_h\|_0^2 + \|f\|_0^2), \tag{7.3.19}$$

where $\alpha_0 > 0$ is a constant. Note that $|||u_h|||_0$ and $\|u_h\|_0$ are equivalent, view $\phi(t) = |||u_h|||_0$ as a unknown function, and integrate the above inequality, then we have

$$\begin{aligned} &\|u_h(t)\|_0^2 + \alpha_0 \int_0^t |u_h|_1^2 \mathrm{d}t \\ \leq\ & C\Big(\|u_h^0\|_0^2 + \int_0^t \|f\|_0^2 \mathrm{d}t\Big),\ 0 \leq t \leq T. \end{aligned} \tag{7.3.20}$$

Therefore the semi-discrete solution $u_h(t)$ is stable respect to the initial value and the right-hand side.

Now we turn to deal with the error of the semi-discretization. Write

$$u - u_h = \rho_h + e_h, \tag{7.3.21}$$

$$\rho_h = u - \Pi_h u,\ e_h = \Pi_h u - u_h,$$

Then $e_h$ satisfies

$$\begin{aligned} &\Big(\frac{\partial e_h}{\partial t}, v_h\Big) + a(e_h, v_h) + b_h(e_h, v_h) \\ =\ & (\Pi_h u_t - u_t, v_h) - a(\rho_h, v_h) + (b_h(\Pi_h u, v_h) - b(u, v_h)). \end{aligned} \tag{7.3.22}$$

Take $v_h = \Pi_h^* e_h$, and employ (7.2.19), (7.2.24), (7.2.27) and the $\epsilon$-inequality, then we obtain an inequality similar to (7.3.19):

$$\frac{\partial}{\partial t}|||e_h|||_0^2 + \alpha_0 |e_h|_1^2 \leq C(|||e_h|||_0^2 + h^2|u|_2^2 + h^2|u_t|_1^2).$$

Integrating the above inequality leads to another inequality analogous to (7.3.20):

$$\begin{aligned} &\|e_h(t)\|_0^2 + \alpha_0 \int_0^t |e_h(t)|_1^2 \mathrm{d}t \\ \leq\ & Ch^2\Big(\|e_h^0\|_0^2 + \int_0^t |u(t)|_2^2 \mathrm{d}t + \int_0^t |u_t(t)|_1^2 \mathrm{d}t\Big). \end{aligned} \tag{7.3.23}$$

Connect (7.3.21)-(7.2.23) and note $e_h^0 = 0$, then we have the following error estimate for the semi-discrete solution:

$$\begin{aligned} &\|u - u_h\|_0^2 + \int_0^t |u - u_h|_1^2 \mathrm{d}t \\ \le\ & Ch^2\Big(\|u\|_2^2 + \int_0^t |u|_2^2 \mathrm{d}t + \int_0^t |u_t|_1^2 \mathrm{d}t\Big). \end{aligned} \tag{7.3.24}$$

Next we turn to deal with the fully-discrete scheme. Choose $\theta = 1$ in (7.3.12) to obtain

$$\begin{aligned} &((u_h^k - u_h^{k-1})\tau^{-1}, v_h) + a(u_h^k, v_h) + b_h(u_h^k, v_h) \\ =\ & (f, v_h), \quad \forall v_h \in V_h. \end{aligned}$$

As before, we only need to estimate $e_h^k = \Pi_h u^k - u_h^k$, which satisfies the equation

$$\begin{aligned} &(\bar{\partial}_t e_h^k, v_h) + a(e_h^k, v_h) + b_h(e_h^k, v_h) \\ =\ & (\Pi_h u_t^k - u_t^k, v_h) + (\Pi_h(\bar{\partial}_t u^k - u_t^k), v_h) \\ & + a(-\rho_h^k, v_h) + (b_h(\Pi_h u^k, v_h) - b(u^k, v_h)) \\ =\ & R_{1h}^k + R_{2h}^k + R_{3h}^k + R_{4h}^k. \end{aligned} \tag{7.3.25}$$

Let $v_h = \Pi_h^* e_h^k$. Then

$$|R_{1h}^k| \le Ch^2 |u_t^k|_2 \|e_h^k\|_0 \le Ch^2(\|e_h^k\|_0^2 + |u_t^k|_2^2), \tag{7.3.26a}$$

$$|R_{2h}^k| \le C\tau \|u_{tt}^k\|_0 \|e_h^k\|_0 \le Ch^2(\|e_h^k\|_0^2 + \|u_{tt}^k\|_0^2), \tag{7.3.26b}$$

$$|R_{3h}^k| \le Ch|u^k|_2 |e_h^k|_1 \le C(\epsilon^2 |e_h^k|_1^2 + \epsilon^{-2} h^2 |u^k|_2^2). \tag{7.3.26c}$$

It follows from (7.2.27) that

$$|R_{4h}^k| \le Ch|u^k|_2 |e_h^k|_1 \le C(\epsilon^2 |e_h^k|_1^2 + \epsilon^{-2} h^2 |u^k|_2^2). \tag{7.3.26d}$$

To estimate the left-hand side of (7.3.25), we note

$$\begin{aligned} &\frac{1}{\tau}(e_h^k, \Pi_h^* e_h^k) - \frac{1}{\tau}(e_h^{k-1}, \Pi_h^* e_h^k) + a(e_h^k, \Pi_h^* e_h^k) + b_h(e_h^k, \Pi_h^* e_h^k) \\ \ge\ & \frac{1}{\tau}|||e_h^k|||_0^2 - \frac{1}{\tau}|||e_h^k|||_0 |||e_h^{k-1}|||_0 + \alpha_0 |e_h^k|_1^2 \\ \ge\ & \frac{1}{2\tau}(|||e_h^k|||_0^2 - |||e_h^{k-1}|||_0^2) + \alpha_0 |e_h^k|_1^2. \end{aligned}$$

Choose $\epsilon$ sufficiently small such that the sum of the coefficients of $|e_h^k|_1^2$ in (7.3.26c,d) is less than $\alpha_0$, then there exists an $\alpha_1 > 0$ such that

$$\begin{aligned} & \frac{1}{2\tau}(|||e_h^k|||_0^2 - |||e_h^{k-1}|||_0^2) + \alpha_1 |e_h^k|_1^2 \\ \leq \; & Ch^2(\|e_h^k\|_0^2 + |u^k|_2^2 + |u_t^k|_2^2 + \|u_{tt}^k\|_0^2). \end{aligned} \tag{7.3.27}$$

Find the sum with respect to $k$, notice the equivalence of $|||e_h^k|||_0$ and $\|e_h^k\|_0$, and note $k\tau \leq T$ and $e_h^0 = 0$, then we have

**Theorem 7.3.1** *Suppose $u \in C^2([0,T]; H^2(\Omega))$, then the backward difference solution $\{u_h^n\}$ satisfies the following error estimate:*

$$\max_{1\leq k\leq N} \|u(t_k) - u_h^k\|_0 \leq Ch\|u\|_{X_1}, \tag{7.3.28}$$

*where*

$$\begin{cases} X_1 = C^1([0,T]; H^2(\Omega)) \cap C^2([0,T]; L^2(\Omega)), \\ \|u\|_{X_1} = \|u\|_{C^1([0,T];H^2(\Omega))} + \|u\|_{C^2([0,T];L^2(\Omega))}. \end{cases} \tag{7.3.29}$$

## 7.4 Highly Accurate Generalized Upwind Schemes

The generalized upwind schemes introduced by now are all of first order accuracy. In this section we combine generalized difference methods with higher order upwind schemes in Chapter 6 to construct a class of generalized upwind schemes for convection-dominated diffusion equations, which, in principle, can reach arbitrarily high order accuracy.

### 7.4.1 Construction of the difference schemes

Again we try to solve (7.3.1) and keep all the assumptions on the coefficients and the solution region there. As in §7.2, we assume $T_h$ is a quasi-uniform triangulation, and $T_h^*$ a barycenter or circumcenter dual grid. Here and below, the meanings of the symbols are the same as in the last section unless otherwise stated. For simplicity we only construct linear element scheme. Extensions to higher order elements are self-evident.

As before we construct a trial function space

$$U_h = \{u_h(x) \in C(\bar{\Omega}) : u_h(x)|_K \text{ is linear } \forall K \in T_h \text{ and } u_h|_\Gamma = 0\}.$$

For each inner node $P_i$, there is a basis function $\phi_i(x)$ of $U_h$, which equals to 1 at $P_i$ (cf. Figg. 7.2.1, 7.2.2), and equals to 0 at other nodes. A function $u_h \in U_h$ has the expression

$$u_h(x) = \sum_{i=1}^{N} u_h(P_i)\phi_i(x).$$

The test function space relative to the dual grid $T_h^*$ is given as

$$\begin{aligned} V_h = \ & \{v_h(x) : v_h(x) \text{ is piecewise constant on } \mathrm{T_h^*}; \\ & v_h \text{ vanishes on } K_{P_i}^* \text{ when } P_i \text{ is a boundary node}\}. \end{aligned}$$

For $j = 1, 2, \cdots, N$, the basis function $\psi_j(x)$ is chosen as the characteristic function of $K_{P_j}^*$. Let $\Pi_h$ and $\Pi_h^*$ be the interpolation projectors defined in §7.2. Then for any $u \in C(\overline{\Omega})$ we have

$$\Pi_h u = \sum_{i=1}^{N} u(P_i)\phi_i(x), \ \Pi_h^* u = \sum_{j=1}^{N} u(P_j)\psi_j(x).$$

Let $u_h(\cdot, t) \in U_h$ $(0 < t \leq T)$, the approximation solution of (7.3.1), satisfy formally

$$\int_{K_{P_j}^*} \Big[\frac{\partial u_h}{\partial t} - \mu\Delta u_h + b\cdot\nabla(\Pi_h^* u_h)\Big] v_h \mathrm{d}x = \int_{K_{P_j}^*} f v_h \mathrm{d}x, \ v_h \in V_h. \tag{7.4.1}$$

In terms of Green's formula we have

$$-\mu\int_{K_{P_j}^*} \Delta u_h \cdot v_h \mathrm{d}x = -\mu\int_{\partial K_{P_j}^*} \frac{\partial u_h}{\partial \nu} v_h \mathrm{d}s. \tag{7.4.2}$$

As in Chapter 6, we apply Green's formula to the convection term in the following fashion:

$$\begin{aligned}&\int_{K^*_{P_j}} b\cdot\nabla(\Pi^*_h u_h)v_h \mathrm{d}x\\ =\ & -\int_{K^*_{P_j}}(\Pi^*_h u_h)\mathrm{div}(bv_h)\mathrm{d}x+\int_{\partial K^*_{P_j}} b\cdot\nu(\Pi^*_h u_h)v_h\mathrm{d}s.\end{aligned} \tag{7.4.3}$$

Denote by $(\Pi^*_h u_h)^+$ and $(\Pi^*_h u_h)^-$ the upwind and downwind values (cf. §6.2) of $\Pi^*_h u_h$ across the boundary $\partial K^*_{P_j}$. Then, it follows from $\nabla(\Pi^*_h u_h)=0$ and (7.4.3) that

$$\begin{aligned}&\int_{K^*_{P_j}}(\Pi^*_h u_h)\mathrm{div}(bv_h)\mathrm{d}x\\ =\ & \int_{(\partial K^*_{P_j})_-}(b\cdot\nu)(\Pi^*_h u_h)^-\mathrm{d}s+\int_{(\partial K^*_{P_j})_+}(b\cdot\nu)(\Pi^*_h u_h)^+\mathrm{d}s.\end{aligned} \tag{7.4.4}$$

Next we replace $\Pi^*_h u_h$ in the second term on the right-hand side of (7.4.3) by $(\Pi^*_h u_h)^+$, and substitute (7.4.4) into the first term to obtain

$$\int_{K^*_{P_j}} b\cdot\nabla(\Pi^*_h u_h)v_h\mathrm{d}x\approx\int_{(\partial K^*_{P_j})_-}(b\cdot\nu)[\Pi^*_h u_h]v_h\mathrm{d}s, \tag{7.4.5}$$

where

$$[\Pi^*_h u_h]=(\Pi^*_h u_h)^+-(\Pi^*_h u_h)^-.$$

Finally, inserting (7.4.2) and (7.4.5) into (7.4.1) yields a semi-discrete generalized upwind scheme:

$$\begin{aligned}&\int_{K^*_{P_j}}\frac{\partial u_h}{\partial t}v_h\mathrm{d}x\\ =\ & \mu\int_{\partial K^*_{P_j}}\frac{\partial u_h}{\partial\nu}v_h\mathrm{d}s-\int_{(\partial K^*_{P_j})_-}(b\cdot\nu)[\Pi^*_h u_h]v_h\mathrm{d}s+\int_{K^*_{P_j}} f v_h\mathrm{d}x.\end{aligned} \tag{7.4.6}$$

Let us introduce the following bilinear forms:

$$a(u,v_h)=-\sum_{j=1}^{N}v_h(P_j)\int_{\partial K^*_j}\mu\frac{\partial u}{\partial\nu}\mathrm{d}s, \tag{7.4.7}$$

$$\begin{aligned} b(u,v_h) &= \sum_{j=1}^{N}\int_{K^*_{P_j}}(b\cdot\nabla u)v_h\mathrm{d}x \\ &= \sum_{j=1}^{N}v_h(P_j)\int_{\partial K^*_{P_j}}b\cdot\nu u\mathrm{d}s - \int_\Omega uv_h\mathrm{div}b\mathrm{d}x, \end{aligned} \tag{7.4.8a}$$

$$b_h(u_h,v_h) = \sum_{j=1}^{N}\int_{(\partial K^*_{P_j})_-}(b\cdot\nu)[\Pi^*_h u_h]v_h\mathrm{d}s. \tag{7.4.8b}$$

Then, the solution $u$ to (7.3.1) satisfies

$$\left(\frac{\partial u}{\partial t},v_h\right)+a(u,v_h)+b(u,v_h)=(f,v_h),\quad \forall v_h\in V_h. \tag{7.4.9}$$

(7.4.6) can be written as

$$\left(\frac{\partial u_h}{\partial t},v_h\right)+a(u_h,v_h)+b_h(u_h,v_h)=(f,v_h),\quad \forall v_h\in V_h. \tag{7.4.6$'$}$$

**Remark 1** If we adopt a higher order finite element space $U_h$, and a corresponding polynomial function space $V_h$ on $T^*_h$ (cf. §3.4 and §3.5), then $\nabla(\Pi^*_h u_h)\neq 0$, and (7.4.8b) should be modified as

$$\begin{aligned} & b_h(u_h,v_h) \\ = & \sum_{j=1}^{N}\int_{K^*_{P_j}}b\cdot\nabla(\Pi^*_h u_h)v_h\mathrm{d}x+\sum_{j=1}^{N}\int_{(\partial K^*_{P_j})_-}(b\cdot\nu)[\Pi^*_h u_h]v_h\mathrm{d}s. \end{aligned} \tag{7.4.8c}$$

Take a time step size $\tau=T/N$ ($N$ is a positive integer), and the nodes $t_k=k\tau$ ($k=0,1,\cdots,N$). Use $u^k_h$ for the approximation of $u(x,k\tau)$ and introduce the symbols $u^{k,\theta}_h=\theta u^k_h+(1-\theta)u^{k-1}_h$ $(0\le\theta\le 1)$, then a class of fully-discrete generalized upwind schemes approximating (7.3.1) is:

$$(\bar\partial_t u^k_h,v_h)+a(u^{k,\theta}_h,v_h)+b_h(u^{k,\theta}_h,v_h)=(f,v_h),\quad \forall v_h\in V_h. \tag{7.4.10}$$

This leads to an explicit forward and an implicit backward scheme as $\theta=0,1$ respectively:

$$(\bar\partial_t u^k_h,v_h)+a(u^{k-1}_h,v_h)+b_h(u^{k-1}_h,v_h)=(f,v_h), \tag{7.4.11}$$

$$(\bar{\partial}_t u_h^k, v_h) + a(u_h^k, v_h) + b_h(u_h^k, v_h) = (f, v_h). \qquad (7.4.12)$$

**Remark 2** It should be pointed out that the schemes (7.4.6) and (7.4.10) are slightly different from (7.3.8) and (7.3.12). The techniques in the previous section result in only first order accuracy schemes, while schemes with arbitrary orders can be deduced by the methods in this section by choosing $U_h$ as higher order element spaces.

**Remark 3** As in §6.3, one may extend the methods in this section to systems of convection-diffusion equations by introducing a viscosity term (cf. [B-60]).

### 7.4.2 Convergence and error estimate

As in §7.3, we once again assume that $-\frac{1}{2}\mathrm{div}b \geq \sigma_0 > 0$ and that $T_h^*$ is a barycenter dual grid. By §3.2, there exists a constant $\alpha_0 > 0$ such that

$$a(\bar{u}_h, \Pi_h^* \bar{u}_h) \geq \alpha_0 \|\bar{u}_h\|_1^2, \quad \forall \bar{u}_h \in U_h. \qquad (7.4.13)$$

Noting (6.2.15), we have a constant $\gamma_0 > 0$ such that

$$b_h(v_h, v_h) \geq \gamma_0(\|v_h\|_0^2 + \|v_h\|_{\partial\Omega}^2). \qquad (7.4.14)$$

First we need to evaluate the difference of $b(u, v_h)$ and $b_h(u, v_h)$. To this end, we observe that according to the approximation procedure of the convection term

$$\begin{aligned} b_h(u, v_h) = \; & -\sum_{j=1}^{N} \int_{K_{P_j}^*} \Pi_h^* u \mathrm{div}(b v_h) \mathrm{d}x \\ & + \sum_{j=1}^{N} \int_{\partial K_{P_j}^*} b \cdot \nu (\Pi_h^* u)^+ v_h \mathrm{d}s, \end{aligned}$$

where $(\Pi_h^* u)^+$ stands for the upwind value across the boundary $\partial K_{P_j}^*$. Subtract it from (7.4.8a) to get

$$\begin{aligned} & b(u, v_h) - b_h(u, v_h) \\ = \; & -\sum_{j=1}^{N} \int_{K_{P_j}^*} (u - \Pi_h^* u) v_h(P_j) \mathrm{div} b \mathrm{d}x \end{aligned}$$

$$\begin{aligned}
&+\sum_{j=1}^{N}\Big[\int_{(\partial K^*_{P_j})_+} b\cdot\nu(u-u(P_j))v_h(P_j)\mathrm{d}s \\
&+\int_{(\partial K^*_{P_j})_-} b\cdot\nu(u-(\Pi_h^*u)^+)v_h(P_j)\mathrm{d}s\Big] \qquad (7.4.15)\\
=\ & I_{1h}+I_{2h},
\end{aligned}$$

where

$$\begin{aligned}
|I_{1h}| &= \Big|\sum_{j=1}^{N}\int_{K^*_{P_j}}(u-\Pi_h^*u)v_h(P_j)\mathrm{div}b\mathrm{d}x\Big| \\
&= \Big|\int_\Omega (u-\Pi_h^*u)v_h\mathrm{div}b\mathrm{d}x\Big| \qquad (7.4.16a)\\
&\le C\|\mathrm{div}b\|_\infty h|u|_1\|v_h\|_0,
\end{aligned}$$

$$\begin{aligned}
I_{2h} = & \sum_{j=1}^{N}\sum_{l\in\Lambda_j}\int_{\Gamma_{jl}} b\cdot\nu[H(\beta_{jl})(u-u(P_j)) \qquad (7.4.16b)\\
&+(1-H(\beta_{jl}))(u-u(P_l))]v_h(P_j)\mathrm{d}s.
\end{aligned}$$

The above integral line $\Gamma_{jl}$'s are the sides of $K^*_{P_j}$. If $l\in\Lambda_{jl}$, then along $\Gamma_{lj}$ we have

$$\begin{aligned}
&\int_{\Gamma_{lj}} b\cdot\nu[H(\beta_{lj})(u-u(P_l)) \\
&+(1-H(\beta_{lj}))(u-u(P_j))]v_h(P_l)\mathrm{d}s \\
= & -\int_{\Gamma_{jl}} b\cdot\nu[H(\beta_{jl})(u-u(P_j)) \\
&+(1-H(\beta_{jl}))(u-u(P_l))]v_h(P_j)\mathrm{d}s.
\end{aligned}$$

So we have

$$\begin{aligned}
I_{2h} = & \frac{1}{2}\sum_{j=1}^{N}\sum_{l\in\Lambda_j}(v_h(P_j)-v_h(P_l))\int_{\Gamma_{jl}} b\cdot\nu \\
&\cdot[H(\beta_{jl})(u-u(P_j))+(1-H(\beta_{jl}))(u-u(P_l))]\mathrm{d}s.
\end{aligned}$$

Similar to (7.2.27) one has

$$|I_{2h}| \leq C\|b\|_\infty h|u|_2|\bar{u}_h|_1, \ v_h = \Pi_h^* \bar{u}_h, \ \bar{u}_h \in U_h. \tag{7.4.17}$$

Summarizing, we have a constant $C > 0$ such that

$$|b(u, \Pi_h^* \bar{u}_h) - b_h(u, \Pi_h^* \bar{u}_h)| \leq Ch\|u\|_2\|\bar{u}\|_1. \tag{7.4.18}$$

After the above preparations, we readily obtain an estimate like (7.3.28) for the fully-discrete backward difference solution:

$$\max_{1\leq k\leq N} \|u(t_k) - u_h^k\|_0 \leq Ch\|u\|_{X_1}. \tag{7.4.19}$$

**Remark 4** If $T_h$ is an acute triangulation, then the solution of the backward upwind scheme enjoys an extreme value property as well as a uniform convergence.

## 7.5 Upwind Schemes for Nonlinear Convection Problems

The above introduced generalized upwind schemes can be extended to elliptic and parabolic differential equations with a nonlinear convection term. The key point is to employ Osher's split technique of nonlinear functions. Retaining the previous symbols, we consider a nonlinear elliptic equation:

$$\begin{cases} -\mu\Delta u + \nabla \cdot F(x, u) = g(x, u), & x \in \Omega \subset \mathrm{R}^2, \quad (7.5.1a) \\ u = 0, & x \in \Gamma = \partial\Omega, \quad (7.5.1b) \end{cases}$$

where $F(x, u) = (f_1(x, u), f_2(x, u))$ is a smooth function on $\bar{\Omega}\times\mathrm{R}$, satisfying

$$F(x, 0) = 0. \tag{7.5.2}$$

Multiply (7.5.1) by $v$, integrate it on a dual element $K_{P_j}^*$, apply Green's formula, and sum it over $j = 1, 2, \cdots, N$, then we have

$$a(u, v) + b(u, v) = (g, v), \tag{7.5.3}$$

where $u$ and $v$ satisfy (7.5.1b), and

$$a(u,v) = \sum_{j=1}^{N}\Big[\mu \int_{K^*_{P_j}} \nabla u \cdot \nabla v \mathrm{d}x - \mu \int_{\partial K^*_{P_j}} \frac{\partial u}{\partial \nu} v \mathrm{d}s\Big], \tag{7.5.4}$$

$$b(u,v) = -\sum_{j=1}^{N} \int_{K^*_{P_j}} F(x,u)\nabla v \mathrm{d}x + \sum_{j=1}^{N} \int_{\partial K^*_{P_j}} F(x,u)\nu v \mathrm{d}s. \tag{7.5.5}$$

Write $F$ as

$$F(x,u) = \int_0^u \frac{\partial F(x,\bar{u})}{\partial \bar{u}} \mathrm{d}\bar{u}. \tag{7.5.6}$$

Denote by $P_{jl}$ the midpoint of two adjacent nodes $P_j$ and $P_l$, and set

$$\beta^+_{jl}(u) = \int_0^u \max\Big(0, \frac{\partial F(P_{jl},\bar{u})}{\partial \bar{u}} \cdot \nu_{jl}\Big) \mathrm{d}\bar{u}, \tag{7.5.7}$$

$$\beta^-_{jl}(u) = \int_0^u \max\Big(0, -\frac{\partial F(P_{jl},\bar{u})}{\partial \bar{u}} \cdot \nu_{jl}\Big) \mathrm{d}\bar{u}, \tag{7.5.8}$$

where $\nu_{jl}$ is the unit outer normal direction of $\Gamma_{jl} \subset \partial K^*_{P_j}$. For $u_h \in U_h$ and $v_h \in V_h$, we introduce a bilinear form

$$b_h(u_h,v_h) = \sum_{j=1}^{N} v_h(P_j) \sum_{l\in\Lambda_j} \gamma_{jl}[\beta^+_{jl}(u_h(P_j)) - \beta^-_{jl}(u_h(P_l))], \tag{7.5.9}$$

where $\gamma_{jl}$ is the length of $\Gamma_{jl}$; $U_h$ is the piecewise linear element space on $T_h$ satisfying $U_h \subset H^1_0(\Omega)$; and $V_h$ is the piecewise constant function space on $T^*_h$, subject to the zero boundary condition on $\partial\Omega$. Then, a generalized upwind difference scheme approximating (7.5.1) is: Find $u_h \in U_h$ such that

$$a(u_h,v_h) + b_h(u_h,v_h) = (g(x,u_h),v_h), \quad \forall v_h \in V_h. \tag{7.5.10}$$

This is apparently a generalization of Scheme (7.2.13). For a discussion of the monotonicity and convergence of (7.5.10), we refer to [B-85].

For a non−steady-state diffusion equation with a nonlinear convection term:

$$\frac{\partial u}{\partial t} - \mu\Delta u + \nabla \cdot F(x,u) = g(x,u), \tag{7.5.11}$$

we have the following upwind difference scheme:

$$(\bar{\partial}_t u_h^k, v_h) + a(u_h^{k,\theta}, v_h) + b_h(u_h^{k,\theta}, v_h) = (g(x, u_h^{k,\theta}), v_h), \qquad (7.5.12)$$

where $u_h^{k,\theta} = \theta u_h^k + (1-\theta)u_h^{k-1}$, $0 \le \theta \le 1$. (7.5.12) stands for a forward explicit scheme when $\theta = 0$, and a backward implicit scheme when $\theta = 1$.

In order to construct highly accurate upwind schemes, we rewrite (7.5.11) as

$$\frac{\partial u}{\partial t} - \mu\Delta u + \bar{b}(x,u)\cdot\nabla u = \bar{g}(x,u), \qquad (7.5.13)$$

where

$$\bar{b}(x,u) = \Big(\frac{\partial f_1}{\partial u}, \frac{\partial f_2}{\partial u}\Big),$$

$$\bar{g}(x,u) = g(x,u) - \Big(\frac{\partial f_1}{\partial x_1} + \frac{\partial f_2}{\partial x_2}\Big).$$

Define $a(u,v)$ by (7.5.4), and

$$b(u,v) = -\sum_{j=1}^{N}\int_{K_{P_j}^*} u\cdot\nabla(v\bar{b})\mathrm{d}x + \sum_{j=1}^{N}\int_{\partial K_{P_j}^*} (\bar{b}\cdot\nu)uv\mathrm{d}s.$$

Then we can write (7.5.13) in a weak form: Find $u(x,t) \in C([0,T]; H_0^1(\Omega))$ such that

$$\Big(\frac{\partial u}{\partial t}, v\Big) + a(u,v) + b(u,v) = (\bar{g}, v), \quad \forall v \in H_0^1(\Omega). \qquad (7.5.14)$$

In order to construct a upwind scheme, we first discretize the time direction to get

$$(\bar{\partial}_t u^{k-1}, v) + a(u^{k-1}, v) + b(u^{k-1}, v) = (\bar{g}(x, u^{k-1}), v).$$

Then set (cf (7.4.8b))

$$b_h(u_h^{k-1}, v_h) = \sum_{j=1}^{N}\int_{(\partial K_{P_j}^*)_-} \bar{b}(x, u_h^{k-1})\cdot\nu[\Pi_h^* u_h^{k-1}]v_h\mathrm{d}s. \qquad (7.5.15)$$

So our task is to seek $u_h^k \in U_h$ such that

$$(\bar{\partial}_t u_h^k, v_h) + a(u_h^{k-1}, v_h) + b_h(u_h^{k-1}, v_h) = (\bar{g}(x, u_h^{k-1}), v_h),$$

$$\forall v_h \in V_h. \tag{7.5.16}$$

The stability and the convergence of schemes (7.5.12) and (7.5.16) have not been studied yet.

## Bibliography and Comments

The paper [B-21] of Courant et. al. is the most fundamental work on the numerical solution of hyperbolic equations. [B26] combines characteristic methods with finite element, or finite difference, methods, and constructs a kind of upwind scheme on rectangular networks for the convection-dominated diffusion equations. Since 1977, Tabata and others have published a series of papers studying upwind schemes on triangular networks (cf. [B-3,83,84,85] and the references therein). They employ linear finite elements to discretize the diffusion term, and upwind difference schemes to discretize the convection term. Dong Liang ([A-17,18]) uses linear generalized difference method to deal with the diffusion term, and upwind schemes to convection term. Besides the methods discussed in the second section of this chapter, Dong Liang also proposes another class of upwind schemes based on some monotonic schemes, which is similar to some methods appearing in mechanics literature (cf. [B-80]). A class of highly accurate upwind schemes is obtained in [A-28] and [B-61] by approximating the diffusion term by higher order element generalized difference methods, and the convection term by highly accurate upwind schemes. We remark that the methods resulted from the linear case of this class of schemes are not identical to those in §7.2 and §7.3. [B-85] is among the few papers discussing nonlinear convection terms. Finally, we observe that if the diffusion coefficient $\mu = 0$, then the methods in this chapter result in the difference methods for hyperbolic equations.

**Problem 1** Extend the results to a system of convection- dominated diffusion equations on higher dimensions, e.g., on two-dimensional regions.

**Problem 2** Extend the results to higher dimensional, nonlinear, convection-dominated problems.

# Chapter 8

# APPLICATIONS

The first six sections of this chapter are devoted to the applications of generalized difference methods to elastic mechanics, fluid kinetics, electromagnetic fields, coupled sound-heat problems and long wave equations. The last section discusses the hierarchical basis methods for difference equations.

## 8.1 Planar Elastic Problems

Under certain conditions, one can regard the study of an elastic body, at an equilibrium state subject to an outer force, as a planar elastic problem. Let $\Omega$ be a planar region occupied by the elastic body, and $\Gamma = \partial\Omega$ its boundary. There are three groups of state variables: the stress tensor $\sigma = (\sigma_{11}, \sigma_{22}, \sigma_{12})^T$, the strain tensor $\epsilon = (\epsilon_{11}, \epsilon_{22}, \epsilon_{12})^T$, and the displacement tensor $u = (u_1, u_2)^T$. Assume the elastic body is homogeneous and isotropic. Write $\nabla = (\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2})$, denote by $\lambda, \mu > 0$ the Lame constants, and set

$$E(\nabla) = \begin{pmatrix} \partial/\partial x_1 & 0 & \partial/\partial x_2 \\ 0 & \partial/\partial x_2 & \partial/\partial x_1 \end{pmatrix},$$

$$A = \begin{bmatrix} \lambda + 2\mu & \lambda & 0 \\ \lambda & \lambda + 2\mu & 0 \\ 0 & 0 & \mu \end{bmatrix},$$

then $\sigma, \epsilon$, and $u$ satisfy the following three equations (cf. [A-26,16]):

$$\begin{cases} \epsilon = E^T(\nabla)u, & \text{(strain − displacement relation)} \quad (8.1.1) \\ E(\nabla)\sigma + f = 0, & \text{(balance equation)} \quad (8.1.2) \\ \sigma = A\epsilon, & \text{(stress − strain relation)} \quad (8.1.3) \end{cases}$$

where $f$ is the body force.

It is an easy matter to deduce, from Green's formula, the following general Green's formula:

$$\int_\Omega \sigma^T E^T(\nabla)u dx + \int_\Omega (E(\nabla)\sigma)^T u dx = \int_{\partial\Omega} (E(\nu)\sigma)^T u ds, \quad (8.1.4)$$

where $\nu = (\nu_1, \nu_2)^T$ is the unit outer normal vector of $\Gamma$.

Suppose $\Gamma$ is divided into two parts $\Gamma_0$ and $\Gamma_1$. A displacement boundary condition $u = \bar{u}$ is given on $\Gamma_0$, and a surface force condition $E(\nu)\sigma = \bar{P}$ on $\Gamma_1$.

To solve the above system in practice, one usually eliminates some variables in (8.1.1)-(8.1.3), then solves it for the remaining unknowns, yielding accordingly the displacement method, the force method or the mixed method. In the sequel, we describe the generalized difference methods based on the displacement and the mixed methods.

### 8.1.1 Displacement methods

Eliminating $\sigma$ and $\epsilon$ in (8.1.1)-(8.1.3) yields a system of second order elliptic partial differential equations of the displacement $u$:

$$-\mu\nabla u - (\lambda + \mu)\text{grad div } u = f, \quad (8.1.5)$$

where $\text{div}u = \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2}$. Multiply (8.1.5) by $v \in (H_E^1(\Omega))^2$, integrate it over $x \in \Omega$, and make use of Green's formula, then we have an equation in an integral form:

$$a(u, v) - \int_{\partial\Omega} \Big(\mu\frac{\partial u}{\partial \nu} + (\lambda + \mu)(\text{div}u)\nu\Big)v ds = (f, v), \quad (8.1.6)$$

where

$$a(u, v) = \int_\Omega \Big[\mu\nabla u\nabla v + (\lambda + \mu)\text{div}u \cdot \text{div}v\Big] dx. \quad (8.1.7)$$

On the boundary $\Gamma_1$

$$\mu\frac{\partial u}{\partial\nu} + (\lambda+\mu)(\text{div}u)\nu = (E(\nu)\sigma)^T = \bar{P}.$$

Thus, we obtain a variational form of (8.1.5): Find $u \in (H^1(\Omega))^2$, $u|_{\Gamma_0} = u_0$ such that

$$a(u,v) = (f,v) + \int_{\Gamma_1} \bar{P}v\mathrm{d}s, \quad \forall v \in (H_E^1(\Omega))^2. \tag{8.1.8}$$

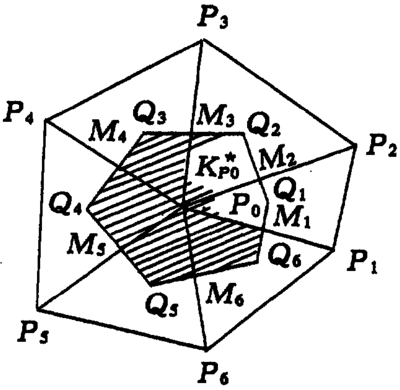



Fig. 8.1.1

To construct a generalized difference scheme, as before let $T_h = \{K\}$ be a triangulation of $\Omega$ such that $\Omega_h = \bigcup_{K\in T_h} K$ is an approximation of $\Omega$. Let $T_h^* = \{K_P^*\}$ be a dual grid of $T_h$, usually a barycenter or a circumcenter dual grid. In the sequel, we assume $T_h^*$ is a circumcenter dual grid. Fig. 8.1.1 shows all the triangular elements in $T_h$ with vertex $P_0$, as well as the dual element surrounding $P_0$. Let $U_h$ be a piecewise linear vector function space related to $T_h$. The interior nodes of $T_h$ are numbered by $1, 2, \cdots, N_0$. The boundary nodes are divided into two groups: The nodes where force conditions are given are numbered by $N_0+1, \cdots, N_1$, and the nodes bearing displacement conditions by $N_1+1, \cdots, N$. Denote by a scalar function $\phi_i(x)$ the basis function of the node $i \in \{1, 2, \cdots, N_1\}$, then a function $u_h \in U_h$

satisfying $u|_{\Gamma_{0h}} = 0$ can be expressed as

$$u_h(x) = \sum_{i=1}^{N_1} u_i \phi_i(x),$$

where $u_i$ is the value of $u_h(x)$ at the $i$-th node $x_i$, and $\Gamma_{0h}$ is an approximation of $\Gamma_0$. Choose $V_h$ as the piecewise constant vector function space corresponding to $T_h^*$, subject to the boundary condition that $v_h \in V_h$ vanishes on the dual elements corresponding to the nodes $i \in \{N_1+1, \cdots, N\}$. Let $\bar{\psi}_j = (\psi_j, \psi_j)^T$ be the dual basis function. The generalized difference equation reads: Find $u_h \in U_h$ such that $u_h = u_0$ on $\Gamma_{0h}$ and

$$a(u_h, \bar{\psi}_j) = (f, \bar{\psi}_j) + \int_{\Gamma_{1h}} \bar{P}\bar{\psi}_j \mathrm{d}s, \; j = 1, 2, \cdots, N_1, \tag{8.1.9}$$

where $\Gamma_{1h}$ is a certain approximation of $\Gamma_1$.

In fact, we can as well work out (8.1.9) in a more direct manner. So we integrate the two sides of (8.1.5), apply Green's formula, and replace $u$ by $u_h$ to obtain

$$-\int_{\partial K_{P_0}^*} \left[\mu \frac{\partial u_h}{\partial \nu} + (\lambda+\mu)(\mathrm{div} u_h)\nu\right] \mathrm{d}s = \int_{K_{P_0}^*} f \mathrm{d}x. \tag{8.1.10}$$

This is the generalized difference equation at the node $P_0$. Let $P_0$ and its adjacent nodes be as in Fig. 8.1.1. We now compute the integrals on the left-hand side of (8.1.10). We divide the first integral into a sum of integrals on the perpendicular bisector segments $\overline{Q_1Q_2}, \overline{Q_2Q_3}, \cdots, \overline{Q_6Q_1}$. For instance, the integral on $\overline{Q_1Q_2}$ looks like

$$-\mu \int_{\overline{Q_1Q_2}} \frac{\partial u_h}{\partial \nu} \mathrm{d}s = -\mu|\overline{Q_1Q_2}|(u_{P_2} - u_{P_0})/|\overline{P_0P_2}|. \tag{8.1.11}$$

Similarly, the second integral of (8.1.10) is divided into a sum of integrals on the fold line segments $\overline{M_1Q_1M_2}, \overline{M_2Q_2M_3}, \cdots, \overline{M_6Q_6M_1}$. For example, on $\overline{M_1Q_1M_2}$,

$$\begin{aligned} & -(\lambda+\mu)\int_{\overline{M_1Q_1M_2}} (\mathrm{div} u_h)\nu \mathrm{d}s \\ = \quad & -(\lambda+\mu)\mathrm{div} u_h(Q_1)\Big(\frac{\overline{P_0P_1}}{|\overline{P_0P_1}|}|\overline{M_1Q_1}| + \frac{\overline{P_0P_2}}{|\overline{P_0P_2}|}|\overline{Q_1M_2}|\Big), \end{aligned} \tag{8.1.12}$$

where

$$\begin{aligned}
&\mathrm{div} u_h(Q_1) = \frac{\partial u_{1h}}{\partial x_1}(Q_1) + \frac{\partial u_{2h}}{\partial x_2}(Q_1) \\
= \; & \frac{1}{2S_{Q_1}}[(x_2(P_1) - x_2(P_2))u_1(P_0) + (x_2(P_2) - x_2(P_0))u_1(P_1) \\
& +(x_2(P_0) - x_2(P_1))u_1(P_2)] + \frac{1}{2S_Q}[(x_1(P_2) - x_1(P_1))u_2(P_0) \\
& +(x_1(P_0) - x_1(P_2))u_2(P_1) + (x_1(P_1) - x_1(P_0))u_2(P_2)].
\end{aligned} \tag{8.1.13}$$

Here $(x_1(P_i), x_2(P_i))$ are the coordinates of the node $P_i$, and $S_{Q_1}$ is the area of the triangular element containing the circumcenter $Q_1$.

Equation (8.1.10) is a generalized difference equation for an interior node. On the boundary $\Gamma_{0h}$ the displacement $u_0$ is given. Extra equations are needed for the nodes on $\Gamma_{1h}$. As an example, suppose a node $\bar{P}_0 \in \Gamma_{1h}$ and its neighbouring nodes are as in Fig. 8.1.2. In such a case, we still have equation (8.1.10), and the line integrals on the left-hand side are computed by use of the formulas (8.1.11) and (8.1.12).

So we finally end up with a system of $(2N_1)$ equations like (8.1.10) together with the displacement boundary condition on $\Gamma_{0h}$.

Compared with the (piecewise linear) finite element method, our generalized difference method here enjoys the same convergence order, and less computational work. One can also employ high order generalized difference methods to approximate planar elastic problems. In practical computation, it might be more convenient to use a barycenter dual grid instead, since it is more accurate and can be extended directly to three-, or arbitrary $n$-, dimensional problems.

### 8.1.2 Mixed methods

Work out $\epsilon = B\sigma$ ($B = A^{-1}$) from (8.1.3), insert it into (8.1.1), and couple it with (8.1.2), then we have the following system of equations of $\sigma$ and $u$:

$$B\sigma - E^T(\nabla)u = 0, \tag{8.1.14}$$

$$E(\nabla)\sigma = f, \tag{8.1.15}$$

where we have replaced $f$ by $-f$, and

$$B = \frac{1}{4\mu(\lambda+\mu)} \begin{bmatrix} \lambda+2\mu & -\lambda & 0 \\ -\lambda & \lambda+2\mu & 0 \\ 0 & 0 & 4(\lambda+\mu) \end{bmatrix}. \tag{8.1.16}$$

In this case the displacement condition remains to be an essential boundary condition, while the force condition is a natural boundary condition. The space $U_h$ of the approximate displacement and the corresponding test function space $V_h$ are constructed as before, with nodal basis functions $\phi_i(x)$ and $\bar{\psi}_i(x)$ respectively.

The three-dimensional tensor $\sigma_k$'s belong to a piecewise linear (vector) function space $M_h$ related to $T_h$, of which the vertexes of $T_h$ are the interpolation nodes and the nodal basis function is $\xi_i(x)$. The corresponding test function space $N_h$ is the piecewise constant space with respect to $T_h^*$, with nodal basis function $\bar{\eta}_i(x)$. Define (cf. [A-26])

$$a(\sigma \times u, \tau \times v) = \int_\Omega \left\{ \tau^T B\sigma - \tau^T E^T u - \sigma^T E^T(\nabla)v \right\} \mathrm{d}x.$$

Then, the generalized difference method for the equations (8.1.14) and (8.1.15) is: Find $u_h \in U_h$, $u_h|_{\Gamma_{0h}} = u_0$ and $\sigma_h \in M_h$ such that

$$a(\sigma_h \times u_h, \bar{\eta}_j \times \bar{\psi}_j) = (f, \bar{\psi}_j) + \int_{\Gamma_{1h}} \bar{P}\bar{\psi}_j \mathrm{d}s, \; j = 1, 2, \cdots, N_1. \tag{8.1.17}$$

Once again, we have a more direct way to deduce the generalized difference equation. Integrate (8.1.14) and (8.1.15) respectively on a dual element $K_{P_0}^*$ and apply the following Green's formulas:

$$\int_{K_{P_0}^*} E^T(\nabla)u_h \mathrm{d}x = \int_{\partial K_{P_0}^*} (E(\nu))^T u_h \mathrm{d}s,$$

$$\int_{K_{P_0}^*} (E(\nabla)\sigma_h)^T \mathrm{d}x = \int_{\partial K_{P_0}^*} (E(\nu)\sigma_h)^T \mathrm{d}s,$$

then we have

$$\int_{K_{P_0}^*} B\sigma_h \mathrm{d}x - \int_{\partial K_{P_0}^*} (E(\nu))^T u_h \mathrm{d}s = 0, \tag{8.1.14$'$}$$

$$\int_{\partial K_{P_0}^*} (E(\nu)\sigma_h)^T \mathrm{d}s = \int_{K_{P_0}^*} f \mathrm{d}x. \qquad (8.1.15)'$$

Suppose $P_0$ and its neighbouring nodes are as in Fig. 8.1.1. Then the double integral on $K_{P_0}^*$ can be divided into a sum of integrals on the intersections of $K_{P_0}^*$ and the adjacent elements respectively. For example, we have

$$\begin{aligned}
&\int_{\Delta P_0 P_1 P_2 \cap K_{P_0}^*} B\sigma_h \mathrm{d}x \\
= \;& \frac{1}{3} B\{(S_{P_0 M_1 Q_1} + S_{P_0 Q_1 M_2})(\sigma_h(P_0) + \sigma_h(Q_1)) \\
& + S_{P_0 M_1 Q_1}\sigma_h(M_1) + S_{P_0 Q_1 M_2}\sigma_h(M_2)\},
\end{aligned}$$

where $S_{ABC}$ denotes the area of a triangle $\triangle ABC$. Similarly, the line integral on $\partial K_{P_0}^*$ can be divided into a sum of the line integrals on the perpendicular bisectors. For instance,

$$\begin{aligned}
&\int_{\overline{Q_1 Q_2}} E(\nu)^T u_h \mathrm{d}s \\
= \;& (E(\nu))^T (|\overline{Q_1 M_2}| u_h(Q_1) + |\overline{Q_1 Q_2}| u_h(M_2) + |\overline{M_2 Q_2}| u_h(Q_2))/2, \\
& \nu = \overline{P_0 P_2}/|\overline{P_0 P_2}|.
\end{aligned}$$

So we end up with a system of $(5N_1)$ generalized difference equations of the forms (8.1.14)′ and (8.1.15)′, plus the displacement boundary conditions on $\Gamma_{0h}$. Like the finite element method, the generalized difference method based on the mixed variational form can obtain both the displacement and the strain simultaneously.

## 8.2 Computation of Electromagnetic Fields

In 1967, A. M. Winslow [B-99] applied a difference method on irregular networks to a two-dimensional quasi-linear Poisson equation representing an electromagnetic field. He allows the network to be a planar curvilinear network, but each node can be the vertex of at most six triangular elements. In 1990, G. Zhao and Y. Liu [A-60] extended Winslow's method to three dimensions, and carried out a numerical experiment for a tetrapolar lens. Their numerical result matches well the theoretical prediction. They adopt cylindrical coordinates. Then they cut the region, for different angles $\theta$, to get some $(r, z)$ planes $D_l$'s, on which is placed a Winslow triangulation with curvilineal sides. Special cones are constructed between different $D_l$'s, referred to as a "secondary network" (corresponding to the dual grid). This paper claims that this method has a bright future in the computation of photoelectronics. We describe in this section in a united framework the generalized difference method for three-dimensional Poisson equations. We only present the results for the Poisson equation in Cartesian coordinates, which can be directly extended to cylindrical or spheroidal coordinates, and even to the second order elliptic equations with variable coefficients. As in the planar case, one can similarly establish the convergence and the error estimate, which are omitted here.

Let $\Omega \subset \mathrm{R}^3$ be a polyhedral region, and $T_h = \{K\}$ a tetrahedral grid of $\Omega$ such that different tetrahedral elements share no common interior and $\Omega = \bigcup_{K \in T_h} K$. As in the case of planar triangular elements, we can analogously introduce a tetrahedral volume coordinates. Let the vertexes of a tetrahedron $K$ be $P_i$ $(i = 1, 2, 3, 4)$. Then for any point $P \in K$ the volume coordinates $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ are defined as

$$\lambda_i = \frac{V_i}{V}, \ i = 1, 2, 3, 4,$$

where $V$ is the volume of $K$, and $V_i$ is the volume of the tetrahedron formed by the point $P$ and the base triangle facing $P_i$ (cf. Fig. 8.2.1). Apparently $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$.

If the Cartesian coordinates of $P_i$ $(i = 1, 2, 3, 4)$ are $(x_i, y_i, z_i)$, and the volume coordinates of $P$ are $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$, then the Cartesian coordinates $(x, y, z)$ of $P$ can be expressed by the volume coordinates as

$$\begin{cases} x = \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \lambda_4 x_4, \\ y = \lambda_1 y_1 + \lambda_2 y_2 + \lambda_3 y_3 + \lambda_4 y_4, \\ z = \lambda_1 z_1 + \lambda_2 z_2 + \lambda_3 z_3 + \lambda_4 z_4, \\ 1 = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4. \end{cases} \tag{8.2.1}$$

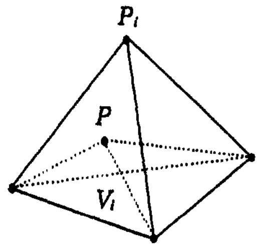



Fig. 8.2.1

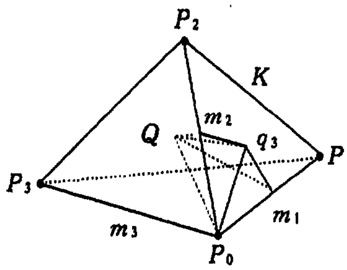



Fig. 8.2.2

On the other hand, the volume coordinates can be expressed by the Cartesian coordinates as

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} = \frac{1}{6V} \begin{bmatrix} X_{14} & Y_{14} & Z_{14} \\ X_{24} & Y_{24} & Z_{24} \\ X_{34} & Y_{34} & Z_{34} \end{bmatrix} \begin{bmatrix} x - x_4 \\ y - y_4 \\ z - z_4 \end{bmatrix}, \qquad (8.2.2)$$

where

$$X_{i4} = \begin{vmatrix} y_{j4} & y_{k4} \\ z_{j4} & z_{k4} \end{vmatrix}, \; Y_{i4} = \begin{vmatrix} z_{j4} & z_{k4} \\ x_{j4} & x_{k4} \end{vmatrix}, \; Z_{i4} = \begin{vmatrix} x_{j4} & x_{k4} \\ y_{j4} & y_{k4} \end{vmatrix},$$

$$x_{j4} = x_j - x_4, \; y_{j4} = y_j - y_4, \; z_{j4} = z_j - z_4.$$

$$(j = i + 1, \; k = j + 1, \; i = k + 1.)$$

$V = |\bar{V}|$ is the area of the element, given by

$$\bar{V} = \frac{1}{6} \begin{vmatrix} x_{14} & x_{24} & x_{34} \\ y_{14} & y_{24} & y_{34} \\ z_{14} & z_{24} & z_{34} \end{vmatrix} = -\frac{1}{6} \begin{vmatrix} 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \\ z_1 & z_2 & z_3 & z_4 \end{vmatrix}. \qquad (8.2.3)$$

Now we define the dual grid $T_h^*$. Let $P_0$ be a node (cf. Fig. 8.2.2). Consider all the tetrahedrons with $P_0$ as a vertex. $K$ is one of them, depicted in Fig. 8.2.2, with vertexes $P_0, P_1, P_2, P_3$. $K$ has three base triangles with the vertex $P_0$, one of which is $\Delta P_0 P_1 P_2$. Denote the midpoints of $\overline{P_0P_1}$ and $\overline{P_0P_2}$ by $m_1$ and $m_2$ respectively, and the barycenter of $\Delta P_0P_1P_2$ by $q_3$. $Q$ represents the barycenter of $K$, with volume coordinates $\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$. Take $P_0$ as a vertex, and $\Delta Qq_3m_1$ and $\Delta Qq_3m_2$ as bases respectively to form two cones. For each of the other two triangles $\Delta P_0P_2P_3$ and $\Delta P_0P_1P_3$, we similarly define two cones. The union of all these six cones constitutes the intersection of the element $K$ and the dual element $K^*_{P_0}$ surrounding $P_0$. By taking $P_0$ to be every (inner and boundary) node, we end up with a (barycenter) dual grid $T^*_h$ related to $T_h$.

The trial function space $U_h$ corresponding to the grid $T_h$ is chosen as the usual finite element space such that $u_h \in U_h$ is a piecewise polynomial of degree $k$ on $T_h = \{K\}$, possessing a global smoothness of a certain degree. Here we restrict ourselves to discuss only the piecewise linear function space with the vertexes of the tetrahedrons as the nodes. On a tetrahedron with vertexes $P_0, P_1, P_2, P_3$, the trial function is of the form:

$$\phi(x, y, z) = u_0\lambda_0 + u_1\lambda_1 + u_2\lambda_2 + u_3\lambda_3, \tag{8.2.4}$$

where $(\lambda_0, \lambda_1, \lambda_2, \lambda_3)$ is the volume coordinates of $(x, y, z)$. Obviously we have that $\phi(P_i) = u_i$ for $i = 0, 1, 2, 3$, that any $u_h \in U_h$ is globally continuous, and that $U_h \subset H^1(\Omega)$. The test function space $V_h$ related to $T^*_h$ is taken as a piecewise polynomial space of degree $\bar{k}$. $V_h$ is not required to have any global smoothness, but it has the same dimension as $U_h$. Each node $P_0$ bears some nodal basis functions, of which the number depends on the type and the number of the interpolation conditions of $U_h$ at $P_0$. For instance, if $U_h$ is a piecewise linear function space, then the nodal basis function of $V_h$ at $P_0$ is

$$\psi_{P_0}(x, y, z) = \begin{cases} 1, & \text{for } (x, y, z) \in K^*_{P_0}, \\ 0, & \text{elsewhere.} \end{cases}$$

If $U_h$ is a standard finite element space, then the basis functions of $V_h$ are of the form

$$\psi_{P_0}(x, y, z) = \begin{cases} \dfrac{1}{l!m!n!}(x - x_0)^l(y - y_0)^m(z - z_0)^n, \\ \qquad\qquad (x, y, z) \in K^*_{P_0}, \\ 0, \quad \text{elsewhere.} \end{cases}$$

Next we consider the generalized difference method for the following Poisson equation

$$\begin{cases} \nabla(a\nabla u) = \dfrac{\partial}{\partial x}\Big(a\dfrac{\partial u}{\partial x}\Big) + \dfrac{\partial}{\partial y}\Big(a\dfrac{\partial u}{\partial y}\Big) + \dfrac{\partial}{\partial z}\Big(a\dfrac{\partial u}{\partial z}\Big) = f, \text{ on } \Omega, \\ u|_{\partial\Omega} = 0, \end{cases} \tag{8.2.5}$$

where $a = a(x, y, z) \geq a_0 > 0$. Assume $U_h$ and $V_h$ are piecewise linear and piecewise constant function spaces, respectively. Integrate (8.2.5) on $K^*_{P_0}$ and use the Gauss formula to obtain

$$\int_{\partial K^*_{P_0}} a\frac{\partial u}{\partial \nu}\mathrm{d}s = \int_{K^*_{P_0}} f\mathrm{d}x\mathrm{d}y\mathrm{d}z, \tag{8.2.6}$$

where $\nu$ is the unit outer normal direction of $\partial K^*_{P_0}$. We see from Fig. 8.2.2 that $\partial K^*_{P_0}$ is cut into six planar triangles with a common vertex (the barycenter) $Q$, by the tetrahedron $K$ with vertexes $P_0, P_1, P_2, P_3$. These six triangles are divided into three pairs, with the barycenters of $\Delta P_0P_1P_2$, $\Delta P_0P_2P_3$, and $\Delta P_0P_1P_3$ respectively as a common vertex. Now we can deduce the surface integral in (8.2.6) into a sum of integrals on these triangles. Notice

$$\frac{\partial u}{\partial \nu} = \nabla u \cdot \nu, \tag{8.2.7}$$

where $\nu$ is the unit outer normal directions of these triangles on $\partial K^*_{P_0}$. Also note that $\nabla u$ is a constant vector on $K$. Thus we may use (8.2.2) and (8.2.4) (changing the numbering 1,2,3,4 into 0,1,2,3) to get

$$\nabla u = \Big(\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}, \frac{\partial u}{\partial z}\Big), \tag{8.2.8}$$

$$\begin{aligned} \frac{\partial u}{\partial x} &= u_0\frac{\partial \lambda_0}{\partial x} + u_1\frac{\partial \lambda_1}{\partial x} + u_2\frac{\partial \lambda_2}{\partial x} + u_3\frac{\partial \lambda_3}{\partial x}, \\ \frac{\partial u}{\partial y} &= u_0\frac{\partial \lambda_0}{\partial y} + u_1\frac{\partial \lambda_1}{\partial y} + u_2\frac{\partial \lambda_2}{\partial y} + u_3\frac{\partial \lambda_3}{\partial y}, \\ \frac{\partial u}{\partial z} &= u_0\frac{\partial \lambda_0}{\partial z} + u_1\frac{\partial \lambda_1}{\partial z} + u_2\frac{\partial \lambda_2}{\partial z} + u_3\frac{\partial \lambda_3}{\partial z}, \end{aligned} \tag{8.2.9}$$

where

$$\begin{cases} \dfrac{\partial \lambda_i}{\partial x} = \dfrac{1}{6V} X_{i4}, \ i = 0, 1, 2, \\ \dfrac{\partial \lambda_3}{\partial x} = -\dfrac{1}{6V}(X_{04} + X_{14} + X_{24}), \\ X_{i4} = \begin{vmatrix} y_{j4} & y_{k4} \\ z_{j4} & z_{k4} \end{vmatrix}, \end{cases} \tag{8.2.10}$$

where the subscripts $j = i+1, \ k = j+1, \ i = k+1$. The representations of $\partial\lambda_i/\partial y$ and $\partial\lambda_i/\partial z$ $(i = 0, 1, 2, 3)$ can be obtained in like manner. The normal vector $\nu$ depends on the different triangles on $\partial K^*_{P_0}$. For example, the outer normal direction of $\Delta Q q_3 m_1$ and $\Delta Q q_3 m_2$ are respectively

$$\frac{\overline{q_3 Q} \times \overline{q_3 m_1}}{|\overline{q_3 Q}||\overline{q_3 m_1}|}, \ \frac{\overline{q_3 m_2} \times \overline{q_3 Q}}{|\overline{q_3 m_2}||\overline{q_3 Q}|}.$$

Here the symbol $\times$ denotes the vector product. As regards the coefficient $a$, we usually take its average value on the vertexes of $\Delta Q q_3 m_i$:

$$a_{3i} = \frac{1}{3}(a(Q) + a(q_3) + a(m_i)), \ i = 1, 2.$$

Summarizing, we can deduce (8.2.5) into the following difference equation:

$$\sum_K (\nabla u)_K \sum a_{3i} \mathrm{meas}(\Delta Q q_3 m_i) \nu_{\Delta Q q_3 m_i} = \int_{K^*_{P_0}} f \mathrm{d}x \mathrm{d}y \mathrm{d}z. \tag{8.2.11}$$

In practice, in the interior of $\Omega$, we often first place a cuboid grid, and then divide each cuboid element into six tetrahedrons, while on the boundary of $\Omega$ we use directly a tetrahedral grid. So we get a difference equation on a rectangular networks.

In [A-60] the equation (8.2.5) is written in cylindrical coordinates:

$$\frac{1}{r}\frac{\partial}{\partial r}\Big(ar\frac{\partial u}{\partial r}\Big) + \frac{1}{r}\frac{\partial}{\partial \theta}\Big(\frac{a}{r}\frac{\partial u}{\partial \theta}\Big) + \frac{\partial}{\partial z}\Big(a\frac{\partial u}{\partial z}\Big) = -p.$$

When $a$ is a constant, this equation is simplified as

$$\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} + \frac{\partial^2 u}{\partial z^2} = -\frac{p}{a}.$$

In [A-60], first they place on $(r, z)$ plane a barycenter dual grid, then they connect certain nodes in between the surfaces $\theta = \theta_l, \theta_m, \theta_n$ into some polyhedrons to finally form a dual element. There it is allowed for different dual elements to have overlapping interiors. Thus their method differs from ours here.

## 8.3 Numerical Simulation of Underground Water Pollution

Underground water is often contaminated by, e.g., the sewage out of factories or mines, and the chemical fertilizer and pesticide in agriculture, which seep into the ground with rain or irrigation and drainage. These solutes in the water may perform a convective motion with respect to the underground water, and/or a diffusive motion due to the density diffusion of the water molecules. In hydrogeology and environmental science, computer applications using mathematical models are widely used to study the law of the motion of the contaminated water. A mathematical model describing the contaminated water, or the water with any chemical solute (e.g., saline-alkali) in general, is the following equation of the solute density $C$ (cf. [A-40]):

$$\frac{\partial(mC)}{\partial t} = \mathrm{div}(mD\mathrm{grad}C) - \mathrm{div}(VmC) - \frac{C'W}{n}, \text{ on } \Omega \subset \mathrm{R}^l. \quad (8.3.1)$$

Here $l = 1, 2$ or $3$. To fix the idea, we take $l = 2$. The other notations are explained below.

$m$ : The saturation thickness of the water-bearing formation, usually depending on $(x, y, z)$.

$V = (V_x, V_y)^T$: The velocity of the water, assumed to be known.

$D = \begin{pmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{pmatrix}$ : The diffusion coefficient tensor, depending on the composition of the solute.

$W$: The amount of the water flooded into (positive) or pumped off from (negative) a unit area of water-bearing formation. In particular, if the water goes in or out through a well $P_0 = (x_0, y_0)$, i.e., $P_0$ is either a source or a sink, then $W = Q\delta(x - x_0, y - y_0)$, where $Q$ is the amount of the water.

$C'$: The density of the solute, which is known for a source, and unknown for a sink.

Besides, the boundary and initial conditions are also needed to be provided.

The first term appearing in the right-hand side of (8.3.1) is referred to as a diffusion term, while the second a convection term, and the third a source term. In fact, the differential equation (8.3.1) may describe many other physical or chemical phenomena, so long as $C$ and the other notations are explained accordingly. For instance, $C$ may denote the mass ratio of chemical composition, the heat enthalpy, the temperature, or the kinetic energy of turbulent flow, etc. (cf. [B-73].)

### 8.3.1 Generalized difference scheme

Let us place a triangulation $T_h = \{K\}$ and its dual grid $T_h^* = \{K^*\}$ (barycenter or circumcenter dual grids, cf. Fig. 8.3.1). Both the sources and the sinks must be taken as nodes. If the coefficient of the diffusion term is discontinuous on a line $L$, then $L$ should be cut into several line segments by some nodes, such that each segment is a side of an element. We assume that $C$ is continuous crossing such an $L$, i.e., $C_+ = C_-$, where $+$ and $-$ denote the two sides of $L$ respectively. The flow of the solute is also assumed to be continuous:

$$(m(D\text{grad}C) \cdot \nu)_+ = (m(D\text{grad}C) \cdot \nu)_-,$$

where $\nu$ is the unit outer normal vector of $L$. Under these assumptions, a generalized differencing of (8.3.1) at a node on $L$ can be done precisely as at other nodes.

Let $T_h^*$ be a barycenter dual grid. A node $P_0$ together with its neighbouring nodes and corresponding dual elements are depicted in Fig. 8.3.1. The trial function space $U_h$ is the piecewise linear function

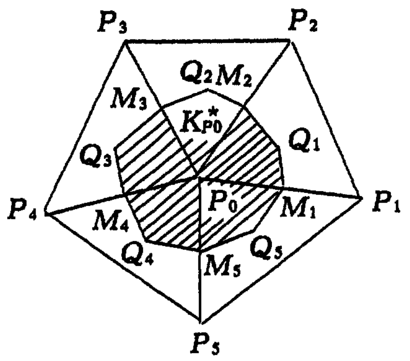



Fig. 8.3.1

space related to $T_h$ with the vertexes of the elements as the nodes, and the test function space is the piecewise constant space corresponding to $T_h^*$. Integrate (8.3.1) on $K_{P_0}^*$ to obtain

$$\begin{aligned} &\int_{K_{P_0}^*} \frac{\partial(mC)}{\partial t} \mathrm{d}x\mathrm{d}y \\ = &\int_{K_{P_0}^*} \mathrm{div}(mD\mathrm{grad}C)\mathrm{d}x\mathrm{d}y - \int_{K_{P_0}^*} \mathrm{div}(VmC)\mathrm{d}x\mathrm{d}y \\ &- \int_{K_{P_0}^*} \frac{C'W}{n} \mathrm{d}x\mathrm{d}y. \end{aligned} \tag{8.3.2}$$

By Green's formula we have for $C = C_h \in U_h$

$$\int_{K_{P_0}^*} \mathrm{div}(mD\mathrm{grad}C_h)\mathrm{d}x\mathrm{d}y = \int_{\partial K_{P_0}^*} m(D\mathrm{grad}C_h) \cdot \nu \mathrm{d}s, \tag{8.3.3}$$

$$\int_{K_{P_0}^*} \mathrm{div}(VmC_h)\mathrm{d}x\mathrm{d}y = \int_{\partial K_{P_0}^*} mC_h(V \cdot \nu)\mathrm{d}s. \tag{8.3.4}$$

Denote by $\triangle_{Q_i}$ a triangular element with a barycenter $Q_i$. Evaluate the above line integral piecewise on the fold line segments obtained by intersecting the integral line with $\triangle_{Q_i}$. One can readily work out the formula for the piecewise integrals thanks to the linearity of $u_h \in U_h$.

For example, in $\triangle_{Q_1}$ we have

$$\begin{aligned}&\int_{\overline{M_1Q_1M_2}} m(D\text{grad}C_h)\cdot\nu\text{d}s\\ =\ &\int_{\overline{M_1Q_1}} m\text{grad}C_h\cdot D\nu_1\text{d}s+\int_{\overline{Q_1M_2}} m\text{grad}C_h\cdot D\nu_2\text{d}s.\end{aligned} \tag{8.3.3a}$$

where

$$\begin{aligned}\text{grad}C_h=\ &\frac{1}{2\triangle_{Q_1}}((y_{P_1}-y_{P_2})C_0+(y_{P_2}-y_{P_0})C_1\\ &+(y_{P_0}-y_{P_1})C_2+(x_{P_2}-x_{P_1})C_0\\ &+(x_{P_0}-x_{P_2})C_1+(x_{P_1}-x_{P_0})C_2),\end{aligned} \tag{8.3.3b}$$

$$\begin{cases}\nu_1=(y_{Q_1}-y_{M_1},-(x_{Q_1}-x_{M_1}))/|\overline{M_1Q_1}|,\\ \nu_2=(y_{M_2}-y_{Q_1},-(x_{M_2}-x_{Q_1}))/|\overline{Q_1M_2}|,\end{cases} \tag{8.3.3c}$$

where $(x_P, y_P)$ denote the coordinates of a point $P$. Also note

$$\begin{aligned}&\int_{\overline{M_1Q_1M_2}} mC_h(V\cdot\nu)\text{d}s\\ =\ &\int_{\overline{M_1Q_1}} mC_h(V\cdot\nu_1)\text{d}s+\int_{\overline{Q_1M_2}} mC_h(V\cdot\nu_2)\text{d}s,\end{aligned} \tag{8.3.4a}$$

where $\nu_1$ and $\nu_2$ are as in (8.3.3c). On the line segment $\overline{M_1Q_1}$: $y = y_{M_1}+\frac{y_{Q_1}-y_{M_1}}{x_{Q_1}-x_{M_1}}(x-x_{M_1})$, we have

$$C_h=\frac{x_{Q_1}-x}{x_{Q_1}-x_{M_1}}C_{M_1}+\frac{y-y_{M_1}}{y_{Q_1}-y_{M_1}}C_{Q_1}. \tag{8.3.4b}$$

Similarly on $\overline{Q_1M_2}$: $y=y_{Q_1}+\frac{y_{M_2}-y_{Q_1}}{x_{M_2}-x_{Q_1}}(x-x_{Q_1})$, we have

$$C_h=\frac{x_{M_2}-x}{x_{M_2}-x_{M_1}}C_{Q_1}+\frac{y-y_{Q_1}}{y_{M_2}-y_{Q_1}}C_{M_2}. \tag{8.3.4c}$$

Also observe that

$$C_{Q_1}=\frac{1}{3}(C_{P_0}+C_{P_1}+C_{P_2}),\ C_{M_i}=\frac{1}{2}(C_{P_0}+C_{P_i}). \tag{8.3.5}$$

As regards the source term on the right-hand side of (8.3.2), if $P_0$ is not a well, then it is directly computed ($C'$ is known for a source, and unknown for a sink). On the other hand, if $P_0$ is a well, then

$$W = Q\delta(x - x_{P_0}, y - y_{P_0}).$$

In this case

$$\int_{K^*_{P_0}} \frac{C'Q}{n}\delta(x - x_{P_0}, y - y_{P_0})\mathrm{d}x\mathrm{d}y = \frac{C'(P_0)}{n(P_0)}Q(P_0). \tag{8.3.6}$$

Now we have successfully discretized the space variable on the right-hand side of (8.3.2). To further discretize the derivative with respect to $t$ on the left-hand side, we take a time step size $\tau > 0$ and nodes $t_k = k\tau$ $(k = 0, 1, \cdots, K)$. We can use explicit or implicit Euler's methods, or Crank-Nicolson method to approximate (8.3.2). For example, the Crank-Nicolson method gives

$$\begin{aligned}
&\tau^{-1}\int_{K^*_{P_0}} (m^{k+1}C_h^{k+1} - m^k C_h^k)\mathrm{d}x\mathrm{d}y \\
= \ &\frac{1}{2}\int_{K^*_{P_0}} \mathrm{div}(m^{k+1/2}D\mathrm{grad}(C_h^k + C_h^{k+1}))\mathrm{d}x\mathrm{d}y \\
&-\frac{1}{2}\int_{K^*_{P_0}} \mathrm{div}(V^{k+1/2}m^{k+1/2}(C_h^k + C_h^{k+1}))\mathrm{d}x\mathrm{d}y \\
&-\frac{1}{2}\int_{K^*_{P_0}} \frac{W^k}{n}C_h'^k\mathrm{d}x\mathrm{d}y.
\end{aligned} \tag{8.3.7}$$

This together with boundary and initial conditions gives the generalized difference scheme for (8.3.1). Quite a few applications of generalized difference methods are discussed in [A-40].

**Remark** If a circumcenter grid $T_h^*$ is adopted, then the related computation will be simpler, resulting in a difference method on irregular networks, as is called in the underground water computation. As in §8.2, we may extend the generalized difference scheme to tetrahedral, cuboid or triangular prismatic grids on a three-dimensional field. (cf. [A-40].)

### 8.3.2 Generalized upwind difference schemes

Generalized difference schemes have been widely and, generally speaking, satisfactorily used in the underground water computation. But in the computation of contaminated underground water, one often encounters a class of problems where the diffusion coefficient is much less than the convection speed. In such a case, a standard generalized difference method fails to approximate accurately the transitional band that results from the diffusion, and it may bring in undesirable oscillations. Upwind difference schemes are often used to overcome this difficulty. To illustrate the idea, let us investigate a one-dimensional convection-diffusion equation ([A-40]):

$$\frac{\partial C}{\partial t} = D\frac{\partial^2 C}{\partial x^2} - V\frac{\partial C}{\partial x}, \tag{8.3.8a}$$

where $D$ and $V$ are positive constants. The initial and boundary conditions are

$$\begin{cases} C(x,0) = 0, & x > 0, \\ C(0,t) = C_0, & t \geq 0, \\ C(\infty,t) = 0, & t \geq 0. \end{cases} \tag{8.3.8b}$$

The true solution to this problem is

$$C(x,t) = \frac{C_0}{2}\Big\{\mathrm{erfc}\Big(\frac{x - Vt}{2\sqrt{Dt}}\Big) + \exp\Big(\frac{Vx}{D}\Big)\mathrm{erfc}\Big(\frac{x + Vt}{2\sqrt{Dt}}\Big)\Big\}.$$

$$\Big(\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}\mathrm{d}t.\Big)$$

Apply the generalized difference method to (8.3.8a) to obtain the following implicit difference scheme:

$$\frac{C_j^{k+1} - C_j^k}{\tau} = D\frac{C_{j+1}^{k+1} - 2C_j^{k+1} + C_{j-1}^{k+1}}{h^2} - V\frac{C_{j+1}^{k+1} - C_{j-1}^{k+1}}{2h}.$$

Define the Peclet number as

$$Pe = Vh/D.$$

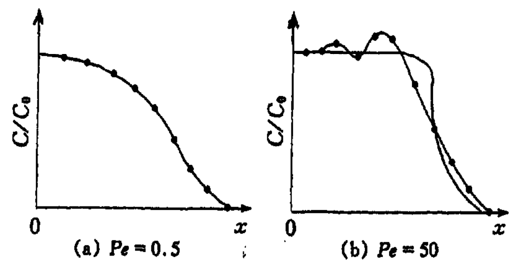



Fig. 8.3.2

For fixed step size $h > 0$, $Pe$ varies with the ratio $V/D$. A comparison of the true solution and the implicit difference solution at $t = 50\tau$ is depicted in Fig. 8.3.2(a) for $Pe = 0.5$ and (b) for $Pe = 50$. We observe that the approximation is fairly good for a small Peclet number. But for a large Peclet number, i.e., for the convection-dominated diffusion case, the density front becomes wider and more level, and undesirable oscillations appear. The too wide and level band is caused by an extra diffusion in the numerical discretization, which can be decreased by a smaller step size. On the other hand, the oscillation is due to the approximation of the convection term by a central differencing. Upwind schemes are often used to eliminate the oscillation as discussed below.

Consider a two-dimensional solute transfer equation

$$\frac{\partial C}{\partial t} = \operatorname{div}(D\operatorname{grad}C) - \operatorname{div}(VC) + I, \tag{8.3.9}$$

where the diffusion tensor $D$ and the convection speed $V$ are known, as in equation (8.3.2), and $I$ is the source term with $I = C'Q\delta(x - x_0, y - y_0)$ at a well. As before, assume $T_h = \{K\}$ is a triangulation, and $T_h^* = \{K_{P_0}^*\}$ is a barycenter dual grid (cf. Fig. 8.3.1). Let $U_h$ be the piecewise linear, globally continuous function space, and $V_h$ the piecewise constant function space. Denote by $\Pi_h^*$ the interpolation projection operator from $U_h$ to $V_h$: For given $C_h \in U_h$, $\Pi_h^* C_h \in V_h$ and $(\Pi_h^* C_h)(P_0) = C_h(P_0)$. Integrate equation (8.3.9) on $K_{P_0}^*$, set $C = C_h \in U_h$, and replace $C_h$ in the convection term by $\Pi_h^* C_h$, then we have

$$\int_{K^*_{P_0}} \frac{\partial C_h}{\partial t} \mathrm{d}x\mathrm{d}y = \int_{K^*_{P_0}} \mathrm{div}(D\mathrm{grad}C_h)\mathrm{d}x\mathrm{d}y$$

$$- \int_{K^*_{P_0}} \mathrm{div}(V\Pi^*_h C_h)\mathrm{d}x\mathrm{d}y + \int_{K^*_{P_0}} I\mathrm{d}x\mathrm{d}y. \tag{8.3.10}$$

The diffusion term on the right-hand side is computed according to (8.3.3) and (8.3.3a-c) with $m = 1$, while the source term according to (8.3.6) with $n = 1$. The convection term is treated as in §7.2. To elaborate, we apply Green's formula

$$\int_{K^*_{P_0}} \mathrm{div}(V\Pi^*_h C_h)\mathrm{d}x\mathrm{d}y = \int_{\partial K^*_{P_0}} (V \cdot \nu)\Pi^*_h C_h \mathrm{d}s. \tag{8.3.11}$$

Set (cf. Fig. 8.3.1) $\Gamma_{0l\delta} = \overline{Q_l M_{l+\delta}}$, $l = 1, 2, \cdots, 5$, $\delta = 0, 1$. Define

$$\beta_{0l\delta} = \int_{\Gamma_{0l\delta}} (V \cdot \nu)\mathrm{d}s, \tag{8.3.12}$$

$$(\partial K^*_{P_0})_- = \Big\{ \bigcup_{\beta_{0l\delta} \le 0} \Gamma_{0l\delta} : 1 \le l \le 5, \delta = 0, 1 \Big\} \ \text{(flow in)},$$

$$(\partial K^*_{P_0})_+ = \Big\{ \bigcup_{\beta_{0l\delta} > 0} \Gamma_{0l\delta} : 1 \le l \le 5, \delta = 0, 1 \Big\} \ \text{(flow out)},$$

$$\beta^+_{0l\delta} = \max\{\beta_{0l\delta}, 0\},\ \beta^-_{0l\delta} = \max\{-\beta_{0l\delta}, 0\}, \tag{8.3.13}$$

and apply the following approximation

$$\int_{\partial K^*_{P_0}} (V \cdot \nu)\Pi^*_h C_h \mathrm{d}s \approx \sum_{\substack{1 \le l \le 5 \\ \delta = 0,1}} \{\beta^+_{0l\delta} C_h(P_0) - \beta^-_{0l\delta} C_h(P_l)\}.$$

Then we have (cf. §7.2)

$$\int_{K^*_{P_0}} \mathrm{div}(V\Pi^*_h C_h)\mathrm{d}x\mathrm{d}y = \sum_{\substack{1 \le l \le 5 \\ \delta = 0,1}} \{\beta^+_{0l\delta} C_h(P_0) - \beta^-_{0l\delta} C_h(P_l)\}. \tag{8.3.14}$$

Finally, discretizing the time yields explicit, implicit, or Crank-Nicolson generalized difference schemes. For example, the Crank-Nicolson scheme reads:

$$\begin{aligned}
&\tau^{-1}\int_{K^*_{P_0}}(C_h^{k+1}-C_h^k)\mathrm{d}x\mathrm{d}y\\
=\ &\frac{1}{2}\int_{K^*_{P_0}}\mathrm{div}(D\mathrm{grad}(C_h^{k+1}+C_h^k))\mathrm{d}x\mathrm{d}y\\
&-\frac{1}{4}\int_{K^*_{P_0}}\mathrm{div}[(V^{k+1}+V^k)\Pi_h^*(C_h^{k+1}+C_h^k)]\mathrm{d}x\mathrm{d}y\\
&+\frac{1}{2}\int_{K^*_{P_0}}(I^{k+1}+I^k)\mathrm{d}x\mathrm{d}y.
\end{aligned}\tag{8.3.15}$$

The diffusion term on the right-hand side is computed according to (8.3.3) and (8.3.3a-c) with $m = 1$, and the convection term according to (8.3.12) and (8.3.14).

The computation will be simpler if a circumcenter dual grid is adopted. In this case, as in Fig. 8.1.1, we once again integrate the equation (8.3.9) on $K^*_{P_0}$ to obtain an equation similar to (8.3.10). The diffusion term is treated analogously as (8.3.3), but the line integral is computed on the pieces $\overline{Q_iQ_{i+1}}$. In particular when $D = \alpha E$ (a scalar matrix)

$$\begin{aligned}
&\int_{K^*_{P_0}}\mathrm{div}(\alpha E\mathrm{grad}C_h)\mathrm{d}x\mathrm{d}y\\
=\ &\alpha\int_{\partial K^*_{P_0}}\frac{\partial C_h}{\partial\nu}\mathrm{d}s=\alpha\sum_{i=1}^{6}\int_{\overline{Q_iQ_{i+1}}}\frac{\partial C_h}{\partial\nu}\mathrm{d}s\\
=\ &\alpha\sum_{i=1}^{6}\frac{C_h(P_{i+1})-C_h(P_0)}{|\overline{P_0P_{i+1}}|}\cdot|\overline{Q_iQ_{i+1}}|.\\
&(P_7=P_1,\ Q_7=Q_1.)
\end{aligned}\tag{8.3.16}$$

The computation of the convection term is similar to (8.3.11). In detail, we have

$$\int_{\partial K^*_{P_0}}(V\cdot\nu)\Pi_h^*C_h\mathrm{d}s=\sum_{i=1}^{6}\int_{\overline{Q_iQ_{i+1}}}(V\cdot\nu)\Pi_h^*C_h\mathrm{d}s.\ (Q_7=Q_1.)$$

Set

$$\beta_{i+1} = \int_{\overline{Q_iQ_{i+1}}} (V \cdot \nu)\mathrm{d}s,$$

$$\beta_{i+1}^{+} = \max\{\beta_{i+1}, 0\},\ \beta_{i+1}^{-} = \max\{-\beta_{i+1}, 0\}.$$

Then we have

$$\begin{aligned} &\int_{\overline{Q_iQ_{i+1}}} (V \cdot \nu)\Pi_h^* C_h \mathrm{d}s \\ = &\ \beta_{i+1}^{+} C_h(P_0) - \beta_{i+1}^{-} C_h(P_{i+1}) \\ = &\begin{cases} \beta_{i+1} C_h(P_0), & \text{as } \beta_{i+1} > 0, \quad \text{(flowing out of } K_{P_0}^*), \\ \beta_{i+1} C_h(P_{i+1}), & \text{as } \beta_{i+1} \le 0, \quad \text{(flowing into } K_{P_0}^*). \end{cases} \end{aligned}$$

This means that on the outer normal direction $\nu$, $C_h$ takes the upwind value. Therefore

$$\int_{K_{P_0}^*} \mathrm{div}(V\Pi_h^* C_h)\mathrm{d}x\mathrm{d}y = \sum_{i=1}^{6} \{\beta_{i+1}^{+} C_h(P_0) - \beta_{i+1}^{-} C_h(P_{i+1})\}, \quad (8.3.17)$$

where $Q_7 = Q_1$ etc.

As pointed out in Chapter 7, in the one-dimensional case, the generalized difference scheme here becomes precisely the usual upwind scheme. If we apply it to (8.3.8), then the oscillation disappears, the density front gets narrower and its position is more accurate.

### 8.3.3 Upwind weighted multi-element balancing method

Sun [A-40] suggests a kind of weighted upwind difference scheme, referred to as a upwind weighted multi-element balancing method. As before, let $T_h = \{K\}$ be a triangulation, and $T_h^* = \{K_{P_0}^*\}$ a barycenter dual grid. A node $P_0$ and its adjacent points distribute as in Fig. 8.3.1. Choose any an element $K_{Q_i} = \triangle P_0 P_i P_{i+1}$ with $P_0$ as a vertex and $Q_i$ as the barycenter. Connect $Q_i$ with the three vertexes to form three smaller triangles: $\triangle P_0 P_i Q_i, \triangle P_i P_{i+1} Q_i,\ \triangle P_{i+1} P_0 Q_i$ (cf. Fig. 8.3.3). Denote the value of $C$ at $P_l$ by $C_{P_l}$ $(l = 0, i, i+1)$.

Take their weighted average

$$\begin{cases} C_{Q_i} = \omega_0 C_{P_0} + \omega_i C_{P_i} + \omega_{i+1} C_{P_{i+1}}, \\ \omega_0 + \omega_i + \omega_{i+1} = 1, \end{cases} \tag{8.3.18}$$

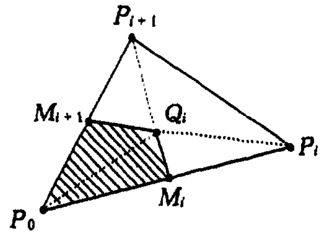



Fig. 8.3.3

where $\omega_l$ stands for the upwind weight of $P_l$ and its value is to be determined later on. The restriction on $K_{Q_i}$ of an element $C_h \in U_h$ is a piecewise linear function. More precisely, it is linear respectively on the three sub-triangles, satisfying the interpolation conditions:

$$C_h(P_l) = C_{P_l} \ \ (l = 0, i, i+1), \ C_h(Q_i) = C_{Q_i}.$$

In addition, it satisfies the first boundary condition on the exterior boundary. $V_h$ remains to be the piecewise constant function space, and vanishes at the dual elements where the first boundary condition is given.

Next, let us extend the generalized difference method to the convection-diffusion equation (8.3.9). An integration of this equation on $K^*_{P_0}$ leads to

$$\begin{aligned} \int_{K^*_{P_0}} \frac{\partial C_h}{\partial t} \mathrm{d}x\mathrm{d}y = & \int_{K^*_{P_0}} \mathrm{div}(D\mathrm{grad}C_h)\mathrm{d}x\mathrm{d}y \\ & - \int_{K^*_{P_0}} \mathrm{div}(VC_h)\mathrm{d}x\mathrm{d}y + \int_{K^*_{P_0}} I\mathrm{d}x\mathrm{d}y. \end{aligned} \tag{8.3.19}$$

For the diffusion term we have ($M_6 = M_1$)

$$\int_{K^*_{P_0}} \mathrm{div}(D\mathrm{grad}C_h)\mathrm{d}x\mathrm{d}y = \int_{\partial K^*_{P_0}} (D\mathrm{grad}C_h)\cdot\nu\mathrm{d}s$$
$$= \sum_{i=1}^{5}\Big\{\int_{\overline{Q_iM_i}}(D\mathrm{grad}C_h)\cdot\nu\mathrm{d}s + \int_{\overline{Q_iM_{i+1}}}(D\mathrm{grad}C_h)\cdot\nu\mathrm{d}s\Big\}. \tag{8.3.20}$$

For the convection term

$$\int_{K^*_{P_0}} \mathrm{div}(VC_h)\mathrm{d}x\mathrm{d}y = \int_{\partial K^*_{P_0}} (V\cdot\nu)C_h\mathrm{d}s$$
$$= \sum_{i=1}^{5}\Big\{\int_{\overline{Q_iM_i}}(V\cdot\nu)C_h\mathrm{d}s + \int_{\overline{Q_iM_{i+1}}}(V\cdot\nu)C_h\mathrm{d}s\Big\}. \tag{8.3.21}$$

Expressing $C_h$ on $\triangle P_0P_iQ_i$ and $\triangle P_0Q_iP_{i+1}$ in terms of $C_{P_0}, C_{P_i}$ and $C_{P_{i+1}}$, we can obtain an ordinary differential equation of $C_{P_0}(t)$. We may further discretize the time $t$ to obtain explicit or implicit difference schemes. For details, see [A-40].

It is an interesting question how to choose the weight coefficients $\omega_i$. If we take $\omega_i = \frac{1}{3}$, then (8.3.19) is nothing but an ordinary generalized difference scheme without any upwind weighting. Next, we present a method to determine the upwind weighting for each element. Let $V$ be an average velocity vector of the element $\triangle P_0P_iP_{i+1}$. (cf. Fig. 8.3.3.) For instance, we can set $V = V(Q_i)$. Let $V_{0,i}, V_{i,i+1}$ and $V_{i+1,0}$ be projections of $V$ onto $\overline{P_0P_i}$, $\overline{P_iP_{i+1}}$, $\overline{P_{i+1}P_0}$, respectively. Define the local Peclet number

$$\tau_{0,i} = V_{0,i}|\overline{P_0P_i}|/(\alpha_i|V|), \tag{8.3.22}$$

where $|V|$ is the length of vector $V$ and $\alpha_i$ is the local diffusion on $\triangle P_0P_iP_{i+1}$. $\tau_{0i} > 0$ means that $P_0$ is on the upwind side of $P_i$, and $\tau_{0i} < 0$ the downwind side. Similarly we can define $\tau_{i,i+1}$ and $\tau_{i+1,0}$. Write $\tau_0 = \tau_{0,i} + \tau_{i+1,0}$, indicating, both qualitatively and quantitatively, the upwind or downwind position of $P_0$ with respect to $P_i$ and $P_{i+1}$. $\tau_i$ and $\tau_{i+1}$ can be defined in like manner. Finally we define

$$\omega_0 = \frac{1}{3}(1+\lambda\tau_0) = \frac{1}{3} + \lambda\frac{V_{0,i}|\overline{P_0P_i}| - V_{i+1,0}|\overline{P_0P_{i+1}}|}{\alpha_i|V|}, \tag{8.3.23a}$$

$$\omega_i = \frac{1}{3}(1 + \lambda\tau_i) = \frac{1}{3} + \lambda\frac{V_{i,i+1}|\overline{P_iP_{i+1}}| - V_{0,i}|\overline{P_0P_i}|}{\alpha_i|V|}, \quad (8.3.23b)$$

$$\omega_{i+1} = \frac{1}{3}(1 + \lambda\tau_{i+1}) = \frac{1}{3} + \lambda\frac{V_{i+1,0}|\overline{P_0P_{i+1}}| - V_{i,i+1}|\overline{P_iP_{i+1}}|}{\alpha_i|V|}. \quad (8.3.23c)$$

Obviously $\omega_0 + \omega_i + \omega_{i+1} = 1$. Here the coefficient $\lambda > 0$ remains to be chosen. Numerical experiments indicate that $\lambda$ should be neither too large nor too small. A too large $\lambda$ will cause extra numerical diffusion, while a too small one will not be good enough to prevent the oscillation of the solution. A proper value of $\lambda$ might be between 0.004-0.005, as suggested by the numerical experiments.

## 8.4 Stokes Equation

Consider a Navier-Stokes equation describing an $n$-dimensional viscous incompressible flow

$$\begin{cases} \dfrac{\partial u}{\partial t} - \mu\Delta u + (u \cdot \nabla)u + \text{grad}p = f, & (8.4.1a) \\ \text{div}u = 0, & (8.4.1b) \end{cases}$$

where $u = (u_1, \cdots, u_n)^T$ is the flow velocity, $p$ the pressure, $\mu > 0$ the viscosity coefficient, and $f = (f_1, \cdots, f_n)^T$ the density of the body force. Assume that the velocity $u$ is small enough such that the nonlinear convection term can be ignored, then equation (8.4.1) is reduced to a Stokes equation:

$$\begin{cases} \dfrac{\partial u}{\partial t} - \mu\Delta u + \text{grad}p = f, \\ \text{div}u = 0. \end{cases}$$

In this section, let $\Omega \subset R^2$ and restrict our attention to the steady-state case, i.e., $\frac{\partial u}{\partial t} = 0$. So we consider the steady-state Stokes equation:

$$\begin{cases} -\mu\Delta u + \text{grad}p = f, \text{ on } \Omega \subset \mathrm{R}^2, & (8.4.2a) \\ \text{div}u = 0. & (8.4.2b) \end{cases}$$

Although it is only a linear equation, it has attracted people's attention due to the presence of the incompressible condition. We suppose $u$ satisfies the first boundary condition: $u = 0$ when $x = (x_1, x_2) \in \partial\Omega$. Let $H_0^1(\Omega)$ be a usual Sobolev space, and

$$L_0^2(\Omega) = \Big\{q \in L^2(\Omega) : (q, 1) = \int_\Omega q\mathrm{d}x = 0\Big\}.$$

Define

$$a(u, v) = \mu \sum_{i,j=1}^{2} \Big(\frac{\partial u_i}{\partial x_j}, \frac{\partial v_i}{\partial x_j}\Big),$$

$$b(v, p) = \sum_{i=1}^{2} \Big(\frac{\partial p}{\partial x_i}, v_i\Big), \quad c(v, q) = -(q, \mathrm{div} v).$$

Then, a variational form for (8.4.2) is: Find $(u, p) \in (H_0^1(\Omega))^2 \times L_0^2(\Omega)$ such that

$$\begin{cases} a(u, v) + b(v, p) = (f, v), & \forall v \in (H_0^1(\Omega))^2, \quad (8.4.3a) \\ c(u, q) = 0, & \forall q \in L_0^2(\Omega). \quad (8.4.3b) \end{cases}$$

Assume that $\Gamma = \partial\Omega$ is Lipschitz continuous, that $\Omega$ is a convex domain, and that $f \in (L^2(\Omega))^2$. Then (8.4.3) possesses a unique solution $(u, p) \in (H_0^1(\Omega))^2 \times (H^1(\Omega) \cap L_0^2(\Omega))$, and there is a constant $C$ independent of $(u, p)$ and $f$ such that (see [B-34] for details)

$$\|u\|_{2,\Omega} + \|p\|_{1,\Omega} \le C\|f\|_{0,\Omega}.$$

### 8.4.1 Nonconforming generalized difference method

Now let us construct a generalized difference method approximating (8.4.2). Let $\Omega$ be a convex polygonal region, and $T_h = \{K_Q\}$ a triangulation of $\Omega$, where $K_Q$ is a triangular element with barycenter $Q$. Take the midpoints of the sides of the element as nodes. Denote the interior nodes of $\Omega$ by $P_1, \cdots, P_M$, and the boundary nodes on $\Gamma$ by $P_{M+1}, \cdots, P_N$. The trial function space related to $U_h$ is a piecewise linear function space, with $P_1, \cdots, P_N$ as the interpolation nodes and with zero value at boundary nodes. Clearly an element in $U_h$ is not necessarily globally continuous. Corresponding to each interior node

$P_i$, there is a nodal basis function $\phi_i(x)$ satisfying $\phi_i(P_j) = \delta_{ij}, 1 \leq i \leq M, 1 \leq j \leq N$. So every $u_h \in (U_h)^2$ has an expression:

$$u_h(x) = \sum_{i=1}^{N} u_h(P_i)\phi_i(x),\ x \in \Omega. \tag{8.4.4}$$

Next we turn to construct the dual grid and the test function space. Let $P_1 \in K_{Q_1} \cap K_{Q_2}$ be an interior node, where $K_{Q_1} = \triangle A_1A_2A_3$ and $K_{Q_2} = \triangle A_1A_2A_4$. Connect $Q_1$ and $A_1$, $Q_1$ and $A_2$, $Q_2$ and $A_1$, and $Q_2$ and $A_2$ respectively to form a tetragon $K^*_{P_1} = \square A_1Q_2A_2Q_1A_1$ containing $P_1$ in its interior. This tetragon is called a dual element containing $P_1$ (cf. Fig. 8.4.1(a)). If $P_1$ is a boundary node, as shown in Fig. 8.4.1(b), then the dual grid containing $P_1$ is a triangle $K^*_{P_1} = \triangle A_1A_2Q_1$. The entire dual elements constitute a new grid $T^*_h = \{K^*_{P_i}\}$ of $\Omega$, referred to as a dual grid. The test function space related to $T^*_h$ is chosen as the piecewise constant function space on $T^*_h$, subject to the zero boundary condition. The nodal basis function $\psi_j$ $(1 \leq j \leq M)$ is the characteristic function of $K^*_{P_j}$. A function $v_h \in (V_h)^2$ can be expressed as

$$v_h(x) = \sum_{j=1}^{M} v_h(P_j)\psi_j(x),\ x \in \Omega. \tag{8.4.5}$$

As before, we use $\Pi^*_h$ to denote the interpolation projection operator from $U_h$ onto $V_h$: $\Pi^*_h\phi_i = \psi_i,\ 1 \leq i \leq M$.

We also design a subspace $W_h$ for the pressure $p_h$, which contains all the piecewise constant functions related to $T_h$, that is, $p_h \in W_h$ equals to a constant $p_h(Q_i)$ on each element $K_{Q_i}$ $(i = 1, 2, \cdots, \bar{N}$, $\bar{N}$ being the number of the elements).

Let $h_Q$ and $\rho_Q$ be the maximum length of the sides and the diameter of the inscribed circle of the element $K_Q$, respectively. We require the grid $T_h$ to be quasi-uniform, i.e., there are constants $\gamma_i > 0,\ i = 0, 1$, such that

$$\gamma_0 h \leq h_Q \leq \gamma_1 \rho_Q,\ \ \forall K_Q \in T_h. \tag{8.4.6}$$

The generalized difference scheme given below is nonconforming since $U_h \not\subset H^1_0(\Omega)$. For $u_h \in (U_h)^2$, $v_h \in (V_h)^2$ and $p_h, q_h \in W_h$,

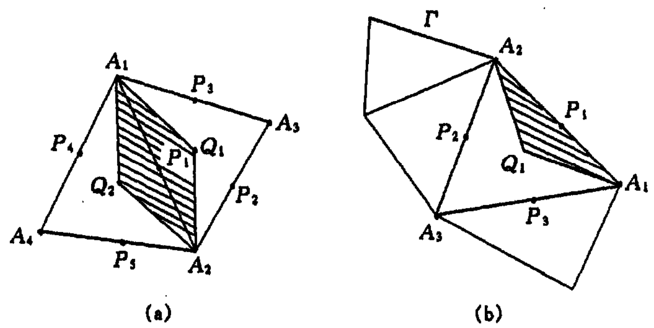



Fig. 8.4.1

set

$$\begin{aligned} a(u_h, v_h) &= -\mu \sum_{i=1}^{M} \int_{K_{P_i}^*} \Delta u_h \cdot v_h \mathrm{d}x \\ &= -\mu \sum_{i=1}^{M} \int_{\partial K_{P_i}^*} \frac{\partial u_h}{\partial \nu} \cdot v_h \mathrm{d}s \\ &= -\mu \sum_{i=1}^{M} v_h(P_i) \cdot \int_{\partial K_{P_i}^*} \frac{\partial u_h}{\partial \nu} \mathrm{d}s, \end{aligned} \tag{8.4.7}$$

$$b(v_h, p_h) = \sum_{i=1}^{M} v_h(P_i) \cdot \int_{\partial K_{P_i}^*} p_h \nu \mathrm{d}s, \tag{8.4.8}$$

$$c(u_h, q_h) = -\sum_{k=1}^{N} q_h(Q_k) \int_{K_{Q_k}} \mathrm{div} u_h \mathrm{d}x, \tag{8.4.9}$$

$$(f, v_h) = \sum_{i=1}^{M} v_h(P_i) \cdot \int_{K_{P_i}^*} f \mathrm{d}x. \tag{8.4.10}$$

Now, we are in a position to introduce a generalized difference method approximating the Stokes equation (8.4.2): Find $(u_h, p_h) \in (U_h)^2 \times W_h$ such that

$$\begin{cases} a(u_h, v_h) + b(v_h, p_h) = (f, v_h), & \forall v_h \in (V_h)^2, \quad (8.4.11a) \\ c(u_h, q_h) = 0, & \forall q_h \in W_h. \quad (8.4.11b) \end{cases}$$

Here the bilinear forms $a, b$ and $c$ are computed according to (8.4.7)-(8.4.9). For instance, we may take $v_h = \begin{pmatrix} \psi_j \\ 0 \end{pmatrix}$ or $v_h = \begin{pmatrix} 0 \\ \psi_j \end{pmatrix}$, and $q_h = \chi_k$ (the characteristic function of $K_{Q_k}$). (8.4.11) is a linear system of $(2M + \bar{N})$ equations with a (discretized) velocity field at the nodes $P_i$'s, and a pressure field at the barycenters $Q_k$'s. So this is a kind of alternative scheme. This system is symmetric since $b(\Pi_h^* u_h, p_h) = c(u_h, p_h)$ as we shall show below.

### 8.4.2 Convergence and error estimate

For $w_h \in (U_h)^2$, define $D_{x_i h} w_h$ $(i = 1, 2)$ as a piecewise function, identical to $\frac{\partial w_h}{\partial x_i}$ in the interior of each element $K_Q \in T_h$. Write

$$(\text{grad}_h u_h, \text{grad}_h w_h) = (D_{x_1 h} u_h, D_{x_1 h} w_h) + (D_{x_2 h} u_h, D_{x_2 h} w_h),$$

$$\|u_h\|_{1h}^2 = (u_h, u_h) + (\text{grad}_h u_h, \text{grad}_h u_h),$$

$$|u_h|_{1,h}^2 = (\text{grad}_h u_h, \text{grad}_h u_h).$$

**Lemma 8.4.1**

(i) *The seminorm $|u_h|_{1h}$ and the norm $\|u_h\|_{1h}$ are equivalent on $(U_h)^2$.*

(ii) $\|u_h\|_0^2 = \|\Pi_h^* u_h\|_0^2$

$$= \frac{1}{3} \sum_{K \in T_h} S_K(|u_h(P_1)|^2 + |u_h(P_2)|^2 + |u_h(P_3)|^2),$$

*where $S_K$ is the area of the element $K$ and $P_i$'s are the midpoints of the three sides of $K$.*

**Proof** (i) can be proved by the Poincaré inequality (cf. [B-86]). A direct integration of the area coordinate expression of the quadratic polynomial leads to (ii). □

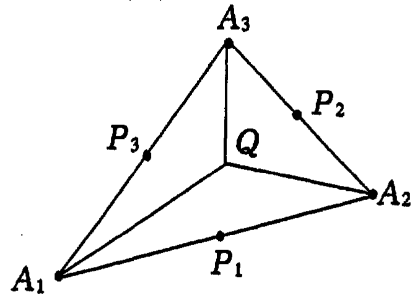



Fig. 8.4.2

**Lemma 8.4.2** *For $u_h, w_h \in (U_h)^2$ we have constants $C_1, C_2 > 0$ such that*

(i) $a(u_h, \Pi_h^* w_h) = a(w_h, \Pi_h^* u_h)$;

(ii) $|a(u_h, \Pi_h^* w_h)| \le C_1 \|u_h\|_{1h} |w_h|_{1h}$;

(iii) $a(u_h, \Pi_h^* u_h) \ge C_2 |u_h|_{1h}^2$.

**Proof** Consider an element $K = K_Q$. As in Fig. 8.4.2 we have

$$a(u_h, \Pi_h^* w_h) = \sum_{K \in T_h} I_K,$$

$$\begin{aligned} I_K = & -w_h(P_1)\Big(\int_{\overline{A_2QA_1}} \frac{\partial u_h}{\partial x_1} \mathrm{d}x_2 - \int_{\overline{A_2QA_1}} \frac{\partial u_h}{\partial x_2} \mathrm{d}x_1\Big) \\ & -w_h(P_2)\Big(\int_{\overline{A_3QA_2}} \frac{\partial u_h}{\partial x_1} \mathrm{d}x_2 - \int_{\overline{A_3QA_2}} \frac{\partial u_h}{\partial x_2} \mathrm{d}x_1\Big) \\ & -w_h(P_3)\Big(\int_{\overline{A_1QA_3}} \frac{\partial u_h}{\partial x_1} \mathrm{d}x_2 - \int_{\overline{A_1QA_3}} \frac{\partial u_h}{\partial x_2} \mathrm{d}x_1\Big) \\ = & -\frac{\partial u_h}{\partial x_1}[w_h(P_1)(x_2(A_1) - x_2(A_2)) \\ & + w_h(P_2)(x_2(A_2) - x_2(A_3)) + w_h(P_3)(x_2(A_3) - x_2(A_1))] \end{aligned}$$

$$+\frac{\partial u_h}{\partial x_2}[w_h(P_1)(x_1(A_1)-x_1(A_2))$$
$$+w_h(P_2)(x_1(A_2)-x_1(A_3))+w_h(P_3)(x_1(A_3)-x_1(A_1))]. \tag{8.4.12}$$

Let $(\lambda_1, \lambda_2, \lambda_3)$ be the coordinates of the barycenter of $K$, and $\mu_i = \lambda_i + \lambda_{i+1} - \lambda_{i+2}$ for $i = 1,2,3$ ($\lambda_4 = \lambda_1, \lambda_5 = \lambda_2$). Then we have $\mu_i(P_j) = \delta_{ij}$ and

$$w_h|_K = w_h(P_1)\mu_1 + w_h(P_2)\mu_2 + w_h(P_3)\mu_3, \tag{8.4.13}$$

$$\begin{cases} \dfrac{\partial \lambda_i}{\partial x_1} = \dfrac{x_2(A_{i+1}) - x_2(A_{i+2})}{2S_K}, \\ \dfrac{\partial \lambda_i}{\partial x_2} = \dfrac{x_1(A_{i+1}) - x_1(A_{i+2})}{2S_K}, \end{cases} \tag{8.4.14}$$

where $i = 1,2,3$ and $A_4 = A_1$, $A_5 = A_2$. Substituting (8.4.13) and (8.4.14) into (8.4.12) yields

$$I_K = \int_K \Big(\frac{\partial u_h}{\partial x_1}\frac{\partial w_h}{\partial x_1} + \frac{\partial u_h}{\partial x_2}\frac{\partial w_h}{\partial x_2}\Big)\mathrm{d}x.$$

Thus

$$a(u_h, \Pi_h^* w_h) = \sum_{K\in T_h}\int_K \mathrm{grad}u_h \cdot \mathrm{grad}w_h \mathrm{d}x = (\mathrm{grad}_h u_h, \mathrm{grad}_h w_h).$$

This implies (i), (ii), (iii) and completes the proof. □

The following Lemma can be proved in like manner.

**Lemma 8.4.3** *For $u_h \in (U_h)^2$ and $p_h \in W_h$ we have ($C_3 > 0$ is a constant)*

(i) $b(\Pi_h^* u_h, p_h) = c(u_h, p_h)$, (8.4.15)

(ii) $|b(\Pi_h^* u_h, p_h)| \le C_3 |u_h|_{1h} |p_h|_0$. (8.4.16)

For $u \in (H_0^1(\Omega))^2$, define its projection onto $(U_h)^2$ (cf. Fig. 8.4.2) by

$$\hat{\Pi}_h u = \hat{u}(P_1)\mu_1 + \hat{u}(P_2)\mu_2 + \hat{u}(P_3)\mu_3, \text{ in each } K \in T_h,$$

$$\hat{u}(P_i) = \frac{1}{|\overline{A_iA_{i+1}}|}\int_{\overline{A_iA_{i+1}}} u(x)\mathrm{d}s, \ i = 1,2,3. \ (A_4 = A_1.)$$

Obviously $\hat{\Pi}_h u|_K = u, \ \forall K \in T_h$ when $u \in (U_h)^2$. So it follows from the interpolation approximation property that

$$|\hat{\Pi}_h u|_{1h} \le C_4 |u|_1.$$

**Lemma 8.4.4**

(i) *It holds for* $u \in (H_0^1(\Omega) \cap C^2(\overline{\Omega}))^2$, *and* $u_h \in (U_h)^2$ *that*

$$|a(u - \hat{\Pi}_h u, \Pi_h^* u_h)| \le C_5 h |u_h|_{1h} |\mathrm{D}^2 u|_{\max}. \tag{8.4.17}$$

(ii) *For* $p \in C^1(\overline{\Omega})$, $u_h \in (U_h)^2$ *and* $p_h \in W_h$ *we have*

$$|b(\Pi_h^* u_h, p - p_h)| \le C_6 \max_{\Omega} |p - p_h| |u_h|_{1h}. \tag{8.4.18}$$

(iii) *If* $u \in (H_0^1(\Omega))^2$, *then*

$$c(u - \hat{\Pi}_h u, q_h) = 0, \ \forall q_h \in W_h. \tag{8.4.19}$$

*Here* $|\mathrm{D}^2 u|_{\max}$ *stands for the maximum norm of the second partial derivatives of* $u$. *Moreover, the equality (8.4.19) is equivalent to*

$$\sum_{k=1}^{\tilde{N}} \int_{K_{Q_k}} q_h \mathrm{div}(u - \hat{\Pi}_h u)\mathrm{d}x = 0. \tag{8.4.20}$$

**Proof** (8.4.17) and (8.4.18) are direct consequences of (8.4.7), (8.4.8), and the expressions of $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$. To show (8.4.19), we denote by $\nu_1, \nu_2$ and $\nu_3$ the unit outer normal directions of $\overline{A_1A_2}$, $\overline{A_2A_3}$ and $\overline{A_3A_1}$ respectively, then

$$\begin{aligned} c(u - \hat{\Pi}_h u, q_h) &= -\sum_{k=1}^{\tilde{N}} q_h(Q_k) \sum_{i=1}^{3} \nu_i \int_{\overline{A_iA_{i+1}}} (u - \hat{\Pi}_h u)\mathrm{d}s \\ &= \sum_{k=1}^{\tilde{N}} q_h(Q_k) \sum_{i=1}^{3} \nu_i \Big[ \int_{\overline{A_iA_{i+1}}} u \mathrm{d}s - |\overline{A_iA_{i+1}}| (\hat{\Pi}_h u)(P_i) \Big] = 0. \end{aligned}$$

Finally, the equivalence of (8.4.20) and (8.4.19) follows from (8.4.9) and the fact $\mathrm{div}\hat{\Pi}_h u \in W_h$. This completes the proof. □

It results from (8.4.15) and (8.4.20) that there is a constant $\gamma > 0$ satisfying

$$\sup \frac{b(\Pi_h^* u_h, q_h)}{|u_h|_{1h}} \geq \gamma \|q_h\|_0, \quad \forall q_h \in W_h. \tag{8.4.21}$$

Making use of Lemmas 8.4.1-4, one may easily prove (cf. [B-34,86]) the following: If $(u,p) \in (C^2(\overline{\Omega}))^2 \cap (H_0^1(\Omega))^2 \times C^1(\overline{\Omega}) \cap L_0^2(\Omega)$ is the solution to (8.4.3), then, for sufficiently small $h > 0$, the nonconforming generalized difference equation (8.4.11) has a unique solution $(u_h, p_h) \in (U_h)^2 \times W_h$, and

$$|u - u_h|_{1h} + \|p - p_h\|_0 \leq Ch(|\mathrm{D}^2 u|_{\max} + |\mathrm{D}p|_{\max}),$$

where $|\mathrm{D}p|_{\max}$ denotes the maximum norm of the first derivative of $p$.

### 8.4.3 A numerical example

In the Stokes problem (8.4.2), take $\Omega = [0,1] \times [0,1]$, $\mu = 1$, $f = (f_1, f_2)$:

$$f_1(x,y) = -6x^2(x-1)^2(2y-1),$$

$$f_2(x,y) = 4x(x-1)(2x-1)[(2x-1)^2 + 2y(y-1)] - f_1(y,x).$$

Its true solution is

$$u_1(x,y) = x^2(x-1)^2 y(y-1)(2y-1),$$

$$u_2(x,y) = -u_1(y,x),$$

$$p(x,y) = 2x(x-1)(2x-1)y(y-1)(2y-1).$$

Divide $\Omega$ into thirty-six small squares with side size $\frac{1}{6}$. Then we use the diagonal lines at an angle of $\pi/4$ to the $x$-axis to further divide each small square into two triangles, ending up with a triangulation $T_h$. The errors of the approximation is given below:

$$\max_j |u_h(P_j) - u(P_j)| = 0.31616975 \times 10^{-2},$$

$$\min_j |u_h(P_j) - u(P_j)| = 0.80926718 \times 10^{-5},$$

$$\max_Q |p_h(Q) - p(Q)| = 0.14746220 \times 10^{-1},$$

$$\min_Q |p_h(Q) - p(Q)| = 0.56510426 \times 10^{-3}.$$

## 8.5 Coupled Sound-Heat Problems

The following hyperbolic-parabolic system describes the flow of compressible fluid with heat transfer:

$$\begin{cases} \dfrac{\partial u}{\partial t} = c\dfrac{\partial}{\partial x}(w - (\gamma - 1)e), \\ \dfrac{\partial w}{\partial t} = c\dfrac{\partial u}{\partial x}, & 0 \le x \le 1, 0 < t \le T, \\ \dfrac{\partial e}{\partial t} = \sigma\dfrac{\partial^2 e}{\partial x^2} - c\dfrac{\partial u}{\partial x}, \end{cases} \tag{8.5.1}$$

where $c$ and $\sigma$ are positive constants, and $\gamma > 1$. The initial values

$$u(x,0) = f(x),\ w(x,0) = g(x),\ e(x,0) = h(x)$$

are 1-periodic functions, and the boundary condition is a 1-periodic boundary condition. Write $I = [0,1]$ and $V = \{v \in H^1(I) : v(0) = v(1)\}$. Then, a weak form of (8.5.1) is: Find $u, w, e \in V$ such that for all $v \in V$

$$\begin{cases} (u_t, v) - c(w_x, v) + c(\gamma - 1)(e_x, v) = 0, \\ (w_t, v) - c(u_x, v) = 0, \\ (e_t, v) + c(u_x, v) + \sigma(e_x, v_x) = 0. \end{cases} \tag{8.5.2}$$

Decompose $I = [0,1]$ by $0 = x_0 < x_1 < \cdots < x_N = 1$, and set $I_i = [x_{i-1}, x_i]$, $h_i = x_i - x_{i-1}$, then we have a grid $T_h = \{I_i\}$. The dual grid $T_h^* = \{I_i^*\}$, where $I_i^* = [x_{i-1/2}, x_{i+1/2}]$, $x_{i-1/2} = \frac{1}{2}(x_{i-1} + x_i)$, for $i = 1, 2, \cdots, N-1$, and $I_0^* = [0, x_{1/2}]$, $I_N^* = [x_{N-1/2}, x_N]$. The trial function space $U_h$ related to $T_h$ consists of the piecewise linear continuous functions with period 1. For $i = 1, 2, \cdots, N-1$, the nodal basis functions $\phi_i(x)$'s are defined by

$$\phi_i(x_j) = \delta_{ij},\ 0 \le j \le N,$$

and $\phi_N(x)$ is defined by

$$\phi_N(x_0) = 1,\ \phi_N(x_j) = \delta_{Nj},\ 1 \le j \le N.$$

The test function space $V_h$ is composed of the piecewise constant periodic function on $T_h^*$ with the nodal basis function $\psi_j(x)$ ($1 \le j \le N$) being the characteristic function on $I_j^*$. Any $u_h, w_h, e_h \in U_h$ can be expressed as

$$u_h(x,t) = \sum_{i=1}^{N} u_i(t)\phi_i(x),$$

$$w_h(x,t) = \sum_{i=1}^{N} w_i(t)\phi_i(x),$$

$$e_h(x,t) = \sum_{i=1}^{N} e_i(t)\phi_i(x).$$

Now, a generalized difference scheme approximating (8.5.2) is: Find $u_h, w_h, e_h \in U_h$ such that for $1 \le j \le N$

$$\begin{cases} (u_{ht}, \psi_j) - c(w_{hx}, \psi_j) + c(\gamma - 1)(e_{hx}, \psi_j) = 0, \\ (w_{ht}, \psi_j) - c(u_{hx}, \psi_j) = 0, \\ (e_{ht}, \psi_j) + c(u_{hx}, \psi_j) + \sigma(e_{hx}, \psi_{jx}) = 0. \end{cases} \tag{8.5.3}$$

Write $v_h(t) = (u_1(t), \cdots, u_N(t), w_1(t), \cdots, w_N(t), e_1(t), \cdots, e_N(t))^T$, then one can rewrite (8.5.3) as

$$M\frac{\mathrm{d}v_h(t)}{\mathrm{d}t} - Kv_h(t) = 0, \tag{8.5.4}$$

where

$$M = \begin{bmatrix} M_1 & & \\ & M_1 & \\ & & M_1 \end{bmatrix},\ K = \begin{bmatrix} 0 & cK_1 & -c(\gamma-1)K_1 \\ cK_1 & 0 & 0 \\ -cK_1 & 0 & \sigma K_2 \end{bmatrix}$$

are $(3N \times 3N)$ matrices, and $M_1, K_1, K_2$ are $(N \times N)$ matrices: The entry at the $j$-th row and $i$-th column of $M_1$ is $(\phi_i, \phi_j)$, the one of $K_1$

is $(\phi_{ix}, \psi_j)$, and the one of $K_2$ is $(\phi_{ix}, \psi_{jx})$ (in the sense of generalized functions). The initial value is

$$v_h(0) = (f(x_1), \cdots, f(x_N), g(x_1), \cdots, g(x_N), h(x_1), \cdots, h(x_N))^T.$$

(8.5.4) is a system of ordinary differential equations. Various kinds of numerical methods for ordinary differential equations can be used here for a further discretization. For example, the modified Euler's method gives a fully-discretized generalized difference scheme:

$$M\frac{v_h^{n+1} - v_h^n}{\tau} = K\frac{v_h^{n+1} + v_h^n}{2}. \tag{8.5.5}$$

This is a six-point symmetric scheme (Crank-Nicolson scheme). If the solution to (8.1.1) is smooth, and the step size $h_i = h$, then the truncation error of (8.5.5) is $O(\tau^2 + h^2)$. In this case (8.5.5) can be written as

$$\begin{aligned} & u_{j-1}^{n+1} + 6u_j^{n+1} + u_{j+1}^{n+1} + \frac{2c\tau}{h}(w_{j-1}^{n+1} - w_{j+1}^{n+1}) \\ & -\frac{2c(\gamma-1)\tau}{h}(e_{j-1}^{n+1} - e_{j+1}^{n+1}) \\ = \; & u_{j-1}^n + 6u_j^n + u_{j+1}^n + \frac{2c\tau}{h}(w_{j+1}^n - w_{j-1}^n) \\ & -\frac{2c(\gamma-1)\tau}{h}(e_{j+1}^n - e_{j-1}^n), \end{aligned} \tag{8.5.6a}$$

$$\begin{aligned} & \frac{2c\tau}{h}(u_{j-1}^{n+1} - u_{j+1}^{n+1}) + w_{j-1}^{n+1} + 6w_j^{n+1} + w_{j+1}^{n+1} \\ = \; & \frac{2c\tau}{h}(u_{j+1}^n - u_{j-1}^n) + w_{j-1}^n + 6w_j^n + w_{j+1}^n, \end{aligned} \tag{8.5.6b}$$

$$\begin{aligned} & \frac{2c\tau}{h}(u_{j+1}^{n+1} - u_{j-1}^{n+1}) + \left(1 - \frac{4\sigma\tau}{h^2}\right)e_{j-1}^{n+1} \\ & +\left(6 + \frac{8\sigma\tau}{h^2}\right)e_j^{n+1} + \left(1 - \frac{4\sigma\tau}{h^2}\right)e_{j+1}^{n+1} \\ = \; & \frac{2c\tau}{h}(u_{j-1}^n - u_{j+1}^n) + \left(1 + \frac{4\sigma\tau}{h^2}\right)e_{j-1}^n \\ & +\left(6 - \frac{8\sigma\tau}{h^2}\right)e_j^n + \left(1 + \frac{4\sigma\tau}{h^2}\right)e_{j+1}^n, \end{aligned} \tag{8.5.6c}$$

where

$$u_k = u_{N+k},\ w_k = w_{N+k},\ e_k = e_{N+k},\ k = 0, 1, \cdots.$$

If we use Euler's method in place of (8.5.6a,b), while keeping (8.5.6c) as it is, then (8.5.6a,b) are replaced by

$$\begin{aligned} & u_{j-1}^{n+1} + 6u_j^{n+1} + u_{j+1}^{n+1} \\ = \ & u_{j-1}^n + 6u_j^n + u_{j+1}^n - \frac{4c\tau}{h}(w_{j-1}^n - w_{j+1}^n) \\ & -\frac{4c\tau(\gamma-1)}{h}(e_{j+1}^n - e_{j-1}^n), \end{aligned} \tag{8.5.6a$'$}$$

$$\begin{aligned} & w_{j-1}^{n+1} + 6w_j^{n+1} + w_{j+1}^{n+1} \\ = \ & -\frac{4c\tau}{h}(u_{j+1}^n - u_{j-1}^n) + w_{j-1}^n + 6w_j^n + w_{j+1}^n. \end{aligned} \tag{8.5.6b$'$}$$

Its truncation error is $O(\tau+h^2)$. This scheme is implicit, but (8.5.6a, b)$'$ are both triple diagonal matrices for the unknowns $\{u_j^{n+1}\}$ and $\{w_j^{n+1}\}$ respectively. If we work out $\{u_j^{n+1}\}$ and insert it into (8.5.6c), then (8.5.6c) also becomes a triple diagonal matrix for $\{e_j^{n+1}\}$, which is easy to solve numerically.

Next we try to investigate the stability of (8.5.6) as with the usual difference methods. By the separation of variables it is easy to deduce the amplification matrix of (8.5.6):

$$G(\xi,h) = (a^2d - b^2d - (\gamma-1)ab^2)^{-1}\cdot$$
$$\begin{bmatrix} a^2d + b^2d + (\gamma-1)ab^2 & 2abcd & -2(\gamma-1)a^2b \\ 2abd & a^2d + b^2d - (\gamma-1)ab^2 & -2(\gamma-1)ab^2 \\ -2a^2b & -2ab^2 & (\gamma-1)ab^2 + a^2l - b^2l \end{bmatrix}$$

where

$$a = 2(3+\cos\xi),\ b = \mathrm{i}\frac{4rhc}{\sigma}\sin\xi,\ (\mathrm{i} = \sqrt{-1})$$

$$d = 2(3+4r) + 2(1-4r)\cos\xi = 4\Big(2 - (1-4r)\sin^2\frac{\xi}{2}\Big),$$

$$l = 2(3-4r) + 2(1+4r)\cos\xi = 4\Big(2 - (1+4r)\sin^2\frac{\xi}{2}\Big),$$

$$r = \frac{\sigma\tau}{h^2} = \text{constant}.$$

Since $a \geq 4$, $d \geq 4$ and $b^2 = O(h^2) = O(\tau)$, we have

$$\begin{aligned} G(\xi, h) &= \begin{bmatrix} 1 & \dfrac{2b}{a} & \dfrac{-2(\gamma-1)b}{d} \\ \dfrac{2b}{a} & 1 & 0 \\ -\dfrac{2b}{d} & 0 & \dfrac{l}{d} \end{bmatrix} + O(\tau) \\ &= G_0(\xi, h) + O(\tau). \end{aligned}$$

By [A-27] and [B-74], the matrix family $\{G^n(\xi,h)\}$ $(0 < \tau \leq \tau_0, |\xi| \leq \pi, 0 < n\tau \leq T)$ is uniformly bounded, if and only if $\{G_0^n(\xi,h)\}$ $(0 < \tau \leq \tau_0, |\xi| \leq \pi, 0 < n\tau \leq T)$ is uniformly bounded. Similarity transformations are employed in [A-56] to show the uniform boundedness of $\{G_0^n(\xi,h)\}$, and hence scheme (8.5.6) is stable for any grid ratio $r > 0$. This together with the consistency of the scheme (truncation error $O(\tau^2 + h^2)$) guarantees the convergence of $(u_h, w_h, e_h)$ to $(u, w, e)$ with an error estimate:

$$\max_{0\leq t\leq T} \|u(\cdot,t) - u_h(\cdot,t)\|_0 + \max_{0\leq t\leq T} \|w(\cdot,t) - w_h(\cdot,t)\|_0$$

$$+ \max_{0\leq t\leq T} \|e(\cdot,t) - e_h(\cdot,t)\|_0 \leq C(\tau^2 + h^2).$$

(cf. [A-56] and [B-74] for details.)

Similarly one can show that the scheme (8.5.6a)′, (8.5.6b)′, (8.5.6c) is absolutely stable, and its convergence order is $O(\tau + h^2)$.

**Remark** This section considers only a linear element generalized difference scheme for one-dimensional problems. In principle, high order element schemes can be constructed for triangulations in higher dimensions. But we point out that the discussions of this section on convergence and stability are analogous to that of standard finite difference methods, which can not be extended directly to arbitrary triangulations and high order elements. The equation (8.5.1) is a coupled hyperbolic-parabolic system. It remains to be an open question how to extend the methods in Chapters 5 and 6 to study the convergence and the error estimate of the generalized difference scheme approximating (8.5.1).

## 8.6 Regularized Long Wave Equations

This section is devoted to the generalized difference solutions of the following initial and boundary values problem of a regularized long wave equation:

$$\begin{cases} \dfrac{\partial u}{\partial t} - \gamma \dfrac{\partial^2}{\partial x^2}\dfrac{\partial u}{\partial t} = \dfrac{\partial f(u)}{\partial x}, & a < x < b,\ 0 < t \le T, \quad (8.6.1a) \\ u(a,t) = u(b,t) = 0, & 0 < t \le T, \quad (8.6.1b) \\ u(x,0) = u_0(x), & a < x < b, \quad (8.6.1c) \end{cases}$$

where $f(u) = \alpha u + \frac{1}{2}\beta u^2$; $\gamma > 0$ and $\alpha$, $\beta$ are given constants; $u_0(x) \in C^2(I)$, $I = [a,b]$, $u_0(a) = u_0(b) = 0$. It is confirmed in [B-7] that problem (8.6.1) has a unique solution $u(x,t) \in C^1([0,T]; C^2(I))$. Define

$$a(u,v) = \Big(\gamma\frac{\partial u}{\partial x}, \frac{\partial v}{\partial x}\Big) + (u,v) = \gamma\int_a^b u_x v_x \mathrm{d}x + \int_a^b uv\mathrm{d}x. \quad (8.6.2)$$

Then a weak form of (8.6.1) reads: Find $u(x,t) \in C^1([0,T]; H_0^1(I))$ such that

$$\begin{cases} a\Big(\dfrac{\partial u}{\partial t}, v\Big) = \Big(\dfrac{\partial f(u)}{\partial x}, v\Big), & \forall v \in H_0^1(I), \quad (8.6.3a) \\ (u(x,0), v) = (u_0(x), v), & \forall v \in H_0^1(I). \quad (8.6.3b) \end{cases}$$

Next we shall present a Lagrange quadratic generalized difference scheme for (8.6.1) (or (8.6.3)).

### 8.6.1 Semi-discrete generalized difference schemes

Place a grid $T_h = \{I_i\}_1^N$ on $I = [a,b]$, where $I_i = [x_{i-1}, x_i]$, $a = x_0 < x_1 < \cdots < x_N = b$. Set $h_i = x_i - x_{i-1}$, $h = \max h_i$. Piecewise quadratic polynomials vanishing on the boundary $x = a$ and $b$ are chosen as the trial function space $U_h \subset H_0^1(I)$. On each $I_i$, a function $p \in U_h$ is determined uniquely by its values at three interpolation nodes $x_{j-1}, x_{j-1/2}$ and $x_j$. The nodal basis functions are (cf. §2.3)

$$\phi_j(x) = \begin{cases} 2h_j^{-2}(x - x_{j-1})(x - x_{j-1/2}), & x_{j-1} \le x \le x_j, \\ 2h_{j+1}^{-2}(x - x_{j+1})(x - x_{j+1/2}), & x_j \le x \le x_{j+1}, \\ 0, & \text{elsewhere.} \end{cases}$$

$$\phi_{j-1/2}(x) = \begin{cases} 4h_j^{-2}(x - x_{j-1})(x_j - x), & x_{j-1} \le x \le x_j, \\ 0, & \text{elsewhere.} \end{cases}$$

The dual grid $T_h^* = \{I_j^*, I_{j-1/2}^*\}$, where $I_j^* = [x_{j-1/4}, x_{j+1/4}]$, $I_{j-1/2}^* = [x_{j-3/4}, x_{j-1/4}]$. The corresponding test function space $V_h$ contains piecewise constant functions, which vanish on $I_0^* = [a, x_{1/4}]$ and $I_N^* = [x_{N-1/4}, b]$. The nodal basis functions $\psi_j$ and $\psi_{j-1/2}$ are the characteristic functions of $I_j^*$ and $I_{j-1/2}^*$ respectively. The semi-discrete generalized difference scheme approximating (8.6.1) is: Find

$$u_h(x,t) = \sum_{j=1}^{N-1} u_j(t)\phi_j(x) + \sum_{j=1}^{N} u_{j-1/2}(t)\phi_{j-1/2}(x) \in U_h$$

such that

$$\begin{cases} a\Big(\dfrac{\partial u_h}{\partial t}, v_h\Big) = \Big(\dfrac{\partial f(u_h)}{\partial x}, v_h\Big), & \forall v_h \in V_h, 0 < t \le T, \quad (8.6.4a) \\ u_h(x,0) = u_{0h}(x), & a \le x \le b. \quad (8.6.4b) \end{cases}$$

Here $u_{0h}$ is a certain approximation of $u_0(x)$, generally taken as the interpolation projection of $u_0(x)$ onto $U_h$. Now the above bilinear form $a(\frac{\partial u_h}{\partial t}, v_h)$ should be understood in the sense of generalized functions. (8.6.4) is an initial value problem of a system of ordinary differential equations.

We observe that $u_j(t) = u_h(x_j, t)$ and $u_{j-1/2}(t) = u_h(x_{j-1/2}, t)$. Denote by $\Pi_h^*$ the interpolation projection from $U_h$ onto $V_h$. Then, for any $u_h \in U_h$

$$\Pi_h^* u_h = \sum_{j=1}^{N-1} u_j \psi_j + \sum_{j=1}^{N} u_{j-1/2}\psi_{j-1/2}.$$

Let us introduce the following discrete norms:

$$|u_h|_{0h}^2 = |\Pi_h^* u_h|_0^2 = \sum_{j=1}^{N} \frac{h_j}{4}(u_{j-1}^2 + 2u_{j-1/2}^2 + u_j^2),$$

$$|u_h|_{1h}^2 = \sum_{j=1}^{N} \frac{h_j}{2}\Big[\Big(\frac{u_{j-1/2} - u_{j-1}}{h_j/2}\Big)^2 + \Big(\frac{u_j - u_{j-1/2}}{h_j/2}\Big)^2\Big].$$

As in §2.3, it is easy to show that

(i) On $U_h$, $|\cdot|_{0h}$ and $|\cdot|_{1h}$ are equivalent to $|\cdot|_0$ and $|\cdot|_1$ respectively;

(ii) $|a(u_h, \Pi_h^* w_h)| \leq M\|u_h\|_1\|w_h\|_1, \ \forall w_h, u_h \in U_h$;

(iii) $a(u_h, \Pi_h^* u_h) > \alpha\|u_h\|_1^2, \ \forall u_h \in U_h$,

where $M > 0$ and $\alpha > 0$ are constants. By virtue of these observations, the following statement holds:

(iv) If $u \in H_0^1(I) \cap H^3(I)$ is the solution to

$$a(u, v) = (g, v), \quad \forall v \in H_0^1(I),$$

and $u_h \in U_h$ is the solution to

$$a(u_h, v_h) = (g, v_h), \quad \forall v_h \in V_h,$$

then we have the following estimate

$$\|u - u_h\|_1 \leq C_1 h^2 |u|_3. \tag{8.6.5}$$

Here and below, $C_j$'s $(1 \leq j \leq 6)$ denote constants independent of $h$. Now let us consider the unique-solvability and the convergence of the semi-discrete scheme (8.6.4). Define

$$\underline{m} = \inf_{\substack{a\leq x\leq b\\ 0\leq t\leq T}} u(x,t), \ \overline{m} = \sup_{\substack{a\leq x\leq b\\ 0\leq t\leq T}} u(x,t).$$

Take a bounded truncation function $\hat{f}(v) \in C^2(\mathrm{R})$ of $f(v)$ satisfying ($\epsilon > 0$ is a constant)

$$\hat{f}(v) = f(v), \ \ \forall v \in [\underline{m} - \epsilon, \overline{m} + \epsilon],$$

$$\sup_{v\in\mathrm{R}}\{|\hat{f}(v)|, |\hat{f}'(v)|, |\hat{f}''(v)|\} = d_\epsilon < \infty.$$

Replacing $f$ by $\hat{f}$ yields a continuous problem related to (8.6.1)

$$\begin{cases} \dfrac{\partial \hat{u}}{\partial t} - \gamma \dfrac{\partial^2}{\partial x^2}\dfrac{\partial \hat{u}}{\partial t} = \dfrac{\partial \hat{f}(\hat{u})}{\partial x}, & a < x < b,\ 0 < t \le T, \quad (8.6.6a) \\ \hat{u}(a,t) = \hat{u}(b,t) = 0, & 0 < t \le T, \quad (8.6.6b) \\ \hat{u}(x,0) = u_0(x), & a < x < b, \quad (8.6.6c) \end{cases}$$

and a semi-discrete problem

$$\begin{cases} a\Big(\dfrac{\partial \hat{u}_h}{\partial t}, v_h\Big) = \Big(\dfrac{\partial \hat{f}(\hat{u}_h)}{\partial x}, v_h\Big), & \forall v_h \in V_h, 0 < t \le T, \quad (8.6.7a) \\ \hat{u}_h(x,0) = u_{0h}(x), & a \le x \le b. \quad (8.6.7b) \end{cases}$$

**Lemma 8.6.1** *The unique solution $u(x,t)$ of problem (8.6.1) is identical to the unique solution of problem (8.6.6).*

**Proof** The solution $u(x,t)$ of (8.6.1) obviously solves (8.6.6). To show the uniqueness, let $\hat{u}(x,t)$ be any solution to (8.6.6). Then

$$\frac{\partial(u-\hat{u})}{\partial t} - \gamma \frac{\partial^2}{\partial x^2}\frac{\partial(u-\hat{u})}{\partial t} = \frac{\partial \hat{f}(u)}{\partial x} - \frac{\partial \hat{f}(\hat{u})}{\partial x}.$$

Multiply it by $(u-\hat{u})$ and then integrate it for $x$ to obtain

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}(|u-\hat{u}|_0^2 + \gamma|u-\hat{u}|_1^2) \le d_\epsilon |u-\hat{u}|_0 |u-\hat{u}|_1,$$

and hence

$$\frac{\mathrm{d}}{\mathrm{d}t}(|u-\hat{u}|_0^2 + \gamma|u-\hat{u}|_1^2) \le d'_\epsilon(|u-\hat{u}|_0^2 + \gamma|u-\hat{u}|_1^2),$$

where $d'_\epsilon = d_\epsilon \max\{1, \frac{1}{\gamma}\}$. Solving the last inequality gives

$$\begin{aligned} & |u(\cdot,t) - \hat{u}(\cdot,t)|_0^2 + \gamma|u(\cdot,t) - \hat{u}(\cdot,t)|_1^2 \\ \le\ & \exp(d'_\epsilon T)(|u(\cdot,0) - \hat{u}(\cdot,0)|_0^2 + \gamma|u(\cdot,0) - \hat{u}(\cdot,0)|_1^2) = 0, \end{aligned}$$

which implies $u \equiv \hat{u}$. This completes the proof. □

**Lemma 8.6.2** *Problem (8.6.7) has a unique solution $\hat{u}_h(\cdot,t) \in U_h$ defined on* $[0,T]$.

**Proof** By the existence and uniqueness theorem of the solution to ordinary differential equations, equation (8.6.7) has a unique solution $\hat{u}_h(\cdot,t)$ at least on a right-hand neighborhood of $t = 0$. According to the continuation theorem, to extend the solution $\hat{u}_h(\cdot,t)$ to the entire $[0,T]$ we only have to show the boundedness of $\|\hat{u}_h(\cdot,t)\|_1$ for all such $t \in [0,T]$ where $\hat{u}_h(\cdot,t)$ is well-defined. So we integrate (8.6.7a) with respect to $t$ and make use of (8.6.7b) to get

$$a(\hat{u}_h, v_h) = a(u_{0h}, v_h) + \int_0^t \Big(\frac{\partial \hat{f}(\hat{u}_h(\cdot,s))}{\partial x}, v_h(\cdot)\Big) \mathrm{d}s.$$

Choose $v_h = \Pi_h^* \hat{u}_h$, then it follows from (i), (ii) and (iii) that

$$\|\hat{u}_h(\cdot,t)\|_1 \leq \frac{M}{\alpha}\|u_{0h}\|_1 + \frac{d_\epsilon}{\alpha C_2}\int_0^t \|\hat{u}_h(\cdot,s)\|_1 \mathrm{d}s.$$

By virtue of Gronwall's inequality

$$\|\hat{u}_h(\cdot,t)\|_1 \leq \frac{M}{\alpha}\exp\left(d_\epsilon'' T\right)\|u_{0h}\|_1.$$

Here $d_\epsilon'' = d_\epsilon/(\alpha C_2)$.

After these preparations, we are able to use a method similar to that in §5.1 to obtain

$$\begin{aligned} & \|u(\cdot,t) - \hat{u}_h(\cdot,t)\|_1 \\ \leq\ & C_3\Big[\|u_0 - u_{0h}\|_1 + h^2(|u_0|_3 + \int_0^t \Big|\frac{\partial u(\cdot,s)}{\partial s}\Big|_3 \mathrm{d}s\Big]. \end{aligned}$$

Finally, it should be pointed out that, for sufficiently small $h$, the above $\hat{u}_h(x,t)$ is precisely the solution $u_h(x,t)$ of the difference equation (8.6.4). In fact, by the imbedding theorem and the above estimate we have

$$\begin{aligned} & \sup_{\substack{a\leq x\leq b \\ 0\leq t\leq T}} |u(x,t) - \hat{u}_h(x,t)| \\ \leq\ & C_4 \sup_{0\leq t\leq T} \|u(\cdot,t) - \hat{u}_h(\cdot,t)\|_1 \to 0, \text{ as } h \to 0. \end{aligned}$$

Thus for $h > 0$ sufficiently small

$$\underline{m} - \epsilon \le \hat{u}_h(x,t) \le \overline{m} + \epsilon,\ a \le x \le b,\ 0 \le t \le T.$$

Consequently

$$\hat{f}(\hat{u}_h(x,t)) = f(u_h(x,t)).$$

This implies $\hat{u}_h(x,t) = u_h(x,t)$.

To sum up, we have

**Theorem 8.6.1** *If the solution $u(x,t)$ of the problem (8.6.1) satisfies $u(x,t) \in C^1([0,T]; H^3(I))$, then for sufficiently small $h > 0$, the semi-discrete scheme (8.6.4) possesses a unique solution $u_h(\cdot,t) \in U_h$ defined on $[0,T]$, satisfying the following error estimate for $0 \le t \le T$:*

$$\begin{aligned} &\|u(\cdot,t) - u_h(\cdot,t)\|_1 \\ \le\ & C_5\Big[\|u_0 - u_{0h}\|_1 + h^2\Big(|u_0|_3 + \int_0^t \Big|\frac{\partial u(\cdot,s)}{\partial s}\Big|_3 \mathrm{d}s\Big)\Big]. \end{aligned} \tag{8.6.8}$$

### 8.6.2 Fully-discrete generalized difference schemes

Take the time step size $\tau = T/M$ and $t_k = k\tau$ $(k = 0,1,\cdots,M)$. A fully-discrete generalized difference scheme for (8.6.1) is: Find $u_h^k \in U_h$ such that

$$\begin{cases} a(\bar{\partial}_t u_h^k, v_h) = \dfrac{1}{2}\Big(\dfrac{\partial f(u_h^k)}{\partial x} + \dfrac{\partial f(u_h^{k-1})}{\partial x}, v_h\Big),\ v_h \in V_h, & \text{(8.6.9a)} \\ \qquad k = 1,2,\cdots,M, & \\ u_h^0 = u_{0h}, & \text{(8.6.9b)} \end{cases}$$

where $\bar{\partial}_t u_h^k = (u_h^k - u_h^{k-1})/\tau$. As before, we also consider the auxiliary problem

$$\begin{cases} a(\bar{\partial}_t u_h^k, v_h) = \dfrac{1}{2}\Big(\dfrac{\partial \hat{f}(u_h^k)}{\partial x} + \dfrac{\partial \hat{f}(u_h^{k-1})}{\partial x}, v_h\Big),\ v_h \in V_h, & \text{(8.6.10a)} \\ \qquad k = 1,2,\cdots,M, & \\ u_h^0 = u_{0h}. & \text{(8.6.10b)} \end{cases}$$

**Lemma 8.6.3** *For sufficiently small $\tau > 0$, problem (8.6.10) has a unique solution $\{u_h^k\} \subset U_h$, satisfying*

$$\|u_h^k\|_1 \le C_6 \|u_{0h}\|_1, \ k = 1, 2, \cdots, M. \tag{8.6.11}$$

**Proof** Suppose $u_h^{k-1}$ is given. By the property (iii) of $a(u_h, \Pi_h^* v_h)$, for any $u_h \in U_h$ we have a unique $Gu_h \in U_h$ such that

$$a(Gu_h, \Pi_h^* w_h) = a(u_h^{k-1}, \Pi_h^* w_h) + \frac{\tau}{2}\Big(\frac{\partial \hat{f}(u_h)}{\partial x} + \frac{\partial \hat{f}(u_h^{k-1})}{\partial x}, \Pi_h^* w_h\Big),$$
$$\forall w_h \in V_h. \tag{8.6.12}$$

If we choose $w_h = Gu_h$, then we have by (i)-(iii) that

$$\begin{aligned} \|Gu_h\|_1 &\le \frac{M}{\alpha}\|u_h^{k-1}\|_1 + \frac{\tau}{2} d_\epsilon''(\|u_h\|_1 + \|u_h^{k-1}\|_1) \\ &= \Big(\frac{M}{\alpha} + \frac{\tau}{2} d_\epsilon''\Big)\|u_h^{k-1}\|_1 + \frac{\tau}{2} d_\epsilon'' \|u_h\|_1. \end{aligned}$$

Let $\tau > 0$ be sufficiently small such that $(1 - \frac{\tau}{2} d_\epsilon'') > 0$, and let $R$ be a constant satisfying $R > (1 - \frac{\tau}{2} d_\epsilon'')^{-1}(\frac{M}{\alpha} + \frac{\tau}{2} d_\epsilon'')\|u_h^{k-1}\|_1$. Then $\|Gu_h\|_1 \le R$ when $\|u_h\|_1 < R$. But $G$ is a continuous mapping. So by Brouwer fixed point theorem, $G$ has a fixed point $u_h^k \in U_h$, which is exactly the solution to (8.6.10).

(8.6.10) implies that

$$a(u_h^k, v_h) = a(u_h^0, v_h) + \frac{\tau}{2}\sum_{l=1}^{k}\Big(\frac{\partial \hat{f}(u_h^l)}{\partial x} + \frac{\partial \hat{f}(u_h^{l-1})}{\partial x}, v_h\Big), \ \ \forall v_h \in V_h.$$

Taking $v_h = \Pi_h^* u_h^k$ yields

$$\|u_h^k\|_1 \le \frac{M}{\alpha}\|u_{0h}\|_1 + \frac{\tau}{2} d_\epsilon'' \sum_{l=1}^{k}(\|u_h^l\|_1 + \|u_h^{l-1}\|_1),$$

and consequently

$$\|u_h^k\|_1 \le (1 - \frac{\tau}{2} d_\epsilon'')^{-1}\Big[\Big(\frac{M}{\alpha} + \frac{\tau}{2} d_\epsilon''\Big)\|u_{0h}\|_1 + d_\epsilon'' \tau \sum_{l=1}^{k-1} \|u_h^l\|_1\Big].$$

So the discrete Gronwall's inequality guarantees the existence of a constant $C_6 > 0$ which validates (8.6.11).

Finally, we deal with the uniqueness of the solution. Let $u_h^k$ and $\hat{u}_h^k$ be two solutions of (8.6.10). Then

$$\begin{aligned}
& a(u_h^k - \hat{u}_h^k, v_h) \\
= \ & a(u_h^{k-1} - \hat{u}_h^{k-1}, v_h) + \frac{\tau}{2}\Big(\frac{\partial \hat{f}(u_h^k)}{\partial x} - \frac{\partial \hat{f}(\hat{u}_h^k)}{\partial x}, v_h\Big) \\
& + \frac{\tau}{2}\Big(\frac{\partial \hat{f}(u_h^{k-1})}{\partial x} - \frac{\partial \hat{f}(\hat{u}_h^{k-1})}{\partial x}, v_h\Big), \quad \forall v_h \in V_h.
\end{aligned}$$

Take $v_h = \Pi_h^*(u_h^k - \hat{u}_h^k)$ and note the boundedness of the solution, then we have

$$\begin{aligned}
\|u_h^k - \hat{u}_h^k\|_1 \leq \ & \frac{M}{\alpha}\|u_h^{k-1} - \hat{u}_h^{k-1}\|_1 \\
& + C\tau(\|u_h^k - \hat{u}_h^k\|_1 + \|u_h^{k-1} - \hat{u}_h^{k-1}\|_1).
\end{aligned}$$

So for sufficiently small $\tau > 0$

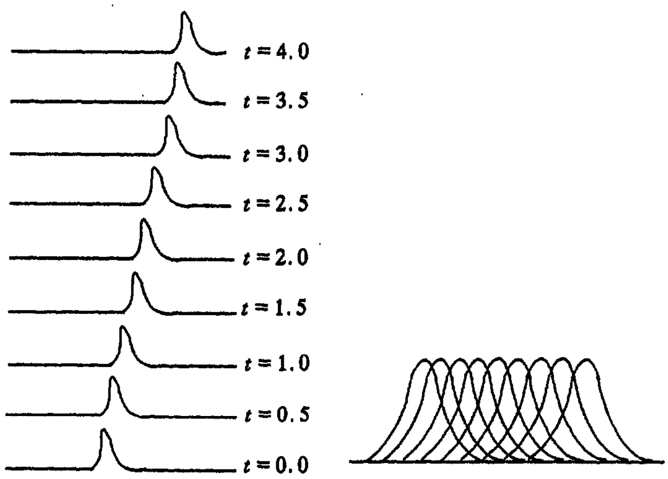
$$\|u_h^k - \hat{u}_h^k\|_1 \leq C\|u_h^{k-1} - \hat{u}_h^{k-1}\|_1,$$

where $C > 0$ is a general constant. This recursive relation implies $u_h^k = \hat{u}_h^k$. This completes the proof. □

Having done these preparations, we can prove the following theorem as in §5.2 and Theorem 8.6.1.

**Theorem 8.6.2** *Assume that the solution to (8.6.1) is sufficiently smooth. Then for small enough $h$ and $\tau$, the fully-discrete scheme (8.6.9) has a unique solution $\{u_h^k\}$, and the following error estimate holds:*

$$\begin{aligned}
& \|u_h^k(\cdot) - u(\cdot, t_k)\|_1 \\
\leq \ & C\Big[\|u_0 - u_{0h}\|_1 + h^2\Big(|u_0|_3 + \int_0^{t_k}\Big|\frac{\partial u(\cdot,t)}{\partial t}\Big|_3 \mathrm{d}t\Big) \\
& + \tau^2\int_0^{t_k}\Big\|\frac{\partial^2 u(\cdot,t)}{\partial t^2}\Big\|_1 \mathrm{d}t\Big], \quad k = 1, 2, \cdots, M.
\end{aligned} \tag{8.6.13}$$




(a) (b)

Fig. 8.6.1 Propagation of isolated waves

### 8.6.3 Numerical experiments

We use the generalized difference scheme (8.6.9) to approximate the regularized long wave equation (8.6.1), with $\alpha = \beta = 1$, $\gamma = 0.1$. In particular, we investigate the propagation of single isolated waves and the collision of double isolated waves, so as to check the efficiency of the scheme.

#### Propagation of single isolated waves

Set $a = 0$ and $b = 20$ in (8.6.1) and define the initial function as

$$u_0(x) = 3(c_0 - 1)(\text{sech}X_0)^2, \; X_0 = \sqrt{\frac{c_0 - 1}{4\gamma c_0}}(x - d_0),$$

where $c_0 = 2$, $d_0 = 8$. Choose the step sizes $h = 0.1$ and $\tau = 0.05$ in Scheme (8.6.9). The numerical results are depicted in Fig. 8.6.1(a). To investigate the propagation of the waves, all the waves are depicted simultaneously in Fig. 8.6.1(b). We observe from the figures that the isolated wave propagates forward with velocity $c_0 = 2$ and amplitude

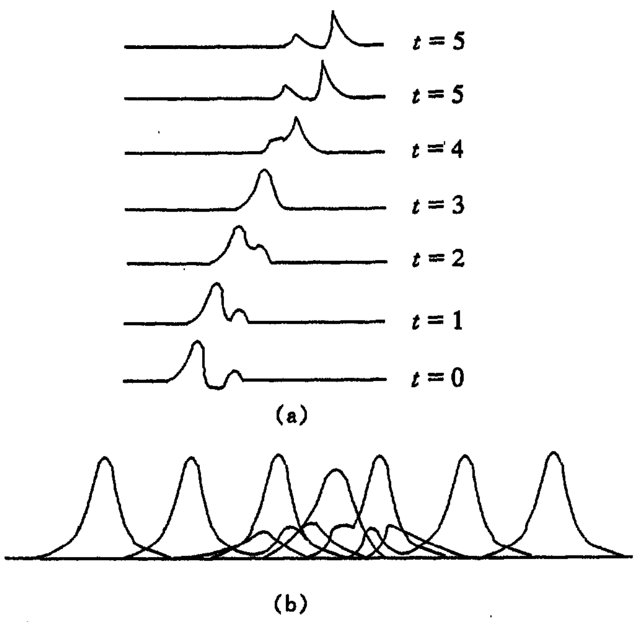



Fig. 8.6.2 Collision of double isolated waves

$3(c_0 - 1) = 3$, the wave form keeping unchanged. The numerical result matches quite well the theoretical isolated wave solution:

$$u(x,t) = 3(c_0 - 1)\text{sech}^2 X, \; X = \sqrt{\frac{c_0 - 1}{4\gamma c_0}}(x - c_0 t - d_0).$$

**Collision of double isolated waves**

Choose $a = 0, b = 30, h = 0.1, \tau = 0.05$, and the initial function

$$u_0(x) = \sum_{i=1}^{2} 3(c_i - 1)(\text{sech} X_i)^2, \; X_i = \sqrt{\frac{c_i - 1}{4\gamma c_i}}(x - d_i),$$

where $c_1 = 3, d_1 = 6, c_2 = 1.5, d_2 = 11.5$. The numerical results are shown in Fig. 8.6.2(a),(b). Amplify it twenty times in the longitudinal direction and truncate the heads of the waves to get Fig. 8.6.3.

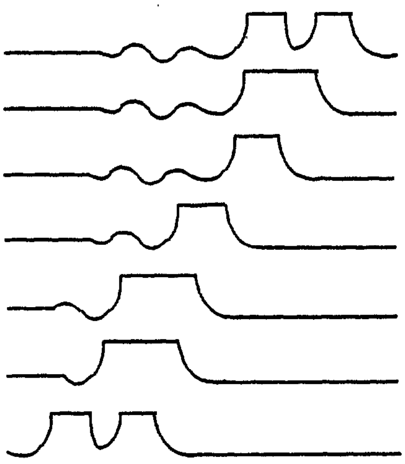



Fig. 8.6.3

One observes from this that before the collision, the big wave has a velocity $c_1 = 3$ and an amplitude $3(c_1 - 1) = 6$; and the small wave has a velocity $c_2 = 1.5$ and an amplitude $3(c_2 - 1) = 1.5$. During collision, the big wave gradually gets lower, while the small wave gets higher. After the collision, there appear once again a big wave and a small wave with almost identical amplitude and velocity as before the collision. But now apparently, the position of the big wave is (about 0.6) ahead of the position it would be at according to the original speed $c_1$, while the position of the small wave is (about 1.1) behind according to the speed $c_2$. We also notice the appearance of tail waves of slight vibrations.

## 8.7 Hierarchical Basis Methods

By now, we have established a theory for generalized difference methods, almost parallel to that for finite element methods. Generalized difference methods differs significantly from classical difference methods since they possess a variational form or a generalized Galerkin form. This advantage not only helps to establish the theory for generalized difference methods by use of finite element techniques, but also brings in the possibility to extend some algorithms designed for finite element methods to finite difference methods. As an example, we discuss in this section the hierarchical basis methods for finite difference equations. (cf. [A-29] and [B-56].)

It is well-known that the condition number of the coefficient matrix of the discrete equation is $O(h^{-2})$ ($h > 0$ being the maximum step size of the grid), when we use the usual finite element or finite difference methods to solve a self-adjoint, positive definite, second order elliptic, planar, boundary value problem. Great progress was made in [B-100] by introducing a hierarchical basis for finite elements, resulting in a condition number $O((\log\frac{1}{h})^2)$. In the sequel, we shall first write the difference equation into a generalized difference form, then we shall make use of the hierarchical basis techniques to improve the condition number in like manner.

## 8.7.1 Hierarchical Basis

Let $\Omega$ be a polygonal region with boundary $\Gamma = \partial\Omega$, and let $\overline{\Omega} = \Omega \cup \Gamma$. Suppose $T_0$ is an initial triangulation of $\Omega$. Starting from $T_0$, we construct successively a series of triangulations of $\Omega$: $T_0, T_1, \cdots$. Each $T_{k+1}$ is a uniform re-decomposition of the previous $T_k$, meaning that $T_{k+1}$ is obtained by dividing each triangle of $T_k$ into four equal smaller triangles with the three original vertexes and the three midpoints of the sides as new vertexes (cf. Fig. 8.7.1).

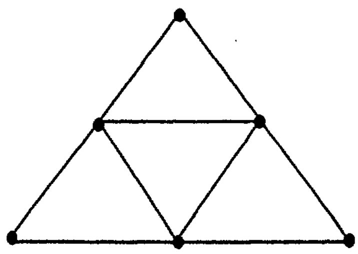


Fig. 8.7.1

Denote by $N_k$ the set of all the nodes, i.e., the vertexes of the triangular elements of $T_k$. $F_k$ denotes the piecewise linear, continuous function space relative to $T_k$. Call $u \in F_k$ a $k$-th hierarchy finite element function. For any $u \in C(\overline{\Omega})$, we use $I_k u \in F_k$ for the interpolation projection of $u$ onto $F_k$, that is,

$$I_k u \in F_k, \ (I_k u)(x) = u(x), \ \text{when } x \in N_k. \tag{8.7.1}$$

Then, any $u \in F_j$ has the following decomposition

$$u = I_0 u + \sum_{k=1}^{j} (I_k u - I_{k-1} u). \tag{8.7.2}$$

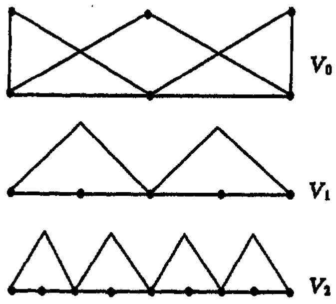



Fig. 8.7.2

Here each term is a rapidly oscillating function related to different hierarchies. $I_0 u$ is the finite element function relative to the initial grid $T_0$. The function $(I_k u - I_{k-1} u) \in F_k$ vanishes on $N_{k-1}$. Set $V_k = \{u \in F_k; u(x) = 0 \text{ when } x \in N_{k-1}\}$. Apparently $V_k$ is the range of $I_k - I_{k-1}$, and (8.7.2) means that $F_j$ is a direct sum of $V_0 = F_0, V_1, \cdots, V_j$. A standard finite element method takes the nodal basis functions of $F_j$ as a basis. Here we introduce a hierarchical basis of $F_j$, consisting of the nodal basis functions of $V_0 = F_0, V_1, \cdots, V_j$ (the corresponding nodes are respectively $N_0, N_1 \backslash N_0, \cdots, N_k \backslash N_{k-1}, \cdots, N_j \backslash N_{j-1}$). Some one-dimensional hierarchical basis functions are depicted in Fig. 8.7.2.

With respect to the decomposition (8.7.2), let us introduce a seminorm for $u \in F_j$

$$|u|^2 = \sum_{k=1}^{j} \sum_{x \in N_k \backslash N_{k-1}} |(I_k u - I_{k-1}) u(x)|^2. \tag{8.7.3}$$

Actually, $|u|$ is the Euclidian norm of the vector of the coefficients in the hierarchical basis expansion of $u$ relative to $F_j$, of all the hierarchies except the initial one. Suppose the dimension of $V_k$ is $m_k$ ($k = 0, 1, \cdots, j$), and the nodal basis functions are $\phi_{ki}$ ($i = 1, 2, \cdots, m_k$). Then any $u \in F_j$ has the expression

$$u = \sum_{i=1}^{m_0} v_{0i}\phi_{0i} + \sum_{k=1}^{j}\sum_{i=1}^{m_k} v_{ki}\phi_{ki}, \tag{8.7.4}$$

where

$$\begin{cases} v_{0i} = u(x_i), & x_i \in N_0, \\ v_{ki} = (I_k u - I_{k-1}u)(x_{ki}), & x_{ki} \in N_k \backslash N_{k-1}. \end{cases} \tag{8.7.5}$$

If we unify the symbols $\phi_{ki}$'s as $\phi_1, \cdots, \phi_{m_0}, \phi_{m_0+1}, \cdots, \phi_{m_0+m_1}, \cdots, \phi_m$, where $m = \sum\limits_{k=0}^{j} m_k$ is the dimension of $F_j$, then

$$u = \sum_{i=1}^{m_0} v_i\phi_i + \sum_{i=m_0+1}^{m} v_i\phi_i. \tag{8.7.6}$$

So it is clear that

$$|u|^2 = \sum_{i=m_0+1}^{m} |v_i|^2. \tag{8.7.3$'$}$$

Denote by $\|\cdot\|_0$ and $\|\cdot\|_1$ the norms of the Sobolev spaces $H^0 = L^2$ and $H^1$ respectively. The following important result is given in [B-100].

**Theorem 8.7.1** *There exist constants $K_1^*, K_2^* > 0$, dependent on the diameter of $\Omega$ and the lower bound of the inner angles of the triangular elements but independent of the hierarchy number $j$, such that the following inequality holds for all $u \in F_j$:*

$$K_1^*(j+1)^{-2}(\|I_0 u\|_1^2 + |u|^2) \le \|u\|_1^2 \le K_2^*(\|I_0 u\|_1^2 + |u|^2). \tag{8.7.7}$$

### 8.7.2 Application to difference equations

Take the grid $T_h = T_j$ and the trial function space $U_h = F_j$. Choose a dual grid $T_h^*$ of $T_h$, and the corresponding test function space $V_h$ as the piecewise constant function space related to $T_h^*$. Let $a(u,v)$ be a bilinear form associated with a symmetric and positive definite second order elliptic operator on $\Omega \subset R^2$. The generalized difference method then reads: Find $u_h \in U_h$ such that

$$a(u_h, v_h) = (f, v_h), \quad \forall v_h \in V_h. \tag{8.7.8}$$

In particular, taking $T_h$ as a rectangular grid results in a classical difference equation. Denote by $\Pi_h^*$ the interpolation projection operator from $U_h$ onto $V_h$. Then $a(u_h, \Pi_h^* \bar{u}_h)$ $(\bar{u}_h, u_h \in U_h)$ is symmetric, and there exist positive constants $\gamma_0$ and $\gamma_1$ such that

$$\gamma_0 \|u_h\|_1^2 \leq a(u_h, \Pi_h^* u_h) \leq \gamma_1 \|u_h\|_1^2. \tag{8.7.9}$$

Inspired by this observation, we introduce for $u \in F_j$ the following norms:

$$|||u|||_1^2 = a(u_h, \Pi_h^* u_h); \; |||u|||_0^2 = \|I_0 u\|_0^2 + |u|^2. \tag{8.7.10}$$

Then (8.7.9) can be written as

$$C_1 \|u\|_1^2 \geq |||u|||_1^2 \geq C_2 \|u\|_1^2, \quad \forall u \in F_j. \tag{8.7.11}$$

This together with Theorem 8.7.1 leads to

**Theorem 8.7.2** *There exist constants $K_1$ and $K_2$ independent of the hierarchy number $j$ such that*

$$K_1 (j+1)^{-2} |||u|||_0^2 \leq |||u|||_1^2 \leq K_2 |||u|||_0^2, \quad \forall u \in F_j. \tag{8.7.12}$$

*The constants $K_1$ and $K_2$ here depend on the lower bound of the inner angles of all the triangular elements, the diameter of $\Omega$ and the constants $C_1$ and $C_2$. But they are independent of the regularity of the boundary value problems.*

The coefficient matrix (the stiff matrix) of the generalized difference equation (8.7.8) has the following two expressions according to different bases of $U_h$ and $V_h$.

**Representation by nodal basis**

Let $u_h = \sum_i u_i \phi_{x_i}$, where $\phi_{x_i}$ is the nodal basis of $x_i \in N_j$ and $u_i = u_h(x_i)$. We rewrite (8.7.8) in the form

$$\sum_i a(\phi_{x_i}, \Pi_h^* \phi_{x_l}) u_i = (f, \Pi_h^* \phi_{x_l}) = \hat{b}_l, \quad \forall x_l \in N_j. \tag{8.7.13}$$

Set $u = (u_1, u_2, \cdots, u_m)^T$, $\hat{b} = (\hat{b}_1, \hat{b}_2, \cdots, \hat{b}_m)^T$, and (a matrix) $\hat{A} = [a(\phi_{x_i}, \Pi_h^* \phi_{x_l})]_{m \times m}$. Then, (8.7.13) can be written as a vector form: $\hat{A}u = \hat{b}$. This gives a usual (nodal basis) generalized difference equation.

**Representation by hierarchical basis**

The difference solution $u_h$ can be expressed as (8.7.6) by hierarchical bases. Accordingly we can write (8.7.8) as

$$\begin{aligned} &\sum_{i=1}^{m_0} a(\phi_i, \Pi_h^* \phi_l) v_i + \sum_{i=m_0+1}^{m} a(\phi_i, \Pi_h^* \phi_l) v_i \\ = \; & (f, \Pi_h^* \phi_l) = b_l, \; l = 1, 2, \cdots, m. \end{aligned} \tag{8.7.14}$$

Write $v = (v_1, \cdots, v_{m_0}, v_{m_0+1}, \cdots, v_m)^T$, $b = (b_1, b_2, \cdots, b_m)^T$, and the matrix $A = [a(\phi_i, \Pi_h^* \phi_l)]_{m \times m}$. Then (8.7.14) becomes $Av = b$, called a hierarchical basis difference equation. The matrix $A$ is obviously symmetric and positive definite.

Let $B_0$ be the $m_0$ order coefficient matrix on the upper left of the system (8.7.14), and

$$A_0 = \begin{pmatrix} B_0 & 0 \\ 0 & I \end{pmatrix}, \tag{8.7.15}$$

where $I$ is the $(m - m_0) \times (m - m_0)$ identity matrix. It is clear that the norm $|||\cdot|||_0$ in (8.7.10) has the following representation

$$|||v|||_0^2 = \|I_0 v\|_0^2 + |v|^2 = \langle v, A_0 v \rangle.$$

Here $\langle \cdot, \cdot \rangle$ stands for the Euclidian inner product. Now Theorem 8.7.2 can be stated as: For any $m$-dimensional hierarchical basis coefficient vector $x$, there holds (cf. (8.7.10))

$$K_1(j+1)^{-2}\langle x, A_0x\rangle \le \langle x, Ax\rangle \le K_2\langle x, A_0x\rangle. \tag{8.7.16}$$

Since the order $m_0$ of $B_0$ is not large, it is economical to compute the Cholesky decomposition $A_0 = LL^T$. Then we insert this decomposition into (8.7.16) to obtain

$$\begin{aligned} & K_1(j+1)^{-2}\langle L^Tx, L^Tx\rangle \\ \le\ & \langle L^Tx, L^{-1}A(L^{-1})^TL^Tx\rangle \le K_2\langle L^Tx, L^Tx\rangle. \end{aligned}$$

Thus we have the following estimate for the spectral condition number

$$\mathrm{cond}(L^{-1}A(L^{-1})^T) \le \frac{K_2}{K_1}(j+1)^2. \tag{8.7.17}$$

Here the factors $L^{-1}$ and $(L^{-1})^T$ are not very important. They are mainly used to eliminate the geometrical influence of the initial triangulation. Since the low dimensional initial space $F_0$ is fixed and independent of the number of the hierarchies, the spectral condition number of $LL^T$ is independent of number of the hierarchies, too. Furthermore, we note

$$\frac{\langle Ax, x\rangle}{\langle x, x\rangle} = \frac{\langle L^Tx, L^{-1}A(L^{-1})^TL^Tx\rangle}{\langle L^Tx, L^Tx\rangle}\frac{\langle L^Tx, L^Tx\rangle}{\langle x, x\rangle}.$$

Therefore, the condition number of the matrix $A$ amplifies only in an order $O(j^2)$. Notice $j \sim \log\frac{1}{h}$. Hence, as in the case of finite element methods, the spectral condition number is reduced to $O\Big(\big(\log\frac{1}{h}\big)^2\Big)$.

Unfortunately $A$ is no longer a sparse matrix, but it bears a simple relationship with the nodal basis coefficient matrix $\hat{A}$ (cf. (8.7.13)). Let $S$ be such a transformation matrix which transforms the functions in $F_j$ from a hierarchical basis representation into a nodal basis representation. So for any $m$-dimensional vectors $\bar{x}$ and $\bar{y}$ we have

$$\langle \bar{x}, A\bar{y}\rangle = \langle S\bar{x}, \hat{A}S\bar{y}\rangle = \langle \bar{x}, S^T\hat{A}S\bar{y}\rangle, \tag{8.7.18}$$

which implies

$$A = S^T\hat{A}S. \tag{8.7.19}$$

Our difference system is

$$\hat{A}x = \hat{b}. \tag{8.7.20}$$

where $x$ is the difference solution vector. This system is equivalent to

$$Ay = S^T\hat{b}, \quad (\text{or resp. } L^{-1}A(L^{-1})^T y = L^{-1}S^T\hat{b},) \tag{8.7.21}$$

where $A = S^T\hat{A}S$ is the hierarchical basis coefficient matrix, and the nodal basis solution vector

$$x = Sy. \quad (\text{resp. } x = S(L^{-1})^T y.) \tag{8.7.22}$$

Since the condition number of $A$ (resp. $L^{-1}A(L^{-1})^T$) is comparatively small, the convergence rate of an iteration method for (8.7.21) will be remarkably increased.

### 8.7.3 Iteration methods

As we see from (8.7.19), (8.7.21) and (8.7.22), it is a key point how to evaluate $S$. As a transformation matrix from a hierarchical basis representation into a nodal basis representation, $S$ can be decomposed into $S = S_jS_{j-1}\cdots S_1$, where $S_k$ describe the computation of the values of the new nodes on $k$-th hierarchy in terms of the values of the nodes on $(k-1)$-th hierarchy. The diagonal entries of $S_k$ are equal to 1; In the $i$-th row of $S_k$, there are two off-diagonal entries $\frac{1}{2}$ for each $x_i \in N_k\backslash N_{k-1}$; All the other entries are zero. So we can choose some fast algorithms to evaluate $S_kx$ and $S_k^Tx$, with only $O(m)$ operations of multiplications and divisions ([cf. [B-100]). If we choose such iteration methods that involve only operations like $Ax$, then we can avoid the difficulties such as the complex structure and the increased nonzero entries of $A$. The following two examples illustrate the idea.

**Richardson iteration method**

Applying the Richardson method to (8.7.21) yields

$$y^{k+1} = y^k - \omega(S^T\hat{A}Sy^k - S^T\hat{b}), \; k = 0, 1, \cdots. \tag{8.7.23}$$

If we choose the optimal relaxation factor $\omega_{\text{opt}} = 2/(\alpha + \beta)$, where $\alpha$ and $\beta$ are the maximum and minimum eigenvalues respectively of the symmetric and positive definite matrix $A$, then the convergence rate of (8.7.23) is $O(j^{-2})$, and the computational amount of each step is only $O(m)$.

**Conjugate gradient method**

The convergence rate of this method for (8.7.21) is $O(j^{-1})$, and the computational amount of each step is also $O(m)$.

Block Gauss-Seidel method and successive block over-relaxation method can as well be used for (8.7.21). (See [A-29].)

### 8.7.4 Numerical experiments

Consider the first boundary value problem of The Laplacian equation:

$$\begin{cases} \Delta u = u_{xx} + u_{yy} = 0, \text{ on } \Omega = (0,1)\times(0,1), \\ u|_{\partial\Omega} = 0. \end{cases}$$

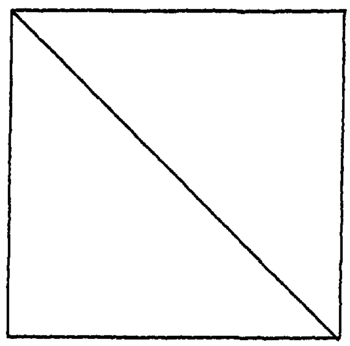


Fig. 8.7.3

Its unique solution is $u = 0$. Based on the initial grid in Fig. 8.7.3, we place a series of densifying triangulations. In this case, the nodal difference equation is precisely the well-known five point difference scheme. The following methods are used: nodal basis Gause-Seidel method (N-GS in short), hierarchical basis Richardson method (H-Richardson), conjugate gradient method (H-CG), block Gauss-Seidel method (H-BGS) and block successive over-relaxation method (H-BSOR). The numbers of iteration steps to reach the accuracy $10^{-3}$ are shown in the following table. We observe that the convergence rates of the hierarchical methods (H-) are significantly higher than that of the usual iteration method (N-GS).

| method | $j=2$ | $j=3$ | $j=4$ | $j=5$ | $j=6$ |
|---|---|---|---|---|---|
| N-GS | 11 | 47 | 191 | 767 | 3067 |
| H-Richardson | 14 | 32 | 58 | 92 | 134 |
| H-CG | 4 | 11 | 16 | 22 | 28 |
| H-BGS | 9 | 13 | 19 | 26 | 35 |
| H-BSOR | 7 | 9 | 10 | 12 | 16 |

## Bibliography and Comments

The paper [B-65] by R.H. MacNeal (1953) is the earliest work studying the difference methods on irregular networks, which was originally proposed to simulate an electric network. A.M. Winslow [B-99] (1967) extended this method to the computation of electromagnetic fields. Further extensions and applications were discussed in [A-60] and [B-72]. A direct extension to three-dimensional regions is presented in this chapter (§8.2).

Early in the seventies, [A-14,49] studied the difference methods on triangular grids for elastic problems, which were called discretization operator methods.

Up to now, fluid mechanics, especially underground fluid mechanics, may be the field where the generalized difference methods apply most often and most successfully, represented by the works [A-38,39,40,13] and [B-29,38]. An important feature of generalized difference methods is that they keep up the mass conservation. Perhaps this is the reason why computational fluid researchers are attracted to them. The application of generalized difference methods in aeromechanics is also fairly successful, cf §6.4 and the corresponding references in the end of the book.

[A-48] proposes a significant extension of the staggered scheme (the nodes of the pressure and the velocity being distributed alternately) for incompressible flows. [A-56] discussed a coupled sound-heat problem, but only in one dimension. [A-4] studied a regularized long wave problem, which was a nonlinear problem, and obtained satisfactory wave forms by a quadratic element generalized difference scheme.

[A-29] and [B-56] extended the hierarchical basis method for finite element methods to finite difference methods, and highlighted an approach to transplant an algorithm for finite element methods to one for finite difference methods. Finite difference methods were used in [A-33] to compute the long time behaviour of dynamical systems, which provided another example of a successful extension of finite element theories to finite difference methods.

# Bibliography

## Part A. (References written in Chinese)


[1] Bao G. and Wu W., Approximate order estimates of generalized difference methods for second order hyperbolic equations, Acta Sci. Natur. Univ. Jilin., 2(1987), 33-42.

[2] Ciarlet P. G., Numerical Analysis of Finite Element Methods, Shanghai Press of Science and Technology, 1978.

[3] Chen M. and Chen Z., Operator Equations and Their Projection Approximations, Guangdong Press of Science and Technology, 1992.

[4] Chen M. and Chen Z., Lagrangian quadratic generalized difference methods for RLW equations and numerical simulation of collision procedure of double isolated waves, Appl. Math. & Numer. Math., 2(1993), 21-32.

[5] Chen M. and Zhang Y., Generalized difference methods for regularized long wave equations, Acta Sci. Natur. Univ. Sunyaseni, 4(1993), 20-27.

[6] Chen Z., A generalized difference method for heat transfer equations, Acta Sci. Natur. Univ. Sunyaseni, 1(1990), 6-13.

[7] Chen Z., A nonconforming generalized difference method for biharmonic equations, Proceedings to the Tianjin Conference of Computational Mathematics, 1991.

[8] Chen Z., Generalized difference methods based on Adini elements for biharmonic equations, Acta Sci. Natur. Univ. Sunyaseni, 1(1993), 21-29.

[9] Chen Z., Variational principle and numerical analysis of generalized difference methods for biharmonic equations, Numer. Math. -J. Chinese Universities, 2(1993), 182-194.

[10] Chen Z., $L^2$ estimates of linear element generalized difference schemes, Acta Sci. Natur. Univ. Sunyaseni, 4(1994), 22-28.

[11] Chen Z., Quadratic generalized difference methods for two point boundary value problems, Acta Sci. Natur. Univ. Sunyaseni, 3(1994), 19-24.

[12] Chen Z., Zienkiewicz element generalized difference schemes and their applications, Acta Sci. Natur. Univ. Sunyaseni, 4(1996), 34-40.

[13] Computing Center of Beijing Agricultural University, Applications of finite element methods in simulation of transformation of soil and solute in earth, Research Report of 8-1-4-1 Project of Chinese Agricultural Department, 1990.

[14] Engeneering mechanics group of Dalian Institute of Technology, Differential operator discretization methods for continuous structure computation, I,II, Acta Sci. Natur. Univ. Dalian Tech., 1,3(1973).

[15] Huang M., Finite Element Methods for Evolution Equations, Shanghai Press of Science and Technology, 1988.

[16] Hu H., Variational Principles of Elastic Mechanics and Their Applications, Scince Press, Beijing, 1981.

[17] Liang D., Upwind generalized difference schemes for convection-diffusion equations, Appl. Math, 4(1990), 456-466.

[18] Liang D., A class of upwind schemes for convection diffusion equations, Math. Numer. Sinica, 1(1991), 133-141.

[19] Li L. and Guo Y., Introduction to Sobolev Spaces, Shanghai Press of Science and Technology, 1981.

[20] Li Q., Generalized difference methods for second order parabolic equations, Acta Sci. Natur. Univ. Shandong., 1(1983), 41-52.

[21] Li Q., Mass concentration generalized Galerkin methods for parabolic equations, Acta Sci. Natur. Univ. Shandong., Special Issue (1984), 31-44.

[22] Li Q., Generalized Galerkin methods for elliptic equations with third kind boundary conditions, Numer. Math. -J. Chinese Universities, 1(1986), 2-11.

[23] Li Q., Generalized Galerkin methods for a class of nonlinear parabolic equations, Numer. Math. -J. Chinese Universities, 2(1986), 97-104.

[24] Li Q., Generalized Galerkin methods for a class of nonlinear hyperbolic equations, Math. Numer. Sinica, 2(1986), 150-158.

[25] Li R., Generalized difference methods for two point boundary value problems, Acta Sci. Natur. Univ. Jilin., 1(1982), 26-40.

[26] Li R., Galerkin Methods for Boundary Value Problems, Shanghai Press of Science and Technology, 1988.

[27] Li R. and Feng G., Numerical Solution of Differential Equations, Publishing House of Higher Educations, China, 1980.

[28] Li R. and Hua X., Generalized difference methods for convection-dominated diffusion problems, Proceedings to Tianjin Conference of Numerical Mathematics, 187-190.

[29] Li R., Liu B., Wu H. and Li F., Applications of hierarchical basis methods to difference equations, Acta Sci. Natur. Univ. Jilin., 1(1995), 9-13.

[30] Li R. and Zhu P., Generalized difference methods for second order elliptic partial differential equations (I) - triangle grids, Numer. Math. -J. Chinese Universities, 2(1982), 140-152.

[31] Li Y., A mixed generalized difference method for biharmonic equations, Acta Sci. Natur. Univ. Jilin., 3(1993), 19-30.

[32] Li Y. and Li R., On a class of Generalized difference methods with BB dual subdivision, Numer. Math. – J. Chinese Univ., 1(1998), 56-68.

[33] Mu G., Long time convergence of generalized difference methods for semilinear parabolic equations, Master Thesis, Math. Insti. Jilin Univ., 1992.

[34] Mu Z., Generalized difference methods discretizing along characteristic curves for convection-dominated diffusion equations, Proceedings to Tianjin Conference of Numerical Mathematics, 264-267.

[35] Ni P., Generalized difference methods for a beam balance equation, Numer. Math. -J. Chinese Universities, 4(1984), 285-295.

[36] Ni P. and Wu W., Superconvergence of generalized Galerkin methods, Numer. Math. -J. Chinese Universities, 2(1986), 153-158.

[37] 515 Research Group, Finite element method and maximum norm principle for instaneouly varing temperature fields, Math. Numer. Sinica, 2(1982), 113-120.

[38] Sun N., Models and Computations of Underground Waters, Geology Press, Bejing, 1981.

[39] Sun N. and Liang W., Multi-element balancing methods simulating underground water contamination, Acta Sci. Natur. Univ. Shandong, 3(1982), 12-32.

[40] Sun N., Underground Water Contamination: Mathematical Models and Numerical Methods, Geology Press, Bejing, 1989.

[41] Tian M. and Chen Z., Quadratic element generalized difference methods for elliptic equations, Numer. Math. -J. Chinese Universities, 2(1991), 99-113.

[42] Wang S., Generalized differnce methods for Schrodinger equations, Acta Sci. Natur. Univ. Shandong., 2(1985), 31-48.

[43] Wang S., Estimation of convergence rate of Crank-Nicolson scheme for parabolic equations, Acta Sci. Natur. Univ. Shandong., 2(1985), 25-30.

[44] Wang S., Generalized difference methods for first order hyperbolic equations, Numer. Math. -J. Chinese Universities, 1(1987), 73-77.

[45] Wang S., Variational principle and $H^1$ error estimate of generalized difference methods for a class of quasilinear hyperbolic equations, Math. Numer. Sinica, 4(1988), 345-355.

[46] Wang S., Variational principle and $H^1$ error estimate of generalized difference methods for a class of nonlinear parabolic equations, Math. Numer. Sinica, 3(1989), 225-230.

[47] Wang S., Generalized difference methods for two-dimensional quasilinear hyperbolic equations, Acta Sci. Natur. Univ. Shandong., 1(1990), 10-17.

[48] Wang S., Nonconforming generalized difference methods for Stokes problems, Chinese J. of Comput. Phy., 2(1993), 129-136.

[49] Wu C., A new method for numerical analysis of bending elastic sheet: the operator discretization method, Scientiarum Sinica, 4(1975), 360-375.

[50] Wu W., Generalized difference methods for a class of elliptic and parabolic variational inequalities, Northeast. Math., 2(1986), 178-185.

[51] Wu W., Mixed generalized difference methods for biharmonic equations, Acta Sci. Natur. Univ. Jilin., 3(1986), 14-22.

[52] Wu W., Error estimates of generalized difference methods for nonlinear parabolic equations, Math. Numer. Sinica, 2(1987), 119-132.

[53] Wu W. and Li R., Generalized difference methods for second order elliptic and parabolic differential equations in one dimension, Annual Math., 5A(3)(1984), 303-312.

[54] Xiang X., Lagrangian quadratic element generalized difference methods for two point boundary value problems, Acta Sci. Natur. Univ. Heilongjiang, 2(1982), 25-34.

[55] Xiang X., Generalized difference methods for second order elliptic equations, Numer. Math. - J. Chinese Universities, 2(1983), 114-126.

[56] Xu Z. and Ma S., Generalized difference methods for coupled sound-heat equations, Appl. Math. - J. Chinese Universities, 4(1987), 385-395.

[57] Ying L. and Teng Z., Hyperbolic conservation laws: differential systems and difference methods, Science Press, Beijing, 1991.

[58] Yuan Y., Stability and convergence of operator discretization mehtods for parabolic equations, Numer. Math. -J. Chinese Universities, 3(1982), 208-221.

[59] Zhang T., Discontinuous finite element methods for first order hyperbolic systems, Acta Sci. Natur. Univ. Northeast. College Tech., 2(1987), 250-257.

[60] Zhao G. and Liu Y., Application of irregular network differenceing in computation of three-dimensional fields, Acta Sci. Natur. Univ. Southeast., 1(1990), 69-74.

[61] Zhu Q. and Lin Q., Superconvergence Theory of Finite Elements, Hunan Press of Science and Technology, 1989.

[62] Zhu P. and Li R., Generalized difference methods for second order elliptic partial differential equations (II) - quadrilateral grids, Numer. Math. -J. Chinese Universities, 4(1982), 360-375.

## Part B. (References written in English)

[1] Adams R. A., Sobolev Space, Academic Press, New York, 1975.

[2] Babuska I. and Aziz A. K., Survey lectures on the mathematical foundation of the finite element method, The Mathematical Foundation of the Finite Element Method with Applications to Partial Differential Equations, Amer. Academic Press, 3-359.

[3] Baba K. and Tabata M., On a conservative upwind finite element scheme for convective diffusion equations, R.A.I.A.O. Numer. Anal., 15(1981), 3-25.

[4] Bank R. E. and Rose D. J., Some error estimates for the box method, SIAM J. Numer. Anal., 24(1987), 777-787.

[5] Baranger J., Maitre J-F. and Oudin F., Connection between finite volume and mixed finite element methods, Math. Model. and Numer. Anal., 30(4)(1996), 445-465.

[6] Benharbit S., Chalabi A. and Vila J. P., Numerical viscosity and convergence of finite volume methods for conservation law with boundary conditions, SIAM J. Numer. Anal., 32(1995), 775-796.

[7] Bona J. L. and Dougalis V. A., An initial and boundary-value problem for a model equation for propagation of long waves, J. Math. Anal. Appl., 75(1980), 503-522.

[8] Book L., Finite difference techniques for vectorized fluid dynamics calculation, Spring-Verlag, New York, Heidelberg, Berlin, 1981.

[9] Cai Z., On the finite volume element method, Numer. Math., 58(1991), 713-735.

[10] Cai Z. and McCormick S., On the accuracy of the finite volume element methods for diffusion equations on composite grid, SIAM J. Numer. Anal., 27(1990), 636-655.

[11] Cai Z., Mandel J. and McCormick S., The finite volume element methods for diffusion equations on general triangulations, SIAM J. Numer. Anal., 28(1991), 392-402.

[12] Champier S., Gallouet T. and Herbin R., Convergence of an upwind finite volume scheme for a nonlinear hyperbolic equation on a triangular mesh, Numer. Math., 66(1993), 139-157.

[13] Chen G., Chen Z. and Xu Y., Numerical computation of a damped slewing beam with tip mass, Commun. Numer. Meth. Engng, 15, 249-261(1999).

[14] Chen Z., The error estimate of generalized difference methods of 3rd-order Hermite type for elliptic partial differential equations, Northeast. Math., J., 8(1992), 127-135.

[15] Chen Z., Superconvergence of generalized difference methods for elliptic boundary value problem, Numer. Math. - J. Chinese Univ., English Series, 3(1994), 163-171.

[16] Chen Z. and Xu Y., The Petrov-Galerkin and iterated Petrov-Galerkin methods for second kind integral equations, SIAM J. Numer. Anal., 35(1998), 406-434.

[17] Ciarlet P., The finite element method for elliptic problems, North-Holland, 1978.

[18] Ciarlet P. and Raviart P., A mixed finite element method for the biharmonic equation, Symposium on mathematical aspects of finite element in partial differential equations, ed. by C. de Boor, Academic Press, New York, 1974, 125-143.

[19] Cockburn B., Coquel F. and Lefloch P. G., Convergence of the finite volume method for multidimensional conservation laws, SIAM J. Numer. Anal., 32(1995), 687-705.

[20] Courant R., Friedrichs K. O. and Lewy H., Uber die partiellen diffenzengleichungen der mathematischen physik, Math. Ann., 100(1928), 32-74.

[21] Courant R., Isaacson E. and Rees M., On the solution of nonlinear hyperbolic differential equations by finite differences, Comm. Pure and Appl. Math., 5(1952), 243-255.

[22] Denkowicz L., On some convergence results for FDM with irregular mesh, Comp. Meth. in Appl. Mech. and Engin., (1984), 344-256.

[23] Derrienx A., Steady Euler simulations using unstructured meshes, in Partial Differential Equations of Hyperbolic Type and Application, World Science Publishing, Singapore, 1987.

[24] Douglas J. and Dupont T., Galerkin methods for parabolic equations, SIAM J. Numer. Anal., 7(1970), 575-625.

[25] Douglas J. and Dupont T.,A Galerkin method for a nonlinear Dirichlet problem, Math. Comput., 29(1975), 689-696.

[26] Douglas J. and Russel T. F., Numerical methods for convection-dominated diffusion problems based on combining the methods of characteristics with finite element or finite difference procedures, SIAM. J. Numer. Anal., 19(1982), 871-885.

[27] Dupont T., $L^2$-estimates for Galerkin methods for second order hyperbolic equations, SIAM J. Numer. Anal., 10(1973), 880-899.

[28] Ewing R., Lazarov R. and Vassilevski P., Local refinement techniques for elliptic problems on cell-centered grids, I. Error analysis, Math. Comp., 56(1991), 437-461.

[29] Eymard R. and Sonier F., Mathematical and numerical properties of control volume finite element schemes for reservoir simulation, SPE Reservoir Engineering, November(1994), 283-289.

[30] Falk R. S. and Osborn J. E., Error estimates for mixed methods, R.A.I.A.O., Numer.Anal., 14(1980), 249-277.

[31] Friedrichs K. O., Symmetric positive linear differential equations, Comm. Pure and Appl. Math., 11(1958), 333-418.

[32] Girault V., Theory of a finite difference method on irregular networks, SIAM J. Numer. Anal., 11(1974), 260-282.

[33] Girault V., Nonelliptic approximations of a class of partial differential equations with Neumann boundary conditions, Math. Comp., 30(1976), 68-91.

[34] Girault V. and Raviart P. A., Finite Element Approximation of the Navier-Stokes Equations, Lecture Notes in Mathematics, No. 749, Spring Verlag, 1979.

[35] Hackbusch W., On first and second order box schemes, Computing, 41(1989), 277-296.

[36] Harten A., High resolution schemes for hyperbolic conservation laws, J. Compt. Phys., 49(1983), 357-393.

[37] Harten A. et.al., Uniformly high order accurate essentially nonoscilatory schemes III, J. Comput. Phys., 71(1987), 231.

[38] Heinemann Z. E., Brand C. W., Munke M. and Chen Y. M., Modeling reservoir geometry with irregular grids, SPE Reservoir Engineering, May(1991), 225-232.

[39] Heinrich B., Finite difference methods on irregular networks, ISNM 82, Birkhauser Verlag, 1987.

[40] Herbin R., An error estimate for a finite volume scheme for a diffusion-convection problem on a triangular mesh, Numer. Meth. for Partial Diff. Equat., 11(1995), 165-174.

[41] Huang M., Stasbility and convergence analysis for box scheme for first order hyperbolic systems, Proceedings of the China-France Symposium on Finite Element Methods, Ed. by Feng K. and Lions J. L., Bejing, China, 696-731, 1982.

[42] Jameson A., Schmidt W. and Tukel E., Numerical solutions to the Euler equation by finite volume methods using Runge-Kutta time-marching schemes, AIAA paper, 81(1981), 1295.

[43] Jameson A. and Baker T. J., Solution of the Euler equations for complex configurations, Proceedings of AIAA 6th Computational Fluid Dynamics Conference, AIAA, New York, 293-302, 1983.

[44] Jameson A. and Mavriplis D., Finite volume solution of the two-dimensional Euler equations on a regular mesh, AIAA J., 24(4)(1986), 611-618.

[45] Kellogg R. B., Difference equations on a mesh arising from a general triangulation, Math. Comp., 18(1964), 203-210.

[46] Kellogg R. B., An error estimate for elliptic difference equations on a convex polygon, SIAM J. Numer. Anal., 3(1966), 79-90.

[47] Krasnoselskii M. A. et.al., Approximate Solution of Operator Equations, Wolters-Noordhoff, 1972.

[48] Kreiss H. O., Manteuffel T.A., Schwartz B, Wendroff B. and White A. B. Jr., Superconvergent schemes on irregular grids, Math. Comp., 47(1986), 537-554.

[49] Kröner D. and Rokyta M., Convergence of upwind finite volume schemes for scalar conservation laws in two dimensions, SIAM J. Numer. Anal., 31(1994), 324-343.

[50] Lascaux P. and Lesaint P., Some nonconforming finite elements for the plate bending problem, Rev. Fr. Auto. Inform. Rech. Oper., R-1(1975), 9-53.

[51] Lazarov R. D. and Mishev I. D., Finite volume methods for reaction-diffusion problems, in F. Benkhaldoun and R. Vilsmeier, editors, Finite Volumes for Complex Applications, 231-240, Hermes, 1996.

[52] Lazarov R. D., Mishev I. D. and Vassilevski P. S., Finite volume methods for convection-diffusion problems, SIAM J. Numer. Anal., 33(1996), 31-55.

[53] Lesaint P. and Paviart P. A., On a finite element method for solving the neutron transport equation, Mathematical Aspects of the Finite Element Method in Partial Differential Equations, Ed. by de Boor C., Academic Press, New York, 89-123, 1974.

[54] Leveque R. J., High resolution finite volume methods on arbitrary grids via wave propagation, J. Comput. Phys., 78(1988), 36-63.

[55] Liang D., $W^{1,p}$- and $L^2$-error estimates for generalized difference methods for second order elliptic equations, Northeast. Math., 6(1990), 235-242.

[56] Li F. and Li R., Multy-level splitting methods solving difference equations, Northeast. Math. J., 13(1997), 229-246.

[57] Li R., On generalized difference method for elliptic and parabolic differetial equations, Proceedings of the Symposium on the Finite Element Method between China and France, Ed. by Feng K. and Lions J.L., Beijing, China, 326-36-, 1982.

[58] Li R., Generalized difference methods for a nonlinear Dirichlet problem, SIAM J. Numer. Anal., 24(1987), 77-88.

[59] Li R., Chen Z. and Wu W., A survey on generalized difference methods and their analysis, Proceedings to Guangzhou Int. Symp. on Comput. Math., eds. by Chen Z., Li Y., Micchelli C. A. and Xu Y., Lecture Notes in Pure and Appl. Math., Vol. 202, 321-337, 1997.

[60] Li R. and Shen J., On a class of high accuracy upwind schemes, Northeast. Math., J., 7(1991), 11-26.

[61] Li R. and Hua X., Generalized upwind difference methods for convection-dominated diffusion problems, Proceedings of the Second Conference on Numerical Methods for Partial Differential Equations, Ed. by Ying L., World Scientific, 83-90, 1991.

[62] Li Y. and Li R., Generalized difference methods on arbitrary quadrilateral networks, to appear in J. of Comp. Math., 1999.

[63] Lin S. Y. and Chin Y. S., An upwind finite volume scheme with a triangular mesh for conservation laws, Nonolinear Analysis (Taibei, 1989), 115-133, World Sci. Publishing, Teaneck, NJ, 1991.

[64] Mackenzie J. A. and Morton K. W., Finite volume solution of convection-difffusion test problems, Math. Comp., 60(1992), 189-220.

[65] MacNeal R. H., An asymmetrical finite difference network, Quart. Appl. Math., 11(1953), 295-310.

[66] Mishev I. D., Finite volume methods on Voronoi meshes, Numer. Meth. for Partial Diff. Equat., 14(2)(1998), 193-212.

[67] Miller J. J. H. and Wang S., A new non-conforming Petrov-Galerkin finite element method with triangular elements for a singularly perturbed advection-diffusion problem, IMA J. Numer. Anal., 14(1994), 257-276.

[68] Mitchell A. R. and Griffiths, The Finite Difference Method in Partial Differential Equations, John Willy & Sons, 1980.

[69] Morton K. W., Finite volume and finite element methods for the steady Euler equations of gas dynamics, The Mathematics of Finite Element and Applications, VI(Uxbridge 1987), 353-377, Academic Press, London-New York, 1988.

[70] Morton K. W. and Süli E., Finite volume methods and their analysis, IMA J. Numer. Anal., 11(1991), 241-260.

[71] Nicolaides R. A., Porsching T. A. and Hall C. A., Covolume methods in computational fluid dynamics, in M. Hafez and K. Oshma, editors, Computational Fluid Dynamics Review, 279-299, John Wiley and Sons, 1995.

[72] Okon E. E., Finite difference approximations for the three-dimensional Laplacian in irregular grids, Z. Angew. Math. Phys., 33(1982), 266-281.

[73] Patankar S. V., Numerical Heat Transfer and Fluid Flow, McGraw-Hill, 1980.

[74] Richtmyer R. D. and Morton K. M., Difference Methods for Initial Value Problems, Interscience Publishers, 1967.

[75] Roe R. L., Approximate Riemann solvers, parameter vectors, and difference schemes, J. Comput. Phys., 43(1981), 357-372.

[76] Schmidt T., Box schemes on quadrilateral meshes, Computing, 51(1993), 271-292.

[77] Scholz R., A mixed method for 4th order problems using linear finite elements, R.A.I.R.O. Anal. Numer., 12(1978), 85-90.

[78] Selmin V., The node-centered finite volume approach: Bridge between finite differences and finite elements, Computer methods in Applied Mechanics and Engineering, 102(1993), 107-138, North-Holland.

[79] Selmin V. and Formaggia L., Unified construction of finite element and finite volume discretizations for compressible flows, Int. J. for Numer. Meth. in Engin., 39(1996), 1-32.

[80] Spalding D. B., A novel finite difference formulation for differential expression involving both first and second derivatives, Int. J. Num. Meth. Engin., 4(1972), 551-559.

[81] Süli E., The accuracy of finite volume methods on distorted partitions, The Mathematics of Finite Element and Applications, VII(Uxbridge, 1990), 253-260, Academic Press, London, 1991.

[82] Süli E., The accuracy of cell vertex finite volume methods on quadrilateral meshes, Math. Comp., 59(1992), 359-382.

[83] Tabata M., Uniform convergence of the upwind finite element approximation for semilinear parabolic problems, J. Math. Kyoto Univ. (JMKYAZ), 18(2)(1978), 327-351.

[84] Tabata M., A finite element approximation corresponding to the upwind finite differencing, Memoirs of Numer. Math., 4(1977), 47-63.

[85] Tabata M., Conservative upwind finite element approximation and its applications, in Analytical and Numerical Approaches to Asympotopic Problem in Analysis, Noth-Holland Publishing Company, 369-381, 1981.

[86] Teman R., Navier-Stokes Equations, Theory and Numerical Analysis, 3nd ed., North-Holland, Amsterdam, 1984.

[87] Thomas J. M. and Trujillo D., Analysis of finite volume methods, Technical Report 19, Universite de Pau et des Pays de L'adour Pau, France, 1995.

[88] Thomas J. M. and Trujillo D., Convergence of finite volume methods, Technical Report 20, Universite de Pau et des Pays de L'adour Pau, France, 1995.

[89] Tseng A. A. and Gu S. X., A finite difference scheme with arbitrary mesh system for solving high-order partial differential equations, Comput. and Structures, 31(1989), 319-328.

[90] Vanselow R., Relation between FEM and FVM applied to the Poisson equation, Computing, 57(1996), 93-104.

[91] Vanselow R. and Scheffier H. P., Convergence analysis of a finite volume method via a new nonconforming finite elment method, Numer. Meth. for Partial Diff. Equat., 14(1998), 213-232.

[92] Van Leer B., Towards the ultimate conservative difference scheme V, A second-order sequel to Godunov's method, J. Comput. Phys., 32(1979), 101-136.

[93] Varga R. S., Matrix Iterative Analysis, Prentice-Hall, 1962.

[94] Vassilevski P. S., Petrova S. I. and Lazarov R.D., Finite difference schemes on triangular cell-centered grids with local refinement, SIAM J. Sci. Stat. Comput., 13(1992), 1287-1313.

[95] Vignal M. H., Convergence of a finite volume scheme for an elliptic-hyperbolic system, Math. Model, and Numer. Anal., 30(1996), 841-872.

[96] Vila J. P., Convergence and error estimates in finite volume schemes for general multidimensional scalar conservation laws I, Explicit monotone schemes, RAIRO Math. Model. Numer. Anal., 28(1994), 267-295.

[97] Wang J. C. T. and Widhopf G. F., A high-resolution TVD finite volume scheme for the Euler equations in conservation form, J. Compt. Phys., 84(1989), 145-173.

[98] Weiser A. and Wheeler M. F., On convergence of block-centered finite differences for elliptic problems, SIAM J. Numer. Anal., 25(1988), 351-375.

[99] Winslow A. M., Numerical solution of the quasilinear Poisson equation in a nonuniform triangle mesh, J. Comp. Phys., 1(1967), 149-172.

[100] Yserentant H., On the multi-splitting of finite element space, Numer. Math., 49(1986), 379-412.

## Part C. (References written in Rusian)

[1] Fryazinov I. V. and Maslyankicina L. A., On the finite approximation of elliptic and parabolic equations on irregular networks, Preprint No. 49, IPM. ANSSSR, (1977), 1-63.

[2] Fryazinov I. V., Construction of finite difference schemes on a pair of irregular networks, Preprint No. 23, IPM. ANSSSR, (1979), 1-57.

[3] Fryazinov I. V., Alternating balance methods and variational-difference schemes on irregular networks, Preprint No. 53, IPM. ANSSSR, (1979), 1-47.

[4] Fryazinov I. V., Balance methods and variational-difference schemes, Diff. Equat., 16(1980), 1332-1342.

[5] Gadunov S. K., A finite difference method for the numerical computation and discontinuous solutions of the equations of fluid dynamics, Math. Sb., 47(1959), 271-305.

[6] Kantorovich L. V. and Krylov V. I., Approximate Methods of Higher Analysis, Moscow, 1952.

[7] Samarski A. A. and Andreev V. B., Finite Difference Methods for Elliptic Equations, Moscow, 1976.

[8] Samarski A. A., Some theoretical results on difference methods, Diff. Equat., 16(1980), 1155-1171.

# Index